PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure et INRIA, UMR 8548

Modelling functions with kernels : from logistic regression to global optimisation

Soutenue par
Ulysse MARTEAU-FEREY
Le 7 septembre 2022

École doctorale n°386
Sciences mathématiques de Paris centre

Spécialité
Mathématiques

Composition du jury :

Francis BACH INRIA-DIENS-PSL	*Directeur de thèse*
Gilles BLANCHARD Université Paris-Saclay	*Rapporteur*
Claire BOYER LPSM, Sorbonne Université	*Examinatrice*
Antonin CHAMBOLLE CEREMADE-CNRS	*Examinateur*
Yohann DE CASTRO ICJ-École Centrale de Lyon	*Examinateur*
Aymeric DIEULEVEUT CMAP	*Examinateur*
Jean-Bernard LASSERRE LAAS-CNRS	*Rapporteur*
Alessandro RUDI INRIA-DIENS-PSL	*Co-encadrant*

1794
ENS | PSL

Remerciements

"Je deumerai longtemps errant dans Césarée"
Racine, *Bérénice*.

Il me paraitrait impensable de commencer ce manuscrit sans remercier les nombreuses personnes sans qui je n'aurais sûrement ni fait de thèse, ni autant aimé ces quatre années.

Tout d'abord Francis et Alessandro, qui ont su à la fois me recruter à un moment ou je n'avais pas les idées très claires, me guider au début puis me laisser mariner pendant des mois. Votre bienveillance tout au long de ce processus et particulièrement pendant les différents passages difficiles de ces années m'a été très précieuse. Merci aussi d'avoir été aussi flexibles et de m'avoir permis de laisser de coté le travail pendant quelques mois pour organiser les quelques événements personnels qui ont aussi marqué ces années de thèse. Merci enfin de m'avoir fait découvrir autant de sujets, et d'avoir partagé avec moi votre curiosité sur des thèmes parfois oubliés mais toujours passionnants. Vous m'avez appris a travailler en m'amusant. Et merci d'avoir créé et entretenu cette petite bulle de bonne humeur qu'est Sierra; je pense qu'elle vous doit beaucoup.

Je souhaites aussi remercier mes deux rapporteurs, Gilles Blanchard et Jean-Bernard Lasserre, qui ont accepté de relire ce pavé. Vos retour et vos commentaires, ainsi que vos travaux, ont été et restent pour moi une grande source d'inspiration. Merci également aux autres membres de mon jury, Aymeric Dieuleveut, Claire Boyer, Antonin Chambolle et Yohann de Castro d'avoir accepté d'assister, tous en présentiel, à ma soutenance.

Je voudrais également remercier tout particulièrement le LAAS, Jean-Bernard Lasserre, Didier Henrion, Victor Magron, de m'avoir permis de découvrir le monde merveilleux de l'optimisation polynomiale, et de ses nombreuses applications, et de m'avoir invité à leurs séminaires et conférences. Un grand merci également à Rémi Gribonval et Yohann de Castro, qui, en m'invitant mi-mai à Lyon, m'ont donné l'entrain et la bouffée d'air frais (et de discussions intéressantes) salvateurs au milieu de la rédaction de thèse.

Ensuite, je souhaiterais remercier l'Inria de nous offrir un cadre de recherche extraordinaire. Un merci particulier aux équipes de communication et médiation, qui m'ont permis de profiter de la recherche différemment, et de découvrir des personnes à l'énergie incroyable. Prendre quelques après midi pour aller parler de sciences à des collégiens en prenant des cafés avec Hélène et Anais et d'autres doctorants sont une expérience que je conseille à tout nouveau doctorant.

Le bureau C406 est passé par différentes phases sociales au cours de mes quatre ans et demi de thèse. Merci a Loucas d'en avoir fait un lieu de débats, intellectuels ou non (en atteste le jeu qui traine encore sur mon bureau), de passage, de bonne humeur, à Raphaël d'en avoir fait un haut lieu du golf parisien, de l'humour et de la sieste post prandiale, a Yann d'en avoir fait un lieu de recherche de vacances en vélo, de git commit et de partage d'experience de vie très intense, et a Céline d'y avoir apporté enfin la baisse de caféine qu'il fallait pour qu'on dorme tous plus de

cinq heures par nuit, tout ça pour aller à l'escalade le plus souvent possible se casser les chevilles afin d'être enfin coincés au bureau pour être productifs.

J'en profite également pour remercier les différents professeurs qui m'ont donné tout au long de ma scolarité le goût des sciences et des mathématiques. Je pense particulièrement à mesdames Kleinman et Lowenfeld.

Pour relacher la pression de ces années ou pour partager les nombreux moments très heureux qu'elles ont contenus, je ne peux pas ne pas remercier tous mes amis, et particulièrement le filtre, avec qui la thèse ou le premier emploi ont été vécus un peu en commun. Merci à Séginus, pour son humour acéré, à Jean pour les soirées à discuter sans fin en errant dans Paris, à Cyril pour les jeux de mots abyssaux et pour sa confiance infinie en la recherche, à Rémy pour ce savant mélange de palabre, sport, maths, cuisine et bled qui a animé cette conversation dans ses heures les plus sombres. Merci à David pour les discussions de cuisine, pour les ballades dans Paris et en montagne, et à Marc pour sa capacité à passer de la barbe sérieuse et impassible aux rires aux éclats en une fraction de seconde. Merci à mes deux collocs, Thibaut et Simon, au lessives du soir et aux gateaux de ramadan en rentrant, a la salle du puzzle, a Aïcha écoute moi, aux barbecues a 100 personnes, aux moelleux au chocolat et aux tartes à la rhubarbe. Comme dit Rémy, la composante connexe s'agrandit : cela ne fait que rendre les choses encore plus amusantes.

Un immense merci à Juliette et Arielle d'avoir partagé avec moi le goût du thé dans un couloir ou cela sentait plutôt la bière. Le bruit et les odeurs de Chirac s'appliquait bien plus dans ce couloir de normaliens qu'à n'importe quel quartier de Paris. Nous avons survécu ensemble. Merci à Juliette de me permettre de mesurer combien la thèse en sciences est un luxe par rapport à celle de lettres, et à Arielle de me rappeler qu'à l'extérieur, la Start-up Nation !

Une dédicace particulière à Joseph, Julie, Louis, Quentin, Timothée, avec qui j'ai beaucoup partagé de choses malgré, pour certains, de bien grandes distances.

Le plus dur reste pourtant à faire : remercier ma famille. Il est évident que je ne serai pas la sans mes parents qui m'ont donné le goût des études et des maths. Votre soutien constant depuis toujours, et en particulier pendant mes années de prépa, m'ont fait passé cette période de la meilleure façon possible, et c'est grâce à vous que je n'ai jamais hésité à travailler ce qui me plaisait le plus, sans trop de prises de tête. Je ne serais pas non plus là sans mes soeurs, avec qui j'ai passé beaucoup de déjeuners ou de séance d'escalade sans leur adresser la parole car n'arrivant pas à me séparer d'un de mes problèmes dans ma tête. Votre joie de vivre, et vos recadrages subtils ont été un soutien précieux. Je sais, j'ai oublié de remercier crapou. Raby, je pense que si je rentre toujours au maximum le midi chez moi pour déjeuner, c'est en grande partie grâce à toi : merci d'avoir fait de la maison un endroit où l'on a toujours voulu rentrer le plus vite possible.

Je ne suis pas sûr que j'aurais persévéré en recherche, et encore moins autant aimé cela, sans la joie et la confiance que Juliette m'a apportées durant ces quatre années. Partager avec toi les moments forts, un peu durs et joyeux, a été et sera toujours inestimable. Ces années de thèse seront toujours indissociables de ces premières années tous les deux.

Et enfin, merci à Bérénice, qui au delà de l'immense joie qu'on a tous les jours de te voir grandir, a su faire en sorte que je sois le plus efficace possible pour préparer ma soutenance.

Contributions, thesis outline and reading guide

Contributions presented in this thesis

This thesis is divided into three parts, each corresponding to two contributions, which are either published or submitted works.

Articles and preprints presented in part I

This part is dedicated to the study of statistics and optmization of empirical risk minimization in the kernel method setting, in the case where the loss is the logistic loss. It presents the following two articles, which have been published.

- Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2294–2340. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/marteau-ferey19a.html
- Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Newton methods for ill-conditioned generalized self-concordant losses. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/60495b4e033e9f60b32a6607b587aadd-Paper.pdf

Articles and preprints presented in part II

This part is dedicated to the development of a model for non-negative functions which we call *PSD models*. The first article defines and analyses the main properties of these models, while the second applies such a model to design an algorithm for i.i.d. sampling from an un-normalized density function. Both of these articles have been published.

- Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12816–12826. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/968b15768f3d19770471e9436d97913c-Paper.pdf
- Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Sampling from arbitrary functions via psd models. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors,

Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of *Proceedings of Machine Learning Research*, pages 2823–2861. PMLR, 28–30 Mar 2022a. URL https://proceedings.mlr.press/v151/marteau-ferey22a.html

Articles and preprints presented in part III

This last part is dedicated to the use of PSD models in the context of global optimization. The first article develops a method for global optimization of regular functions which leverages regularity to break the curse of dimensionality. The second article has a more theoretical flavour and provides sufficient conditions for regular functions to be decomposed as sum of squares of regular functions. This property is crucial for our global optimization scheme. Both of these articles have been submitted and are under review.

- Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations, 2020. URL https://arxiv.org/abs/2012.11978
- Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Second order conditions to decompose smooth functions as sums of squares, 2022b. URL https://arxiv.org/abs/2202.13729

Organisation of the manuscript, and reading guide

We start by a general introduction in chapter 1. It introduces the high level context about machine learning and reproducing kernel Hilbert spaces needed to read this thesis, and provides a simple outline of the rest of the thesis.

The core of the manuscript is divided into three parts. Each part is introduced with either a chapter (for parts I and III) or a small introduction which describes the necessary elements of context, which are sometimes less detailed in the articles (see chapters 2 and 7 as well as the introduction to part II). It is then followed by chapters which are *verbatim* of the articles related to the part, where all the precise results and proofs can be found. Note that we have put all articles and appendices for this thesis to be self-contained.

The final chapter (chapter 10) is a conclusion and highlights the different research directions opened up to us by this thesis.

In a first pass over (part of) this thesis, we invite the reader to read only the introductions; they are made to be self contained in terms of concepts and results. The verbatim articles are divided into a main part, which gives context and results, and an appendix, where most of the technical proofs may be found. As these appendices go into much detail, they are very long; the reader should not be scared by that length as it includes extra information, can repeat itself between articles, and can be overlooked by just reading the main parts of the articles.

Contents

1 General Introduction **9**
- 1.1 Introduction to machine learning . . . 10
- 1.2 Reproducing kernel Hilbert spaces . . . 32
- 1.3 Outline of the thesis . . . 46

I Fast rates and algorithms for generalized self-concordant losses **51**

2 Backround and main results **55**
- 2.1 Fast rates for least-squares . . . 57
- 2.2 Generalized self-concordance . . . 73
- 2.3 Main results and contributions of this part . . . 80

3 Beyond least squares **93**
- 3.1 Introduction . . . 94
- 3.2 Main Assumptions and Results . . . 96
- 3.3 Slow convergence rates . . . 99
- 3.4 Faster Rates with Source Conditions . . . 100
- 3.5 Rates for source and capacity . . . 101
- 3.6 Sketch of the proof . . . 103
- 3.7 Conclusion . . . 104
- 3.A Setting, definitions, assumptions . . . 105
- 3.B Generalized self-concordant losses . . . 107
- 3.C Main result, simplified . . . 110
- 3.D Main result, refined analysis . . . 115
- 3.E Explicit bounds for the simplified case . . . 123
- 3.F Explicit bounds for the refined case . . . 125
- 3.G Additional lemmas . . . 129

4 Second order strikes back **137**
- 4.1 Introduction . . . 138
- 4.2 Backround . . . 141
- 4.3 Globally convergent scheme . . . 143
- 4.4 Application to kernel methods . . . 145
- 4.5 Experiments . . . 147
- 4.A Generalized self-concordance . . . 150
- 4.B Approximate Newton methods . . . 155
- 4.C Globalization scheme . . . 165
- 4.D Non-parametric learning . . . 173

4.E Algorithm . . . 187
4.F Experiments . . . 189
4.G Projected problem . . . 193
4.H Expected versus empirical risk . . . 198
4.I Bounds for Hermitian operators . . . 213

II Positive semidefinite models : theory and applications 219

Introduction : representing non-negative functions in a flexible way 223

5 PSD models 225
5.1 Introduction . . . 225
5.2 Background . . . 226
5.3 Proposed Model for Non-negative Functions . . . 228
5.4 Approximation Properties of the Model . . . 231
5.5 Extensions: Integral Constraints and Output in Convex Cones . . . 232
5.6 Numerical Simulations . . . 234
5.A Notation and basic definitions . . . 237
5.B Proofs and additional discussions . . . 237
5.C Additional proofs . . . 258
5.D Additional details on the other models . . . 258
5.E Additional details on the experiments . . . 261
5.F Relation to similar work . . . 263

6 Sampling from arbitrary functions via PSD models 265
6.1 Introduction . . . 265
6.2 Backround on Positive Semi-Definite (PSD) models . . . 266
6.3 A sampling algorithm for PSD models . . . 268
6.4 Sampling from arbitrary distributions using PSD models . . . 273
6.5 Experiments . . . 279
6.6 Extensions, future work . . . 280
6.A Definitions and notations . . . 284
6.B Properties of the Gaussian RKHS . . . 289
6.C Properties of Gaussian PSD models . . . 294
6.D The sampling algorithm . . . 297
6.E A general method of approximation and sampling . . . 302
6.F Approximation and sampling using a rank one PSD model . . . 309
6.G Additional experimental details . . . 314

III Sum of squares of functions 317

7 A parallel with moment-SOS hierarchies 321
7.1 Polynomial optimization . . . 322
7.2 Global optimization through kernel sums of squares . . . 332
7.3 Similarities and differences between the two approaches . . . 342

8 Finding global minima via kernel approximations 347
8.1 Introduction . . . 348

8.2 Outline of contributions . . . 349
8.3 Setting . . . 353
8.4 Equivalence of the infinite-dimensional problem . . . 356
8.5 Properties of the finite-dimensional problem . . . 360
8.6 Algorithm . . . 366
8.7 Finding the global minimizer . . . 368
8.8 Extensions . . . 370
8.9 Relationship with polynomial hierarchies . . . 373
8.10 Experiments . . . 375
8.11 Discussion . . . 380
8.A Additional notations and definitions . . . 384
8.B Fundamental results on scattered data approximation . . . 390
8.C Auxiliary results on RKHS . . . 392
8.D The constants of translation invariant and Sobolev kernels . . . 393
8.E Proofs for algorithm 6 . . . 399
8.F Global minimizer. Proofs. . . . 404
8.G Proofs for the extensions . . . 407
8.H Details on the algorithmic setup used in the benchmark experiments . . . 411

9 SOS decompositions of smooth functions 419
9.1 Introduction . . . 419
9.2 Decomposition as sums of squares given second order conditions (Euclidean case) 424
9.3 Global decomposition as a sum of squares for functions on manifolds . . . 431
9.4 Proof of the local decomposition as a sum of squares . . . 437
9.5 Discussion and possible extensions . . . 439
9.A Around partitions of unity and gluing functions . . . 441
9.B Morse lemma . . . 443

Conclusion and perspectives 444

Chapter 1

General Introduction

Contents

1.1 Introduction to machine learning 10
1.2 Reproducing kernel Hilbert spaces 32
1.3 Outline of the thesis 46

In many fields of applied mathematics, from machine learning to physics, we aim to find a certain mathematical object o in a set of mathematical objects $\mathcal{O}$, which will help us tackle a certain task. This object can be a function $f : \mathcal{X} \to \mathcal{Y}$, a point in a set $x \in \mathbb{R}^d$ with particular properties, a trajectory $(x_t)_{t\in T}$, or even a measure μ on a space $\mathcal{X}$. In many cases, this object can be thought of as the solution to an "ideal" optmization problem : we have a functional $F : \mathcal{O} \to \mathbb{R}$, and wish to find an object o_* such that

$$F(o_*) = \inf_{o\in\mathcal{O}} F(o), \tag{1.1}$$

where "ideal" functional F incorporates everything there is to know about the problem. In practice however, we cannot solve Eq. (1.1), because *a)* we only have part of the information about the ideal problem at hand, *b)* we have to be able to perform computations on a computer (on which we cannot necessarily represent the ideal object $o \in \mathcal{O}$ or the function F), and *c)* we have only a finite computational budget, in terms of time and memory, and therefore have to use a finite-time algorithm. Instead of solving the ideal problem Eq. (1.1), we therefore solve a surrogate problem

$$\hat{o} = \mathrm{Alg}\left(\widetilde{F}, \widetilde{\mathcal{O}}\right), \qquad \widetilde{F}(\hat{o}) \approx \inf_{\widetilde{o}\in\widetilde{\mathcal{O}}} \widetilde{F}(\widetilde{o}) \tag{1.2}$$

where Alg is an algorithm which computes an approximation of the solution of the surrogate problem in finite time on a computer. The function $\widetilde{F}$ is usually called the **surrogate function** and incorporates *a priori* information on the objective F, the set $\widetilde{O}$ is called a **model** for $\mathcal{O}$ and incorporates a priori information (like shape constraints or regularity of the target), and is representable in a concise way on a computer, and the optimization algorithm Alg is taylored to the surrogate problem, and has a given time and space complexity.

As we already see in this simple setting, algorithms, models, surrogate problems are interconnected. Understanding the quality of the returned object $\hat{o}$, measured by the quantity $\mathcal{R}(\hat{o}) = F(\hat{o}) - \inf_{o\in\mathcal{O}} F(o)$, requires an understanding of all these components and how they relate. Breaking down the roles of these different parts is crucial to understand the limits of each one. In this thesis, in parts I to III, we will try to look at all these aspects, to fully grasp the applied mathematics problem at hand. We will formulate the ideal problem Eq. (1.1), then present a model as well as

a surrogate problem for the function (this can be done in different steps, as in part I), before designing an algorithm whose complexity we can control. We hope that the reader will be able to see and put together these different steps, and that it will expose the interactions between these different phases, how they can limit or help each other out.

This introduction is structured into three parts. In Sec. 1.1, we will introduce the main applied mathematical setting of this thesis. We will almost completely focus on the statistical machine learning setting which is the setting of parts I and II. In particular, we will present a framework for constructing a surrogate problem from an ideal statistical problem, by assuming that instead of knowing the entire distribution (which would be the ideal case), we only have access to it through samples. We will also present basic **optimization algorithms**, to approximately solve the surrogate problem. We will end this section by briefly presenting a simplified setting of part III.

In Sec. 1.2, we will present the main classes of models which we will explore in this thesis, linear models and **kernel methods**. The goal is to center the attention of the reader on the importance of modelling, and how it interacts with both statistics and optimization questions.

Finally, in Sec. 1.3, we will give a brief outline of the thesis, and of the content of each part.

1.1 Machine learning and a little bit of global optimization

In this section, we will present some backround on the problems we worked on in this thesis. Most of these problems in parts I and II are formulated in the context of statistical machine learning. That is why in Secs. 1.1.1 to 1.1.3, we will present the general machine learning framework and context in terms of ideal problems (expected risk minimization), surrogate problems (empirical risk minimization) and optimization algorithms. Instead, in Sec. 1.1.4, we briefly present these three phases in the context of part III.

In parts I and II, we will deal with problems of the form Eq. (1.1) in the context of machine learning. In this context, the object to find (or to learn) is usually a function $f : \mathcal{X} \to \mathcal{Y}$ from an input space $\mathcal{X}$ to an output space $\mathcal{Y}$.

The specificity of machine learning is that one of the main (if not the main) *a priori* knowledge that we use to learn the function f is **data** : we are given a finite set $(z_i)_{1 \leq i \leq n} \in \mathcal{Z}^n$ of examples z_i living in a set $\mathcal{Z}$ which help us to learn f. For comprehensive introductions to machine learning, one can refer to Shalev-Shwartz and Ben-David (2014); Hastie, Tibshirani, and Friedman (2001). In general, the complexity of a machine learning problem can be roughly summarized by two key quantities.

- d the dimension of the input space $\mathcal{X}$ of the function f (d if $\mathcal{X}$ is a subset of $\mathbb{R}^d$ for instance, or other notions of dimensions such as the one proposed by Vapnik (1995));
- n the number of data points available to learn the function f.

In most modern machine learning tasks, the scale is large, meaning that n is large, and the dimensionality d of the data points is also large. While these differences are not set in stone, the large scale / large dimensional setting of machine learning has driven the development for new methods adapted to these specificities, compared to the traditional signal processing world. We do not mention anything on $\mathcal{Y}$ here. This is because, in this thesis, we will mostly take $\mathcal{Y}$ to be equal to $\mathbb{R}$, a subset of $\mathbb{R}$, or $\{-1, 1\}$ in the case of a classification predictor. The case where $\mathcal{Y}$ is large of infinite dimensional can also be of great importance, but we do not treat it here.

In the literature, we distinguish two main kinds of machine learning tasks.

Supervised learning. (Hastie, Tibshirani, and Friedman, 2001; Vapnik, 1995) In this setting, we want to predict an output $y \in \mathcal{Y}$ from an input $x \in \mathcal{X}$. For example, we might want to predict whether a patient will survive or not ($\mathcal{Y}$ is therefore binary) given its medical record and treatment, encoded in $\mathcal{X}$, whether a particle detected at the CERN through measurements is a Higgs Boson or not (Baldi, Sadowski, and Whiteson, 2014), or what price should be set given customer data. In that case, the data usually takes the form $z_i = (x_i, y_i)$ of examples of input output pairs, the point being that from examples of on inputs-outputs, we can learn a function f which will perform well on new data points x. In supervised learning, a distinction is often made between *classification*, where $\mathcal{Y}$ is finite (as in the Higgs Boson case), and *regression*, where one wishes to learn a continuous variable $\mathcal{Y}$ (as in the sales case). In part I, we handle problems which are typical supervised learning problems, such as logistic regression.

Unsupervised learning. (Shawe-Taylor and Cristianini, 2004; Hastie, Tibshirani, and Friedman, 2001) In this setting, the goal is to directly extract some meaningful information from the data. Examples of unsupervised learning tasks include the learning of a generative model, *i.e.*, given examples $x_1, ..., x_n$ sampled from a certain distribution ρ, create a sampler which samples from a distribution $\hat{\rho}$ close enought to ρ, the learning of the mean of a distribution through samples, the extraction of meaningful low dimensional features from high dimensional data, clustering, etc. In part II, we present a model which can tackle certain unsupervised learning tasks, such as the task of generating samples.

Whether it be in a supervised or unsupervised setting, most machine learning problems boil down to learning a function $f : \mathcal{X} \to \mathcal{Y}$ from a finite amount of seen data, in order to use it on other unseen data. In order to formalize the notion of seen and unseen data, as well as to understand how one may use data to infer information, the complexity of the full problem is often encapsulated into a probability distribution, to which we have limited access through the data : this is the statistical learning theory framework.

1.1.1 Statistical learning framework

In the statistical learning setting, the data points $z_1, ..., z_n$ are assumed to be realisations of a random variable Z on $\mathcal{Z}$ (therefore equipped with a σ-algebra), with ρ the associated probability measure on $\mathcal{Z}$. Implicitly, the statistical learning setting assumes that all the information on the data can be contained in the random variable Z. One of the very standard assumptions is that the data points $z_1, ..., z_n$ are independent and identically distributed (i.i.d.) samples from Z.

Assumption 1.1 (iid samples)**.** *The samples $z_1, ..., z_n$ are i.i.d. samples from Z.*

We will make this assumption in this thesis. In practice, this assumption may not be satisfied. In particular, when dealing with time series, the independence assumption may not be true (think of a Markov chain). Moreover, in some settings, the data is collected from multiple sources, which induces a difference in distribution depending on the source. This is the case when dealing with medical data coming from different medical centers for instance, where images are acquired using different tools; the distribution of the images depends on the acquisition device, and hence the data used are not identically distributed.

Supervised learning. In the particular case of supervised learning, the data usually takes the form of input-output pairs, *i.e.*, the samples (x_i, y_i) are assumed to be drawn from a joint

distribution $Z = (X, Y)$. Denote with $\rho_{\mathcal{X}}$ the probability measure of X and assume that ρ can be decomposed as $d\rho = \rho(dy|x)d\rho_{\mathcal{X}}$ for a certain $\rho(\cdot|\cdot) : \mathfrak{S}(\mathcal{Y}) \times \mathcal{X} \to [0, 1]$ where $\mathfrak{S}(\mathcal{Y})$ denotes the σ-algebra of $\mathcal{Y}$ and $p(\mathcal{Y}|x) = 1$, $\rho_{\mathcal{X}}$-almost everywhere (a.e.). This is simply the formal definition of conditional probabilities. There are two sources of randomness in the supervised learning setting : *a)* a randomness of the input data, from which we draw the x_i, and *b)* a randomness of the output data conditioned on the input, *i.e.*, the fact that y_i is drawn from $\rho(dy|x)$. If $Y = f(X)$ a.e., then Y is actually deterministic knowing x and $\rho(dy|x) = \delta_{f(x)}$ is simply a dirac for almost every x. In this case, we assume that Y can fully be recovered from X. However, it is typically not the case, because X does not contain enough information to predict Y, or because the observations are noisy.

Generalization error. Assume first that we are in the supervised learning setting, and that the data follows the distribution ρ of (X, Y). The goal is to find a good predictor $f : \mathcal{X} \to \mathcal{Y}$, such that $f(X)$ is a good approximation of Y. A natural space to look for this predictor is the set $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. To measure the quality of a prediction $f(x)$ compared to $y \in \mathcal{Y}$, we define a loss function *a priori*. A loss function will be a map $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{M}(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}_+$, such that $\ell((x, y), f)$ quantifies the error performed when approximating y by $f(x)$. The **risk** or **generalization error** of a function f is then defined as

$$\mathcal{R}(f) := \mathbb{E}\left[\ell((X, Y), f)\right], \qquad f \in \mathcal{M}(\mathcal{X}, \mathcal{Y}), \tag{1.3}$$

where the expectation is taken over ρ. We also define $\mathcal{R}_* := \inf_{f \in \mathcal{M}(\mathcal{X},\mathcal{Y})} \mathcal{R}(f)$ to be the smallest possible risk. If this optimum is reached for a certain function f_* (note that there can be none, one or more than one), f_* is called an optimal predictor. The **excess risk** of a function f is defined as the quantity $\mathcal{R}(f) - \mathcal{R}_*$. Going back to the introduction of this chapter, finding f_* is therefore our ideal problem Eq. (1.1).

More generally, going beyond the previous setting, we will consider losses of the form $\ell : \mathcal{Z} \times \mathcal{F} \to \mathbb{R}_+$, where $\mathcal{Z}$ is the data space (equipped with the random variable Z), $\mathcal{F}$ is the space of functions (predictors, probability distributions, probabity samplers or even single points) suited for the problem. We assume $\ell(\cdot, f)$ is measurable for all $f \in \mathcal{F}$ and is the loss function. As before, $\ell(z, f)$ will measure how well f performs on data point z. We will also use the notation $\ell_z(f)$ in the place of $\ell(z, f)$. Note that in the supervised learning setting described above, we would define $\ell_z(f) = \ell_{(x,y)}(f)$ and $\mathcal{F}$ would be $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ *a priori*. Note that this formulation allows some flexibility : the output space of f might not be $\mathcal{Y}$ in the supervised setting (this is the case in certain classification problems). The general "ideal" goal is therefore to solve

$$\inf_{f \in \mathcal{F}} \mathcal{R}(f), \qquad \mathcal{R}(f) := \mathbb{E}\left[\ell(Z, f)\right], \tag{1.4}$$

and to find f with small excess risk $\mathcal{R}(f) - \mathcal{R}_*$.

Losses. As can been seen in Eq. (1.4), the choice of the loss defines what we consider to be a good inference from data and must therefore be chosen according to the problem at hand. However, as we will see in Sec. 1.1.3, the choice of the loss also impacts optimization procedures to solve the surrogate problem and must therefore be chosen with this in mind as well. We now present different settings and the associated typical losses.

Regression and square loss. In the setting of regression, that is in supervised learning where Y takes values in a sub-interval of $\mathbb{R}$ or more generally a subset of $\mathbb{R}^k$, one often uses the square

loss : $\ell_{x,y}(f) = l(y, f(x)) = \frac{1}{2}\|y - f(x)\|^2$. It is the most standard loss for regression and solving Eq. (1.4) is called the least squares regression problem. Note that as soon as $Y \in L^2$ (which seems a natural hypothesis in order for the risk to be defined), there is a solution to the ideal problem and it is simply the projection of Y onto $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ seen as a closed linear subspace of $L^2(\mathcal{X} \times \mathcal{Y}, \rho)$, also called the conditional expectation of Y given X, or the **Bayes predictor**. It is defined as

$$f_*(X) = \mathbb{E}[Y|X], \quad f_* \in \underset{f \in \mathcal{M}(\mathcal{X},\mathbb{R})}{\arg\min}\ \mathcal{R}(f) = \tfrac{1}{2}\mathbb{E}\left[\|Y - f(X)\|^2\right].$$

One can decompose the variable $Y = f_*(X) + \varepsilon$, where ε is called the noise and is orthogonal to the input variable X, in the sense that $\mathbb{E}[\varepsilon|X] = 0$. Note that $\mathcal{R}_* = \mathbb{E}\left[\varepsilon^2\right]$. In part I, and more specifically in Sec. 2.1, we will go into much greater detail in the analysis of least-squares regression, from a statistical but also computational perspective, and give the necessary references. In particular, we will see that the regularity of the Bayes predictor, as well as the hypotheses on the noise ε play a crucial role in this analysis.

Binary classification. In binary classification, we wish to predict a binary variable Y which has values in $\mathcal{Y} = \{-1, 1\}$ (this choice is arbitrary). The natural loss for such problems would be the $0-1$ loss, that is the loss $\ell_{x,y}(f) := \mathbf{1}_{f(x)\neq y}$ which penalizes f by 1 if the prediction is incorrect. However, while this loss is ideal from the point of view of the task, it neither smooth nor convex, and can be hard to optimize (see Sec. 1.1.3 for more details). Moreover, it is sometimes easier to parametrize functions f with vector valued outputs, and to allow f to take its values in $\mathbb{R}$ and not strictly $\{-1, 1\}$. This is easily remedied by allowing f to be real-valued and predict the output according to its sign $\operatorname{sgn}(f(x))$. When f is linearly parametrized, f is called a separating hyperplane, as $\{f(x) = 0\}$ defines the boundary between the two classes.

With our choice of $\mathcal{Y}$, the $0-1$ loss can simply be expressed as $\mathbf{1}_{yf(x)<0}$ even in the case where f is real-valued. The typical losses used in binary classification are of the form $\varphi(yf(x))$, where φ is a surrogate for $\mathbf{1}_{t<0}$. They include the Hinge loss $\varphi(t) = \max(0, 1-t)$ which is the convex relaxation of the indicator function, and is convex but is not smooth, and the logistic loss $\varphi(t) = \log(1 + e^{-t})$, which is smooth and convex. In this thesis, we will not discuss the hinge loss or Support Vector Machines in general. However, they are central losses and algorithms in the classification setting; a good reference can be found in the work by Steinwart and Christmann (2008). On the other hand, logistic regression will be at the center of part I. In Fig. 1.1, we show these three different loss functions (the corresponding φ) to illustrate their different convexity and smoothness properties.

Multi-class classification. In multi-class classification, we wish to predict a variable whose values lie in a finite set $\{0, 1, ..., K\}$ where $K \geq 1$. Of course, the $0-1$ loss is still the ideal loss, and can also be defined, and has the same issues than in binary classification in terms of smoothness and optimization difficulty. However, if ρ can be decomposed as a conditional law $\rho(dy|x)d\rho_{\mathcal{X}}$, the conditional law $\rho(dy|x)$ is necessarily a multinomial law whose conditional probability can be represented by a vector $(p_0(x), p_1(x), ..., p_K(x))$ on the simplex ($p_y \geq 0$ and $\sum_y p_y = 1$). One of the typical ways to proceed is to use multiclass logistic regression, that is to look for functions $f_y : \mathcal{X} \to \mathbb{R}$ such that $p_y(x) = \rho(y|x) \propto \exp(-f_y(x))$. The loss of $f = (f_y)_{y\in\mathcal{Y}}$ is then defined as

$$\ell_{x,y}(f) = -\log\left(\frac{e^{-f_y(x)}}{\sum_{i=0}^{K} e^{-f_i(x)}}\right).$$

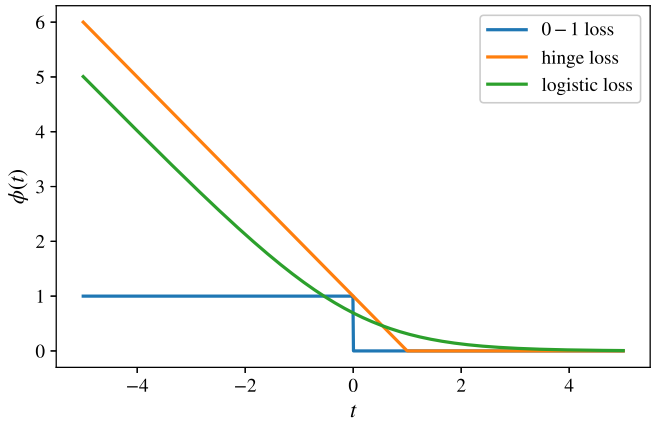

Figure 1.1: Different losses for binary classification

Note that $f : \mathcal{X} \to \mathbb{R}^{\mathcal{Y}}$ is not a predictor per say; however, given f, one defines the associated predictor $\tilde{f}$ simply by predicting $\tilde{f}(x) := \arg\min_{y \in \mathcal{Y}} f_y(x)$, that is the y with highest inferred probability, proportional to $e^{-f_y(x)}$. We will deal with these losses in more detail in part I, for certain classes of models.

In the case where $K = 1$ (binary classification), this is equivalent to logistic regression by taking $\mathcal{Y} = \{-1, 1\}$ and $f(x) = f_{-1}(x) - f_1(x)$.

Density estimation. This time, in a non-supervised setting, the goal is to perform density estimation, *i.e.*, try to approximate a density p on a space $\mathcal{X}$ with respect to a base measure $d\nu$ through i.i.d. samples $x_1, ..., x_n$ from $p\, d\nu$. In that case, the model $\mathcal{F}$ of functions should contain functions which are non-negative and sum to one with respect to $d\nu$. This is a typical case where one uses the negative log-likelihood loss:

$$\ell_x(f) = -\log(f(x)).$$

Such an approach is actually quite general and comes from statistics, as we will see in the following point. We use this loss in chapter 5 in part II to learn densities.

Log-likelihood. In classical statistics, it is common that the space $\mathcal{F}$ parametrizes a set of probability densities on $\mathcal{Z}$ with respect to a base measure $d\nu$, denoted with $\mathcal{P}_{\mathcal{F}} = \{p_f(d\nu) \in \mathscr{M}_1(\mathcal{Z}) \ : \ f \in \mathcal{F}\}$, where $\mathscr{M}_1(\mathcal{Z})$ is the set of probability measures on $\mathcal{Z}$. Given samples from the unknown distribution $\rho \in \mathscr{M}_1(\mathcal{Z})$, it is classical to look for $p_f \in \mathcal{P}_{\mathcal{F}}$ (or equivalently for $f \in \mathcal{F}$) which maximizes the expected log likelihood :

$$\max_{f \in \mathcal{F}} \mathcal{L}(f) := \mathbb{E}\left[\log(p_f(Z))\right] = \mathrm{KL}(p_f || \nu),$$

which can be seen as a *Kullback-Leibler* divergence between p_f and ν (for more details, see Sec. 2.1 in chapter 2). In machine learning, we adopt the point of view of minimizing a loss, and instead of maximizing the log-likelihood, we usually minimize the negative log-likelihood. Note that logistic regression, as presented above, or multi-class logistic regression is equivalent to this for conditional likelihoods $p(y|x)$.

Least-squares regression can also be seen as a maximization of the log-likelighood, for the model $\mathcal{P}_{\mathcal{F}} := \{\rho \in \mathscr{M}_1(\mathcal{X} \times \mathcal{Y}),\ d\rho = \exp(-\|y - f(x)\|^2/2)(d\rho_{\mathcal{X}} \otimes dy)\}$, that is we are looking for a conditional distribution of $Y|x$ as a Gaussian centered at $f(x)$.

1.1.2 Constructing a surrogate problem in statistical learning : empirical risk minimization

As explained in Sec. 1.1.1, we have access to the distribution of Z only through samples $z_1, ..., z_n$, and are therefore unable to directly solve Eq. (1.4). Instead, one can approximate this error by replacing ρ with the empirical distribution constructed from samples $\widehat{\rho} := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and define the so-called *empirical risk*

$$\widehat{\mathcal{R}}_n(f) := \mathbb{E}_{\widehat{\rho}}[\ell_Z(f)] = \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(f), \quad f \in \mathcal{F}. \tag{1.5}$$

For example, in the case of the square loss, the empirical risk can be written

$$\widehat{\mathcal{R}}_n(f) = \tfrac{1}{n}\sum_{i=1}^n \|f(x_i) - y_i\|^2, \quad f \in \mathcal{F}. \tag{1.6}$$

One could then think that simply minimizing the empirical risk over the class $\mathcal{F}$ would lead a small excess risk. However, there are two limiting factors to this approach : *a)* the space $\mathcal{F}$ can be so complex that it is not representable on a computer and *b)* the space $\mathcal{F}$ can be too large. If we go back to the example of supervised learning where $\mathcal{F} = \mathcal{M}(\mathcal{X}, \mathcal{Y})$, in the case of the empirical square loss Eq. (1.6), any interpolating function minimizes the empirical risk (and some of them could be wild). Note that in this introduction, $\mathcal{F}$ simply denotes a large set of functions on which the loss function is defined, and is typically given by the problem : $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ for supervised learning in general, $L^2(\mathcal{X}, \rho_X)$ for least-squares regression (since we know the minimizer to be in that space), the set of measurable functions which sum to one in the case of density estimation.

It is therefore important to minimize the empirical risk over a smaller space $\mathcal{H} \subset \mathcal{F}$ of functions adapted to the problem and to the amount of data we have at hand. A first surrogate problem in this spirit is therefore the following *empirical risk minimization* (e.r.m.) problem on $\mathcal{H}$:

$$\text{find } \widehat{f}_n \in \mathcal{H} \text{ such that } \widehat{f}_n \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(f). \tag{1.7}$$

As explained above, the choice of $\mathcal{H}$ is crucial. It must satisfy the following properties :

- the functions in $\mathcal{H}$ need to be representable on a computer;
- the e.r.m. problem Eq. (1.7) needs to be approximately solved by an optimizer;
- $\mathcal{H}$ must be small enough so that the solutions are not ill-behaved (in other words, $\mathcal{H}$ must incorporate enough *a priori* information compared to the amount of data);
- $\mathcal{H}$ must be large enough to approximate $\mathcal{F}$ well.

A classical yet powerful model for classes $\mathcal{F}$ of real-valued functions is to consider $\mathcal{H}$ in the form $\mathcal{H} = \{f \mid f(x) = \theta^\top \phi(x)\}$ for a given feature map $\phi : \mathcal{X} \to \mathbb{R}^p$. $\mathcal{H}$ is called a *linear model* with feature map ϕ, and p denotes the number of parameters of the model. In that case, it is easy to see that $f \in \mathcal{H}$ can be identified to the corresponding $\theta \in \mathbb{R}^p$ and is therefore representable on a computer as a vector of size p. For example, one may think of polynomials of degree at most $p-1$ on $\mathbb{R}$, which is a p dimensional set of functions parametrized by its coefficients θ where $\phi(x) = (1, x, ..., x^{p-1})$.

The size of $\mathcal{H}$ is determined both by the complexity of the feature map ϕ and the number p of parameters. Sec. 1.2 will be centered around a certain class of spaces $\mathcal{H}$ used in this thesis : the infinite dimensional counterpart of linear models, reproducing kernel Hilbert spaces (RKHS). We will also consider more complex models $\mathcal{H}$ of functions, based on these linear models, but which allow to approximate classes of functions $\mathcal{F}$ which are not necessarily real-valued in part II and part III.

In the following discussion, we will decompose the excess risk in order to mathematically show the different trade-offs between model capacity and number of samples.

Approximation and estimation error. Recall in this setting that $\mathcal{F}$ denotes a large "ideal" space where we expect to find f_* without assumptions, and that $\mathcal{H}$ is a smaller approximation of $\mathcal{F}$ (this notation is not standard; the notation $\mathcal{F}$ is sometimes used for the model). Let $\mathcal{R}_\mathcal{H} := \inf_{f \in \mathcal{H}} \mathcal{R}(f)$ be the the best achievable performance for functions in $\mathcal{H}$ and recall that the best achievable performance is defined as $\mathcal{R}_* = \inf_{f \in \mathcal{F}} \mathcal{R}(f)$. We can decompose the excess risk of an empirical risk minimizer $\widehat{f_n}$ into two terms :

$$\mathcal{R}(\widehat{f_n}) - \mathcal{R}_* = \underbrace{\mathcal{R}(\widehat{f_n}) - \mathcal{R}_\mathcal{H}}_{\text{estimation error}} + \underbrace{\mathcal{R}_\mathcal{H} - \mathcal{R}_*}_{\text{approximation error}} . \tag{1.8}$$

These two errors are often studied differently and have different meanings.

- The approximation error is the error one makes when modelling $\mathcal{F}$ with $\mathcal{H}$. This error decreases when $\mathcal{H}$ becomes larger.
- The estimation error is intrinsic to $\mathcal{H}$; it is the error that one makes when approximating ρ with the empirical distribution $\widehat{\rho}$, given the estimation space $\mathcal{H}$. For a fixed n, this error becomes larger as $\mathcal{H}$ increases.

In Fig. 1.2 we plot the typical evolution of the error when the size of $\mathcal{H}$ increases. We see that the typical evolution for the excess risk is to first decrease (a regime called underfitting : $\mathcal{H}$ is too small), before going back up again, because $\mathcal{H}$ is too large compared to the number of data points : this is called overfitting.

In Fig. 1.3, we further illustrate this underfitting-overfitting phenomenon in the context of polynomials. We take X to be uniform on $[0, 1]$ and $Y = f_*(X) + \varepsilon$ where f_* is a polynomial of degree 6 and ε is a Gaussian noise. We learn $f \in \mathcal{H}$ through least-squares regression with $n = 20$ samples, where we take $\mathcal{H} = \mathbb{R}_k[X]$ for increasing values of k to model increasing $\mathcal{H}$. The plot illustrates the underfitting phenomenon for small values of k (we see that a line or a degree 3 polynomial is not enough to approximate f_* well) , and the overfitting phenomenon for large values : in particular, f interpolates the points (x_i, y_i), *i.e.*, passes through all points. The graph on the bottom-right corner plots the error as a function of the degree k; we see the same U-shaped tendency described theoretically in Fig. 1.2.

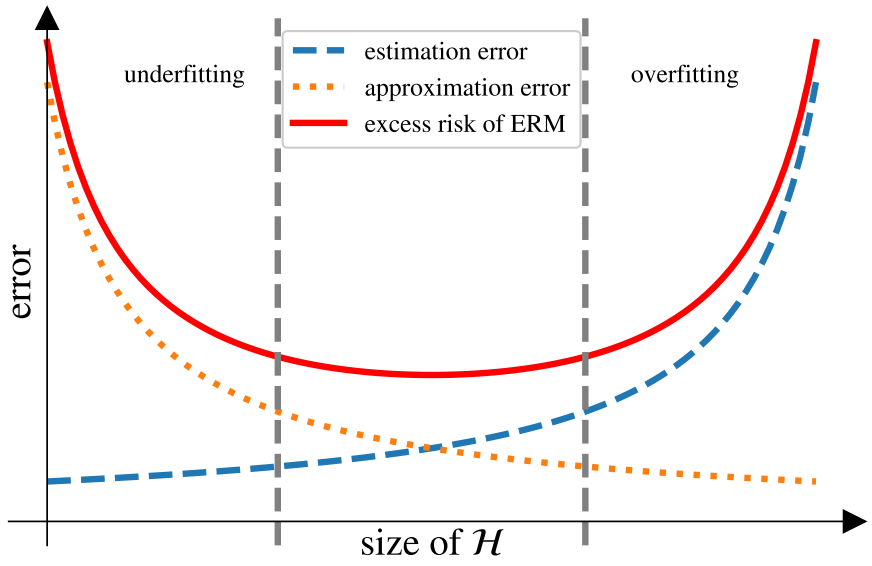

Figure 1.2: Evolution of the approximation error, estimation error and total excess risk for a prototypical machine learning problem.

Remark 1 (train/test)**.** *In practice, to evaluate the risk of a function f, learnt using n data points (throught e.r.m. for example), one uses a test set $z_{n+1}, ..., z_{n+m}$ of additional data (which are not in the training set) and uses the empirical loss on this additional data set to approximate the exptected risk. One can also have multiple test sets in order to obtain error bars on the risk, or perform K-fold cross validation. For more details, see Hastie, Tibshirani, and Friedman (2001).*

Regularization. In order to avoid overfitting, it is necessary to reduce the size of the space $\mathcal{H}$. Another method consists in penalizing functions in $\mathcal{H}$ which are too wild, *i.e.*, solve

$$\text{find } \widehat{f}_{n,\lambda} \in \mathcal{H} \text{ such that } \widehat{f}_{n,\lambda} \in \arg\min_{f \in \mathcal{H}} \widehat{\mathcal{R}}_n(f) + \lambda \Omega(f), \tag{1.9}$$

where Ω is a penalty function, *i.e.*, function controlling the complexity of f. λ is therefore a hyper-parameter which controls the penalization, and hence the effective size of the space (we say that it is a hyper-parameter because it is not optimized on during training but in the model selection phase). This will be made formal in part I and chapter 2. Note that restricting the functions set from $\mathcal{F}$ to $\mathcal{H}$ is actually already a form of penalization with the penalty $\Omega(f) = \iota_{\mathcal{H}}$, where $\iota_{\mathcal{H}}(f) = \begin{cases} 0 \text{ if } f \in \mathcal{H} \\ +\infty \text{ otherwise} \end{cases}$.
When $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ can be identified to a subset of $\Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$ (as in the case of linear models), typical penalties include

$$\Omega(f_\theta) = \|\theta\|_2^2 = \sum_{i=1}^{p} |\theta_i|^2 \text{ or } \Omega(f_\theta) = \|\theta\|_1 = \sum_{i=1}^{p} |\theta_i|.$$

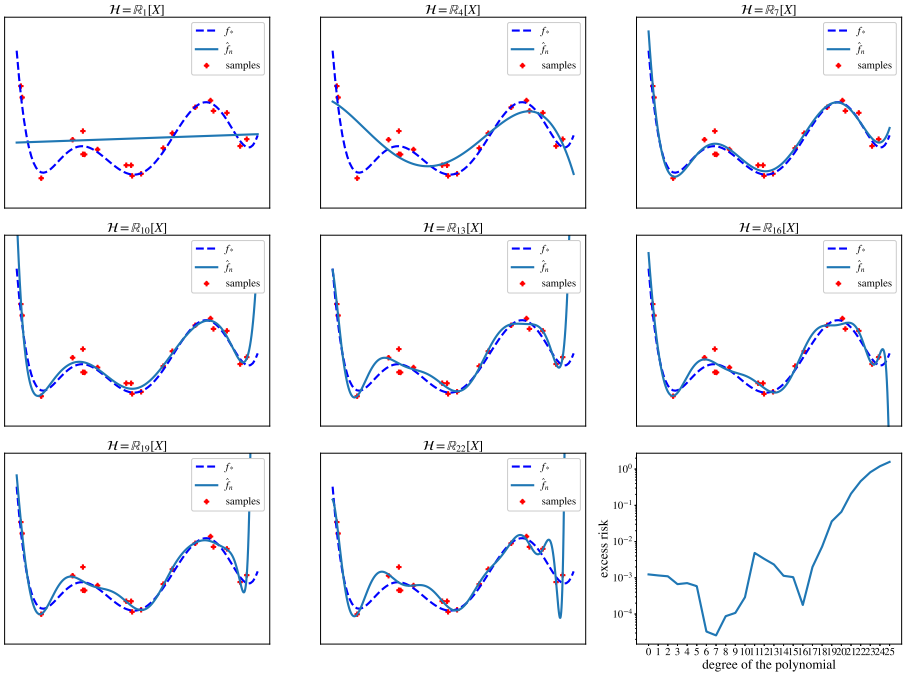

Figure 1.3: Learning from $n = 15$ noisy observation of a polynomial of degree 6, for different maximum degrees.

The first can be used as soon as $\mathcal{H}$ has a Euclidean or Hilbert space structure. Using this regularization with the square loss is called *ridge regression*. It will be at the center of part I, as we will use this regularization to control the size of the space $\mathcal{H}$. It is also beneficial from an optimization point of view, as it induces strong convexity (see Sec. 1.1.3). In Fig. 1.4, we perform the same task as in Fig. 1.3, but instead of making the degree vary, we fix the degree $k = 20$ and perform Ridge Regression for different values of λ. We see that λ really behaves like a size of $\mathcal{H}$ parameter, and that we can balance the fact that $\mathcal{H}$ is too large with some regularization.

The second is often used as a way of find a solution θ_* with a small number of non-zero coefficients. Each coefficient usually corresponds to a feature, which is activated if the coefficient is non-zero; these methods are called LASSO methods Tibshirani (1996); Hastie, Tibshirani, and Friedman (2001). We will briefly use these methods in chapter 5 and chapter 8.

Bounding the approximation error. Controling this error is usually done by assuming something on the distribution ρ of the data. For example, given a loss ℓ and $\mathcal{F}$, we can consider the set of distributions $\mathcal{M}(\mathcal{H}) := \{\rho \in \mathscr{M}(\mathcal{Z}) \ : \ \exists f_{\mathcal{H}} \in \mathcal{H}, \ \mathcal{R}(f_{\mathcal{H}}) = \mathcal{R}_*\}$, that is the set of distributions ρ such that there is an optimal predictor in $\mathcal{H}$. Assuming that $\rho \in \mathcal{M}(\mathcal{H})$ is usually called *well-specified assumption*, which implies $\mathcal{R}_{\mathcal{H}} = \mathcal{R}_*$ and hence a 0 approximation error. Other less stringent assumptions can help control the size of the approximation error.

If $\mathcal{H}$ is chosen to be "dense" in the set $\mathcal{F}$ for a metric linked to the loss (for example the L^2 distance in the case of least-squares regression, see part I, chapter 2), the approximation error can be shown to be zero (this is called a universality property in the context of linear models, see Sec. 1.2). When the approximation error is zero, one usually has to add some regularization, as

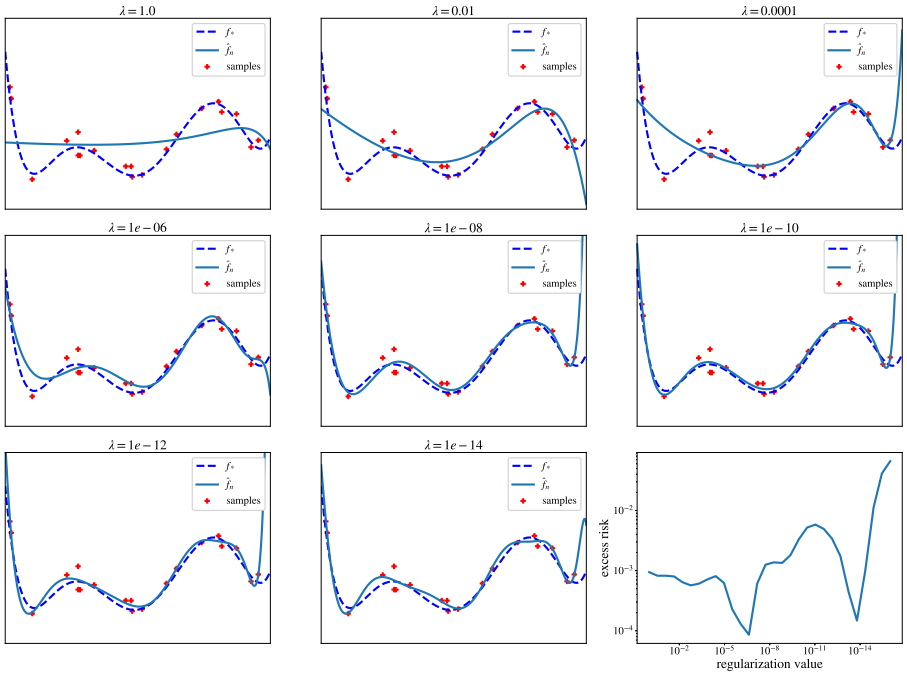

Figure 1.4: Learning from $n = 20$ noisy observation of a polynomial of degree 6, by solving ridge regression on $\mathcal{H} = \mathbb{R}_{20}[x]$ for different regularization values.

the space of functions is very large. In that case, another error plays the role of the approximation error : it is called the bias. This is the approach taken in part I.

Bounding the estimation error. There are two main classes of methods in order to bound the estimation error.

Rademacher complexities. Consider the case of empirical risk minimization (without regularization). One of the way to go, which we will not adopt in this thesis (except for one result in chapter 5 in part II to show statistical properties of the models we introduce), can be to bound the approximation error in the following way :

$$\begin{aligned}\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_{\mathcal{H}}) &= \mathcal{R}(\widehat{f}_n) - \widehat{\mathcal{R}}_n(\widehat{f}_n) + \underbrace{\widehat{\mathcal{R}}_n(\widehat{f}_n) - \widehat{\mathcal{R}}_n(f_{\mathcal{H}})}_{\leq 0} + \widehat{\mathcal{R}}_n(f_{\mathcal{H}}) - \mathcal{R}(f_{\mathcal{H}}) \\ &\leq 2 \sup_{f \in \mathcal{H}} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) \right|,\end{aligned}$$

that is we bound the estimation error by a uniform bound between $\mathcal{R}$ and $\mathcal{R}_n$. In empirical process theory, tools are developed in order to bound this uniform deviation (see Talagrand (1994) for a classical approach), and rely on the study of the tail of certain Gaussian processes. One of the main tools more recently formalized are *Rademacher complexities*, such as

$$\mathrm{Rad}_{z_1,\dots,z_n}(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i \ell_{z_i}(f)) \right| \right], \ \mathrm{Rad}_n(\mathcal{H}) = \mathbb{E}_{z_1,\dots,z_n \sim \rho^{\otimes n}} \left[\mathrm{Rad}_{z_1,\dots,z_n}(\mathcal{H}) \right],$$

where the σ_i are i.i.d. Bernoulli variables such that $\mathbb{P}(\sigma_i = \pm 1) = 1/2$. A Rademacher complexity is a measure of size of the set $\mathcal{H}$ with respect to the samples and the loss. For example, a

symmetrization argument shows that the uniform bound $\sup_{f\in\mathcal{H}}\left|\widehat{\mathcal{R}}_n(f)-\mathcal{R}(f)\right|$ is bounded in expectation by $\mathrm{Rad}_{z_1,...,z_n}(\mathcal{H})$, which bounds the expected estimation errors in the following way :

$$\mathbb{E}\left[\mathcal{R}(\widehat{f}_n)-\mathcal{R}(f_{\mathcal{H}})\right]\leq \mathrm{Rad}_n(\mathcal{H}),$$

where the expectation is taken with respect to the observed data $z_1,...,z_n$, which are i.i.d. samples from ρ. These techniques often provide the baseline to prove so-called slow rates of convergence, that is rates of the form $C/\sqrt{n}$ for a constant C (see for example Mohri, Rostamizadeh, and Talwalkar (2018), the Rademacher complexity obtained in chapter 5, or the discussion at the beginning of chapter 3). They are very general, and cover many loss functions. They can also be used to prove faster rates (when dealing with regularization for example) by defining *localized Rademacher complexities* (Bartlett, Bousquet, and Mendelson, 2005; Sridharan, Shalev-Shwartz, and Srebro, 2009).

Minimax upper and lower rates of convergence. Another point of view on bounding the estimation error is to study so-called "rates". Here, the setting is a bit more involved. Let $\mathcal{M}$ be a class of probability distributions ρ on $\mathcal{Z}$, which is our hypothesis space (we assume that the data measure ρ is in $\mathcal{M}$). $\mathcal{M}$ can be seen as a hypothesis on the regularity of the expected risk minimization problem. Formally, an n-estimator $\widehat{f}_n$ in $\mathcal{H}$ is a function which to n data points $(z_1,...,z_n)\in\mathcal{Z}^n$ maps a function in $\mathcal{F}$ (also denoted $\widehat{f}_n$ by abuse of notation, for example the empirical risk minimizer); an estimator is therefore a random quantity in terms of the data samples $z_1,...,z_n$. Given the samples and a distribution $\rho\in\mathcal{M}$, the performance of the estimator $\widehat{f}_n$ is $\mathcal{R}(\widehat{f}_n)-\mathcal{R}_*=\mathbb{E}_\rho\left[\ell_Z(\widehat{f}_n)\right]-\inf_{f\in\mathcal{H}}\mathbb{E}_\rho\left[\ell_Z(f)\right]$.

Informally an upper/lower/optimal **rate** $(r_n)_{n\in\mathbb{N}}$ is a sequence of positive numbers such that the performance of the best possible estimator $\widehat{f}_n$ is upper/lower/upper-and-lower bounded by a positive constant factor of r_n. Reformulating this, we ask that $r_n^{-1}\left(\mathcal{R}(\widehat{f}_n)-\mathcal{R}_*\right)$ is upper/lower/upper-and-lower bounded by a positive constant as n goes to infinity. Since $r_n^{-1}\left(\mathcal{R}(\widehat{f}_n)-\mathcal{R}_*\right)$ is a random quantity in the data points $z_1,...,z_n$, we define the performance of the the estimator $\widehat{f}_n$ on the distribution ρ relatively to r_n as an expectation of the form $\mathbb{E}_{\rho^{\otimes n}}\left[\psi\left(\frac{\mathcal{R}(\widehat{f}_n)-\mathcal{R}_*}{r_n}\right)\right]$ where ψ is an non-decreasing function. Choices for ψ depend on the use case, but typical choices include $|\cdot|^p,\ p\geq 1$ to have bounds in L^p norm (Blanchard and Mücke, 2018) but also indicators of the form $\mathbf{1}_{t>\tau},\ \tau>0$ to have bounds in probability (Caponnetto and De Vito, 2007; Tsybakov, 2008). The performance of the best possibles estimator $\widehat{f}_n$ is therefore the quantity

$$\mathrm{MinMax}(\psi,\mathcal{H},\mathcal{M},n,r_n):=\inf_{\widehat{f}_n}\sup_{\rho\in\mathcal{M}}\mathbb{E}_{\rho^{\otimes n}}\left[\psi\left(\frac{\mathcal{R}(\widehat{f}_n)-\mathcal{R}_*}{r_n}\right)\right].$$

It is called the minimax performance at n for the rate r_n. We say that (r_n) is :

- a minimax upper rate if $\limsup_{n\to\infty}\mathrm{MinMax}(\psi,\mathcal{H},\mathcal{M},n,r_n)<\infty$;
- a minimax lower rate if $\liminf_{n\to\infty}\mathrm{MinMax}(\psi,\mathcal{H},\mathcal{M},n,r_n)>0$;
- a minimax optimal rate if it is both an upper and a lower rate.

In this thesis, we will mostly focus on upper bounds and in particular *non-asymptotic upper bounds*. Indeed, we will study specific estimators $\widehat{f}_n$, and typically show bounds of the form

$$\mathbb{P}_{\rho^{\otimes n}}\left(\frac{\mathcal{R}(\widehat{f}_n)-\mathcal{R}_*}{r_n}>\tau\right)\leq Ce^{-c\tau\alpha},$$

for a given loss function ℓ, a given set $\mathcal{H}$ of functions and a given model $\mathcal{M}$ of data distributions, and where C and c are explicit and do not depend on n or ρ but only on $\mathcal{M},\mathcal{H},\ell$. These bounds actually show that $(\mathcal{R}(\widehat{f}_n)-\mathcal{R}_*)/r_n$ has an exponential tail. Note that this type of bounds can actually lead to minimax upper rates (see for example Tsybakov (2008); Blanchard and Mücke (2018)). In general, the strategy to find an upper bound is usually to build an estimator which matches that bound. Such non-asymptotic upper bounds and rates can be found in every chapter of the thesis (except perhaps chapter 9); they will not always be formalized in this way, but the spirit is the same.

Finding lower bounds, on the other hand, is a different exercise altogether. As proposed by Tsybakov (2008), the strategy is usually, to build a finite sequence of densities $\rho_1,...,\rho_m$ which are far in terms of risk, but close statistically, so that $\widehat{f}_n$ will be unable to approach all of them well simultaneously. For slightly more details, see chapter 2.

In chapter 2, we will detail minimax optimal rates in the setting of least-squares regression (loss ℓ) and linear methods (model $\mathcal{H}$) for different classes of measures $\mathcal{M}$.

Concluding remarks on surrogate problems. In this section Sec. 1.1.2, we have seen that solving an empirical risk minimization problem can be a good way to find a good estimator while having only access to n samples from the data distribution. However, the excess risk of an empirical risk minimizer depends on the size of $\mathcal{H}$, and is caracterized by a trade-off between the approximation error, *i.e.*, the capacity of $\mathcal{H}$ to approximate all functions of interest, which increases with its size, and the estimation error *i.e.*, the capacity of the empirical risk minimizer to generalize well enough. We have also seen that regularization can be a good way of controlling the capacity of $\mathcal{H}$ which does not require changing the space, but rather penalizing functions already in the space. Finally, we have developed the two main ways of bounding the estimation error, and explained how the results obtained in this thesis, non-asymptotic upper rates, compare with the minimax point of view.

Note that the framework presented here does not contain all forms of machine learning methods or losses. We mentioned two examples, but there are of course many others. First, it happens that the regularization term in e.r.m. is data-dependent. This is the case for example in ranking problems. This leads to what is called *non-decomposable losses*, that is losses which cannot be written as $\sum_{i=1}^n \ell_{z_i}(f)+\Omega(\theta)$ *i.e.*, as a finite sum where each term depends only on one data point plus a penalty, but with sums involving cross terms of the data (Kar, Narasimhan, and Jain, 2014). Second, while the framework for studying e.r.m. here relies on the definition of a space $\mathcal{H}$ of candidate functions, it happens that one uses this space $\mathcal{H}$ to build "learners" (as in random forests, see Hastie, Tibshirani, and Friedman (2001)), and then aggregates these learners into a predictor, which is therefore not in $\mathcal{H}$ anymore. This is called *improper estimation* (see Hastie, Tibshirani, and Friedman (2001); for the definition of such estimator on the logistic regression problem, which we treat in part I, see Mourtada and Gaïffas (2022)).

Finally, the classical approximation/estimation trade-off point of view is sometimes not enough to explain the whole story. Indeed, certain over-parametrized models generalize well, *i.e.*, yield good estimators, even when they interpolate the data. This phenomenon is called *double descent*,

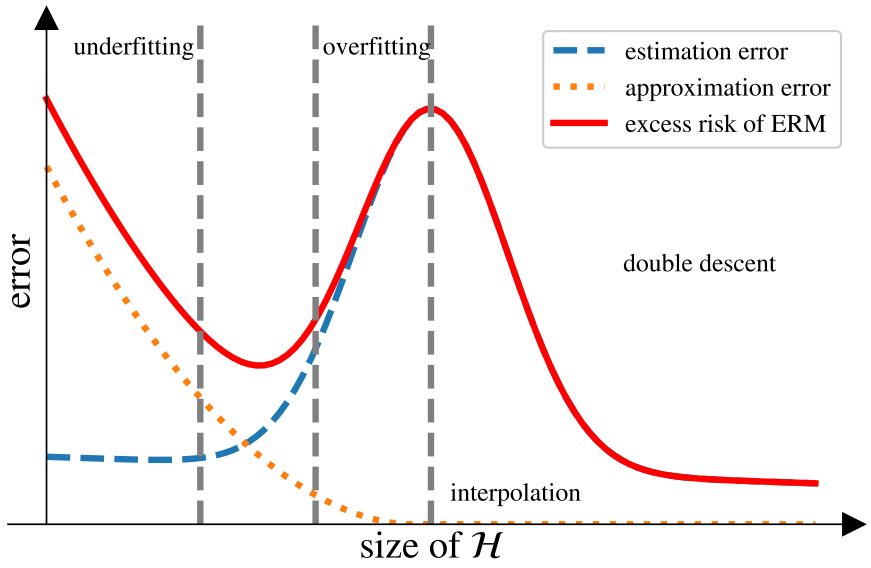

Figure 1.5: Double descent : where the estimation error decreases after interpolation.

and is graphically shown in Fig. 1.5. Recently, theory has made progress in understanding certain situations where double descent does or does not happen. For more details on this subject, one can refer to Belkin, Hsu, Ma, and Mandal (2019); Mei and Montanari (2019).

In all this discussion, we have yet to address a key issue : how do we effectively solve an empirical risk minimization problem such as Eqs. (1.7) and (1.9) ? This is the role of optimization methods, *i.e.*, the development of algorithms to optimize functions. In the next section, we will give a quick introduction to the different optimization settings and methods we will consider in this thesis, and which are used in machine learning in general.

1.1.3 Optimization

Recall that we have started from an ideal problem such as expected risk minimization in Eq. (1.4), which we then transformed into a surrogate problem such as the empirical risk minimization problem Eqs. (1.7) and (1.9) due to the fact that we have only partial access to the data distribution. However, we have yet to address the actual minimization of Eqs. (1.7) and (1.9). This is the role of optimization algorithms. In this section, we adopt the standard notations of optimization; f will not denote the target predictor but the function to optimize, and d will denote the dimension of the set on which we optimize and not the dimension of the input (data) space.

Very generally, an optimization algorithm aims at finding an approximate solution to $\inf_{\theta\in\Theta} f(\theta)$, where $\Theta \subset \mathbb{R}^d$ is a set which can be encoded in a computer. Optimization is a vast field which reflects the wide variety of problems which it can handle. Optimization algorithms may vary according to the way we access the function f, its structure, its regularity or convexity properties, the structure of the set Θ (un-constrained if $\Theta = \mathbb{R}^d$), the precision we want for the returned solution of the algorithm, the time and space complexity available, the means of computation (GPU versus CPU).

The goal of this section is to provide an overview of optimization techniques used or mentioned in this thesis in parts I and II. We will place ourselves in the setting where the function f is a sum of a decomposable function and a regularization term (of the type Eq. (1.9)), and is convex. Hence, unless stated otherwise, we will assume that the optimization problem we want to solve is of the form

$$\min_{\theta \in \Theta} f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta) + \lambda \Omega(\theta), \tag{1.10}$$

where Θ is a convex set of a vector space and the ℓ_i and Ω are convex in θ. In most cases, we will have $\ell_i = \ell_{z_i}$ coming from an empirical problem. The constraint set Θ will be either the whole of $\mathbb{R}^d$ or the positive semidefinite cone of symmetric matrices $A \in \mathbb{R}^{d \times d}$ subject to $A \succeq 0$ (*i.e.*, A is symmetric and $x^\top A x \geq 0$ for all x). Some linear equality constraints will sometimes be added in part II and part III when dealing with probability densities, *i.e.*, a constraint of the form $\theta^\top a = b$.

We will assume that we have access to the ℓ_i, their gradients and sometimes their second order derivatives. We will usually assume easy access to the gradients $\nabla \ell_i(\theta)$, with cost $O(d)$ in time. Note that this means the access to a full gradient will cost at least $O(nd)$ in time. The penalty Ω will be discussed in each case, but we will mostly use the classical square norm penalty $\Omega(\theta) = \frac{1}{2}\|\theta\|^2$, the gradient of which is simply θ and hence is directly accessible, and the Hessian of which is the identity matrix.

Machine learning setting. In the machine learning setting, Bottou and Bousquet (2008) have summarized the specific optimization requirements, which are a bit different than those of the classical optimization community (although since then, the communities have become closer). They highlight two specificities of machine learning problems.

(i) Since the function we wish to minimize is the "ideal" excess risk Eq. (1.4) and not the surrogate Eq. (1.9), we do not need to optimize the surrogate with a better precision than the approximation error. This differs from the classical optimization setting, where the goal is often to have a very precise approximation of the minimum. Thus, using a "bad" optimization may very well do the trick, as it will reach the precision of the approximation error, which is all that is needed for our purposes.

(ii) Complexity matters : indeed, using a rough but fast algorithm in terms of computations per iteration (like a first order method) is likely to yield better results than a precise but slow algorithm. This is due to the fact that since the dimension d and the number of data points n are often huge (and hence the complexity of computing function values, gradients, Hessians), one must limit the number of such computations and leverage hardware architectures, namely GPUs, which perform small computations in a fast parallel way (as opposed to CPU, which can perform large computations such as full gradients of f in Eq. (1.10), but only sequentially).

We will apply the first principle in order to design optimization methods whose goal is to achieve an error of the same order as that of the approximation error, in particular in part I, in chapter 4. The second principle has often been interpreted as a way of saying that first order methods, that is methods relying only on gradient or stochastic gradients (for more details, see below), are the best methods for machine learning. While this is often true, we develop a different approach in chapter 4, based on second order information (that is the Hessian of F), showing that there can indeed be efficient second-order methods for machine learning in certain contexts.

First order methods : the optmization workhorse in machine learning and dependence on the condition number. Here, we briefly present the basic first order optimization techniques. These methods will be a strong baseline for our algoritms developed in part I. In particular, we want to highlight that the speed of convergence of these algorithms depends strongly on the so-called *condition number* of the problem, defined below. In part I, we will argue that certain classical machine learning problems have very high condition number, and that first order methods are not necessarily the best solution in that setting. Moreover, we will implement techniques to reduce the condition number of linear systems, after which one of these first order methods can be used.

Assume here that the regularization is $\Omega(\theta) = \frac{1}{2}\|\theta\|^2$, and that the optimization problem is unconstrained, *i.e.*, $\Theta = \mathbb{R}^d$. Note that we can write f as a finite sum in the form

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta), \qquad f_i(\theta) = \ell_i(\theta) + \frac{\lambda}{2}\|\theta\|^2, \tag{1.11}$$

and that the cost in time and memory of computing a gradient of f_i is still $O(d)$. A **first order method** is a method which relies only on gradient computations; they are the most common methods in machine learning. In particular, the fact that gradients can be computed as fast as function values in a neural network (Hastie, Tibshirani, and Friedman, 2001; Paszke, Gross, Chintala, Chanan, Yang, DeVito, Lin, Desmaison, Antiga, and Lerer, 2017) makes them methods of choice in modern machine learning.

The most well-known method in optimization is *gradient descent* (GD) : the idea is simply to evaluate the gradient at time t and to go in the opposite direction $-\gamma_t \nabla f(\theta_t)$ (the decreasing direction) with a certain stepsize $\gamma_t > 0$ which quantifies how conservative we are with respect to that gradient information (we can take a large stepsize if we know the direction is stable). Given a starting point θ_0, gradient descent simply implements the recursion

$$\theta_{t+1} = \theta_t - \gamma_t \nabla f(\theta_t), \qquad \gamma_t > 0. \tag{1.12}$$

Note that the cost of each iteration here is a priori $O(nd)$ in time and $O(d)$ in memory. In machine learning, as n and d are very large, the computation of a full gradient is quite expensive. To mitigate this problem, another first order algorithm which is used is *stochastic gradient descent* (SGD), *i.e.*,

$$\theta_{t+1} = \theta_t - \gamma_t \nabla f_{i_t}(\theta), \qquad \gamma_t > 0, \tag{1.13}$$

where i_t is selected uniformly at random in $\{1, ..., n\}$ at each iteration. The cost of an iteration here is therefore only $O(d)$ and not $O(nd)$ since we only access one gradient of the f_i, which we use as a gross approximation of the full gradient. When comparing GD to SGD, one therefore has to keep in mind that n iterations of SGD (called one *epoch*) is equivalent to one iteration of GD in term of complexity.

SGD is actually part of a much broader class of method which date back to the 50s, (Robbins and Monro, 1951). A stochastic gradient descent can be implemented as soon as we have an unbiased estimator of the gradient at each timestep. Here we just use the finite sum structure to see it as an expectation and artificially give it a random structure. Note however that in the statistical learning framework, the f_i are random, as they come from i.i.d. samples $z_i \sim Z$,

and that stochastic gradient descent taking $i_t = t$ for n iterations is actually stochastic gradient descent performed on the expected risk minimization problem Eq. (1.4).

The analysis of these methods can be done in many settings, but one is of particular interest to us in this thesis. To do so, we need to make some assumptions on the functions f_i. We say that a function g defined on a convex domain Θ is L-smooth if its gradient is L-Lipschitz, *i.e.*, if

$$\forall(\theta, \theta') \in \Theta^2, \ \|\nabla g(\theta) - \nabla g(\theta')\| \leq L\|\theta - \theta'\|. \tag{1.14}$$

We say that g is μ-strongly-convex if

$$\forall(\theta, \theta') \in \Theta^2, \ \langle \nabla g(\theta) - \nabla g(\theta'), \theta - \theta' \rangle \geq \mu\|\theta - \theta'\|^2. \tag{1.15}$$

We denote with $\mathcal{F}_{L,\mu}$ the class of functions which are L-smooth and mu-strongly-convex. Informally, being L-smooth means being locally upper bounded by a parabola of magnitude L everywhere, while being mu-strongly-convex means being lower bounded by a parabola of magnitude μ everywhere. The prototypical example of a function $f \in \mathcal{F}_{L,\mu}$ is a quadratic of the form

$$f(\theta) = \tfrac{1}{2}\theta^\top \Sigma \theta - b^\top \theta + \text{constant}, \qquad \Sigma^\top = \Sigma, \mu\mathbf{I} \preceq \Sigma \preceq L\mathbf{I}, \tag{1.16}$$

whose minimization is equivalent to finding $\theta_* = \Sigma^{-1}b$. Note that this form of function naturally appears when solving the least-squares problem with a linear model, as we will see in part I and chapter 2, and that first order methods allow to bypass the inversion operation which is very costly when d is large. Note that if the ℓ_i in Eq. (1.10) are all L-smooth, then the f_i all belong to $\mathcal{F}_{L+\lambda,\lambda}$. Moreover, if the f_i belong to $\mathcal{F}_{L,\mu}$, then $F = \frac{1}{n}\sum_{i=1}^n f_i$ also belongs to $\mathcal{F}_{L,\mu}$. One of the key quantities which characterizes how these first-order optimization methods perform is the so-called *condition number*, usually defined as the ratio $\kappa = \frac{L}{\mu}$. Note that $\kappa > 1$ as $\mu \leq L$. In the case of a quadratic Eq. (1.16), the condition number (also denoted with $\kappa_2(\Sigma)$ in the literature) is simply the ratio between the largest and smallest eigenvalues of Σ.

Assuming the f_i are in $\mathcal{F}_{L,\mu}$, the best known algorithms in term of convergence are more involved than the simple GD/SGD updates Eq. (1.12) and Eq. (1.13). In the case of gradient descent, Nesterov (1983) proposes an optimal acceleration (*i.e.*, an acceleration of gradient descent) for the class $\mathcal{F}_{L,\mu}$, where θ_t is updated using a two-step procedure with a so-called "momentum step". Linear convergence (that is showing that $f(\theta_t) - f(\theta_*) \leq \rho^t(f(\theta_0) - f(\theta_*))$ for $\rho < 1$) can be obtained for standard and Nesterov accelerated gradient descent when assuming that $f \in \mathcal{F}_{L,\mu}$. Moreover, in the case where f is a quadratic Eq. (1.16), that is when solving a linear system with a positive definite matrix Σ, there exists a more elaborate first order iterative method named *conjugate gradient* (see chapter 10 of (Golub and Van Loan, 2012) and Theorem 10.2.6), which reaches the same accelerated convergence bounds as that of Nesterov but in a more efficient way. This method will be an important sub-routine in chapter 4.

Proposition 1.1 (Rates of convergence of deterministic gradient descents)**.** *If $f \in \mathcal{F}_{L,\mu}$ and reaches its minimum at θ_*, and $\gamma_t = \gamma := \frac{1}{L}$, the sequence of gradient descent iterates Eq. (1.12) initialized at θ_0 satisfies :*

$$\forall t \in \mathbb{N}, \ f(\theta_t) - f(\theta_*) \leq (1 - 1/\kappa)^t \left(f(\theta_0) - f(\theta_*)\right) \tag{1.17}$$

This means that achieving an error ratio of order ϵ can be done with $t = O(\kappa \log \frac{1}{\varepsilon})$ iterations, i.e., a complexity of order $O(nd\kappa \log \frac{1}{\varepsilon})$ in our context of minimizing Eq. (1.10). Using instead

Nesterov acceleration (see Nesterov (1983)), a better convergence rate (where essentially κ is replaced by $\sqrt{\kappa}$) can be obtained :

$$\forall t \in \mathbb{N},\ f(\theta_t) - f(\theta_*) \leq \frac{1+\kappa}{2}\left(1 - 1/\sqrt{\kappa}\right)^t (f(\theta_0) - f(\theta_*)) \tag{1.18}$$

This means that achieving an error ratio of order ϵ can be done with $t = O(\sqrt{\kappa}\log\frac{\kappa}{\varepsilon})$ iterations, i.e., a complexity of order $O(nd\sqrt{\kappa}\log\frac{\kappa}{\varepsilon})$ in our context of minimizing Eq. (1.10). Finally, when the function f is a quadratic function, the conjugate gradient method iterates θ_t (which are constructed only from function evaluations and gradients of f) satisfy

$$\forall t \in \mathbb{N},\ f(\theta_t) - f(\theta_*) \leq 4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} (f(\theta_0) - f(\theta_*)). \tag{1.19}$$

For stochastic gradient descent, the situation is a bit more complex; proving linear convergence in the strong convexity and smooth setting proves to be hard, and such rates are proved for algorithms which are more involved such as SAG (Roux, Schmidt, and Bach, 2012), SVRG (Johnson and Zhang, 2013) and SAGA (Defazio, Bach, and Lacoste-Julien, 2014). Evaluating the error of θ_t is usually done through a linear combination of $\frac{1}{\mu}\left(\mathbb{E}\left[f(\theta_t)\right] - f(\theta_*)\right)$ and $\mathbb{E}\left[\|\theta_t - \theta_*\|^2\right]$ (instead of simply $f(\theta_t) - f(\theta_*)$ in the "deterministic" setting, since θ_t is now random). For these methods, linear rates can be achieved : $O((n+\kappa)\log\frac{n}{\varepsilon})$ stochastic gradient iterations are necessary to reach an error ratio of order ϵ. Acceleration methods exist in this setting as well, such as Allen-Zhu (2017); Zhou, Shang, and Cheng (2018) for SVRG and Defazio (2016); Zhou, Ding, Shang, Cheng, Li, and Luo (2019) for SAGA. They lead to a a complexity of order $O((n+\sqrt{n\kappa})\log\frac{1}{\epsilon})$ in terms of stochastic gradient iterations, which leads to a total time complexity of order $O(d(n+\sqrt{n\kappa})\log\frac{1}{\epsilon})$.

We see that in spirit, the fastest first order methods to solve Eq. (1.10) under the smooth and strongly convex assumptions are accelerated stochastic methods, which achieve an error ratio of order ε in time $O(d(n+\sqrt{n\kappa})\log\frac{1}{\varepsilon})$. We see that this error crucially depends on κ when the problem is ill-conditioned, *i.e.*, when the regularization λ is small. This will motivate the introduction of a new second order algorithm in chapter 4 in part I.

Second order methods : suited for machine learning ? In this thesis, we will have need of second order methods in two quite different settings. The first, in chapter 4, will be to produce a machine learning method which is not limited by the condition number, as in the case of first order methods. The second, in chapters 6 and 8 will be when dealing with problems with conic constraints, where Θ is a positive semidefinite cone of matrices. In this paragraph, we briefly introduce these methods and explain both their promises and why they are less popular in the machine learning setting. In the next paragraph, we will briefly mention how these methods are used to deal with conic constraints.

The dependence of first order methods on the condition number is due to the fact that we do not take into account that gradients can evolve differently in different directions (*i.e.*, the method does not adapt to the curvature of f). Second order methods aim to solve this problem. The prototypical second order algorithm is called Newton method, and follows the update rule of the type

$$\theta_{t+1} = \theta_t - \gamma_t \Delta_t, \qquad \Delta_t = \nabla^2 f(\theta_t)^{-1} \nabla f(\theta_t), \tag{1.20}$$

where $\gamma_t > 0$ is a stepsize and Δ_t is called the Newton step, a renormalized gradient step. Informally, in directions where the gradient h is stable, $h^\top \nabla^2 f(\theta) h$ will be small, and the newton method will renormalize that direction as if multiplying by large stepsize ($\nabla^2 f(\theta)^{-1} h$ will be large). On the other hand, in directions where the gradient changes quickly, the Newton method will renormalize that direction as if multplying by a small stepsize. This is illustrated by the fact that in the case of a quadratic Eq. (1.16), the Newton method converges in one step with $\gamma_0 = 1$.

Under certain assumptions on the function f, *i.e.*, *self-concordance* (see Nesterov (2018); Boyd and Vandenberghe (2004) and Sec. 2.2 for details), Newton methods can be shown to have a two regimes convergence : a first slow regime, where the loss typically decreases by a constant at every iteration, and a second "quadratic" regime the convergence is quadratic, *i.e.*, of the form $\rho^{\rho^{-t}}$ for $\rho < 1$. The number of iterations needed to get a relative error is therefore $O(\log\log \frac{1}{\varepsilon} + C)$ where C is a constant linked to the first regime. Crucially, in the case of self-concordant functions, C does not depend on the conditioning of the problem and is a fixed constant. In other words, the quadratic convergence happens in a large region around the minimum.

However, in machine learning, and in particular in the context of finite sums in Eq. (1.10), it is almost impossible to know if the target function satisfies the self-concordant property or not (the expectation of self-concordant functions is not self-concordant in general). The classical analysis of Newton methods for such problems rely on other assumptions, and in particular on the *generalized self-concordant* assumption (Bach, 2010) (see Sec. 2.2 for precise definitions). However, these methods and analysis suffer from two key issues.

- The superlinear convergence happens in a small region around the optimum of order $\|\theta - \theta_*\|^2 \leq \lambda$. But getting in that region can take as many steps as that of gradient descent a priori (thus depending on the condition number). Moreover, in machine learning, a great precision is often not needed as it can be the case in classical optimization (recall the discussion at the beginning of Sec. 1.1.3 from Bottou and Bousquet (2008)), and often reaching precision λ is already enough.

- Computing one Newton step is computationally expensive : indeed, the solving of the linear problem; the computing of the Hessian at a given point usually takes time of order $O(nd^2)$, and computing its inverse takes time $O(d^3)$. This is prohibitive both in terms of storage capacity and of time complexity.

To summarize, Newton methods seem to miss the mark in machine learning : they have a large per-iteration complexity, and are fast only after reaching a certain precision which is already good enough for machine learning. However, in part I and in particular in chapter 4, we show that *a)* we can leverage the fast convergence globally, and *b)* that we can drastically reduce the cost of iterations using the structure of Eq. (1.10) and pre-conditioning techniques.

Note that there exists other second-order type methods named Quasi-Newton methods, such as BFGS, which consist in sketching the inverse Hessian little by little at each step rather than computing the entire Hessian and inverting it. However, they are not studied in this thesis. Details on those methods can be found in Nocedal and Wright (2006).

Handling non-smoothness with proximal methods. In part II and in particular in chapter 5, we study problems for which the methods described above, relying on smoothness, cannot be readily applied.

Problem 1. The first is a problem coming from the solving of Eq. (1.10) in the case of positive semidefinite matrices, *i.e.*,

$$\min_{A\in\mathbb{R}^{d\times d},\ A\succeq 0} \sum_{i=1}^{n} \ell_i(A) + n\lambda\Omega(A) \tag{1.21}$$

where the regularization $\Omega(A)$ is a so-called p-schatten norm, that is $\Omega(A) = \|\sigma_i(A)\|_p^p = \sum_{i=1}^d |\sigma_i(A)|^p$ where $\sigma_i(A)$ denotes the eigenvalues of A counted with multiplicity. In particular, when $p = 1$, Ω is the nuclear norm, which is non-smooth. In this first problem, we assume the ℓ_i are L-smooth, and hence the problem is of the form $\min_{A\in\mathbb{S}_d(\mathbb{R})} \mathcal{L}(A) + \lambda n\Omega_+(A)$ where $\mathcal{L}$ is L-smooth and $\Omega_+(A) = \iota_{A\succeq 0}\Omega(A)$ is non smooth.

Problem 2. The second problem comes from a dual formulation of Eq. (1.21) in the case where the penalty $\Omega(A)$ is simply the Frobenius norm $\|A\|_F^2 = \mathrm{Tr}(AA^\top)$. In this case, the dual problem has the form $\sup_{\alpha\in\mathbb{R}^n} -\sum_{i=1}^n l_i^*(\alpha) - \frac{1}{2\lambda n}\Omega^*(\alpha)$ where Ω^* is smooth and where the l_i^* are Fenchel conjugates (see Boyd and Vandenberghe (2004)) of functions which can be logarithms, and hence are not smooth (for more details, see chapter 5, as well as Parikh and Boyd (2014)). Thus, the problem is equivalent to a problem of the form $\min_{\alpha\in\mathbb{R}^n} \mathcal{L}^*(\alpha) + \Omega^*(\alpha)$, where Ω^* is smooth and $\mathcal{L}^*$ is non smooth.

In both these cases, we aim to minimize the function $f(\theta) = g(\theta)+h(\theta)$ where h is not smooth and not even differentiable, and g is L-smooth. There are first order algorithm analog to (accelerated) gradient descent algorithms to solve these problems, called proximal algorithms. These algorithms are based on two basic requirements :

- we have access to gradients $\nabla g(\theta)$;
- we are able to easily compute the *proximal operator* of λh for all $\lambda > 0$ at every point θ_0:

$$\forall\theta_0 \in \mathbb{R}^d, \mathrm{prox}_{\lambda h}(x) = \arg\min_{\theta\in\mathbb{R}^d} h(\theta) + \tfrac{1}{2\lambda}\|\theta - \theta_0\|^2. \tag{1.22}$$

Proximal operators are readily available in many cases (Bach, Jenatton, Mairal, and Obozinski, 2011; Parikh and Boyd, 2014), and in particular in those we will consider in chapter 5. That is the case for instance for the ℓ_1 norm $\|\theta\|_1 = \sum_{i=1}^d |\theta_i|$, where the proximal operator is called the soft thresholding operator :

$$\mathrm{prox}_{\lambda\|\cdot\|_1}(\theta_0) = \mathrm{sgn}(\theta_0) \times \max(|\theta_0| - \lambda, 0),$$

where all operations are considered coordinate-wise. The main idea behind proximal methods is that if one is at a point θ_0, then by L smoothness, we have an upper bound

$$f(\theta) + g(\theta) \leq f(\theta_0) + \langle\theta - \theta_0, \nabla f(\theta_0)\rangle + \tfrac{L}{2}\|\theta - \theta_0\|^2 + g(\theta). \tag{1.23}$$

One can minimize this upper bound in closed form : indeed, if we denote by $p_L(\theta_0)$ this minimizer, we have $p_L(\theta_0) = \mathrm{prox}_{h/L}(\theta_0 - \frac{1}{L}\nabla g(\theta_0))$. We see that $p_L(\theta_0)$ is very close to a gradient descent update Eq. (1.12) with step-size $1/L$. In the case where $h = 0$, this is in fact exactly gradient descent. In the case where f is L-smooth, proximal methods therefore extend the rates of convergence of gradient descent for L-smooth functions to composite losses with the addition of a potentially non-smooth term in the loss. The equivalent algorithm to gradient descent is dubbed ISTA (iterative shrinkage threshold algorithm) : it is defined through the recursion $\theta_{t+1} = p_L(\theta_t)$

and satisfies $f(\theta_t) - f(\theta_*) \leq \frac{L\|\theta_0 - \theta_*\|}{2t}$ (see Theorem 3.1 in Beck and Teboulle (2009)). Beck and Teboulle (2009) also develop the equivalent of acceleration in the proximal setting, dubbed Fast ISTA (or FISTA), which satisfies $f(\theta_t) - f(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|}{(t+1)^2}$ (see Theorem 4.4). It is this method we will use in chapter 5. Note that one cannot leverage strong convexity of f to have linear convergence : these methods will not be adapted to the classical machine learning setting, but rather for more moderate sizes of n and d (as the number of iterations needed will be too high to run for large scale, large dimensional problems).

Handling constraints and interior point methods. For now, we have presented methods for unconstrained convex optimization problems, that is we have always assumed that Eq. (1.10) is solved on the whole of $\mathbb{R}^d$. However, in parts II and III, we use models which are parametrized by a positive semidefinite matrix, and sometimes even impose constraints that certain functions sum to one in order to recover probability densities. The goal of this section is to very briefly present the spirit of such methods, in particular in the case of semidefinite programs (SDPs). For a more complete introduction, we refer to Boyd and Vandenberghe (2004), chapters 10 and 11, as well as to Nesterov and Nemirovskii (1994). Moreover, note that linear equality constraints of the form $A\theta = b$ can often easily be dealt with by either *a)* going to the dual, where the constraints disappear, *b)* using a projected method like projected gradient descent, which projects each gradient step on the affine space defined by the linear equalities and *c)* incorporating them in a Newton method which can be easily adapted to equality constraints. In the few cases where we have to deal with equality constraints in this thesis (mainly in chapters 5 and 8), these are not hard to handle, in the sense that they do not add to the complexity of the problem. The constraints which are hard to deal with in this thesis are the PSD constraints, *i.e.*, handling problems of the form $\min_{A \succeq 0} f(A)$ where A is a $d \times d$ PSD matrix. Note that we do not treat feasability conditions here as we will always be able to exhibit a feasible solution.

Interior point methods. Interior point methods deal with problems of the form $\min_{\theta \succeq 0} f(\theta)$ by adding a self-concordant barrier, that is a function $\frac{1}{t}\phi(\theta)$ where ϕ has domain $\theta \succ 0$ (that is ϕ explodes at the boundary). In a sense, adding this barrier forces the problem to stay in the set $\theta \succeq 0$. The idea is then to construct a sequence of solutions $\theta(t_i)$ for increasing values of t_i in order to approximate the real problem.

In the context of solving semidefinite programs, of the form $\min f(A)$ subject to $A \in \mathbb{S}(\mathbb{R}^p),\ A \succeq 0$ (it is possible to add additional equality constraints), the classical logarithmic barrier used is the log determinant barrier, *i.e.*, $\phi(A) = -\log\det(A)$, which is self-concordant and the domain of which is the set of positive definite matrices. The idea of interior point methods is to iteratively solve the problem

$$A(t) := \arg\min f(A) - \tfrac{1}{t}\log\det(A), \tag{1.24}$$

using a Newton method for a finite sequence $t_1, ..., t_K$ for increasing values t_i. Indeed, the idea is that as t increases, $A(t) \underset{t \to +\infty}{\longrightarrow} A_*$, where A_* denotes an optimal solution. The sequence $A_i := A(t_i)$ is called the central path, and the method is called interior point method because by definition, A_i is a strictly feasible point (*i.e.*, it lies in $A_i \succ 0$ and satisfies the equality constraints if there are some). Under a certain self-concordant hypothesis on f (which will be satisfied when handling linear or quadratic objectives f in parts II and III), as well as a strict feasibility condition, it can be shown that to achieve precision ϵ, one roughly needs $K = \sqrt{d}\log\frac{1}{\varepsilon}$ interior points (Nesterov and Nemirovskii, 1994; Tuncel, 2000).

1.1.4 Special case of global optimization

In this section, we briefly present the setting of part III in terms of ideal, surrogate problems and optimization algorithm. In part III we present methods which deal with global optimization of functions with SDPs. Such methods have been developed for a long time for polynomials (see for example Lasserre (2010) and all the references in chapter 7), and we develop such methods for regular functions. Here, we present the very general setting and principle of such methods in the unconstrained optimization case, while keeping in mind two cases which we will see in part III.

Polynomial case. In this case, we consider the optimization of a polynomial function f of degree at most $2r$ over $\mathbb{R}^d$. We denote with $\mathbb{R}_{2r}[x_1, ..., x_d]$ the set of such polynomial functions. We assume we have access to its coefficients f_α with $|\alpha| \leq 2r$, where $\alpha \in \mathbb{N}^d$ and $|\alpha| = \sum_{i=1}^d \alpha_i$. We define $R_n : f \in \mathbb{R}_{2r}[x_1, ..., x_d] \to \mathbb{R}^n$ where $n = \binom{d+2r}{d}$ the bijective linear map which identifies the polynomial function f to its coefficients $(f_\alpha)_{|\alpha|\leq 2r}$.

Regular function case In this case, we consider the optimization of a regular function f of class C^r on a domain U with access to n function values at points $x_1, ..., x_n$. We define $R_n : f \in C^r(U) \mapsto (f(x_1), ..., f(x_n)) \in \mathbb{R}^n$ the linear map which evaluates f at points $x_1, ..., x_n$.

Ideal problem. Let f be a function defined on a space $\mathcal{X}$ belonging to a vector space $\mathcal{F}$ containing constant functions. The global optimization problem $\min_{x\in\mathcal{X}} f(x)$ can be formulated as a convex optimization problem of the form

$$\begin{aligned} &\sup\ c \\ &\text{subject to } c \in \mathbb{R},\ g = f - c,\ g \geq 0, \end{aligned} \tag{1.25}$$

which is simply stating that the minimum of a function is the maximum lower bound of that function. This will be our ideal problem Eq. (1.1). In the polynomial case, $\mathcal{F} = \mathbb{R}_{2r}[x_1, ..., x_d]$ while in the regular function case, $\mathcal{F} = C^r(U)$.

Surrogate problems. The idea of methods presented in part III is to solve a surrogate problem for Eq. (1.25) by modelling the two constraints $g \geq 0$ and $g = f - c$ which are constraints which have to hold for all $x \in \mathcal{X}$ (and hence imply an infinite number of constraints for our two problems). We therefore proceed in two steps.

Step 1. We model the set of functions $g \geq 0$ with a finite-dimensional model of functions $G_+ = \{g_A\ :\ \forall x \in \mathcal{X},\ g_A(x) = \phi(x)^\top A\phi(x)\}$, parametrized by PSD matrix $A \in \mathbb{S}(\mathbb{R}^p),\ A \succeq 0$ where $\phi : x \in \mathcal{X} \to \mathbb{R}^p$ is a feature map. The PSD constraint on A guarantees that the functions g_A are non-negative. We can see the set G_+ as the set of sum of squares of functions of the form $h_v(x) = v^\top\phi(x)$ for $v \in \mathbb{R}^p$ (this is a direct application of the spectral theorem). In certain cases, we will use a penalty $\Omega(A)$ on the set G_+.

In the case of polynomials, we consider the map of monomials of degree less or equal to r : $\phi(x) = (x^\alpha)_{|\alpha|\leq r} \in \mathbb{R}^p$ for $p = \binom{r+d}{d}$. In this case, $g_A(x) = \sum_{|\alpha|,|\beta|\leq r} A_{\alpha,\beta}x^{\alpha+\beta}$ is a non-negative polynomial of degree at most $2r$ which is a sum of squares of polynomials of degree at most r (see chapter 7). In that case, no penalty Ω is considered.

In the case of regular functions, we will consider models of where the feature map ϕ is defined through a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$: $\phi(x) = (k(x, \tilde{x}_j))_{1 \leq j \leq p} \in \mathbb{R}^p$ (see chapters 5, 6 and 8). G_+ therefore represents the set of sums of squares of functions of the form $\sum_{j=1}^p \alpha_j k(x, \tilde{x}_j)$. We will also consider the penalty $\Omega(A) = \mathrm{Tr}(KA)$ on G_+, where K is a PSD matrix (for more details, see chapter 5.)

Step 2. Replacing with the model for non-negative functions g, the equality constraint becomes $f - c = g_A$. This equality also needs to hold pointwise for all $x \in \mathcal{X}$, which can be infinite.

In the case of polynomials, as f, c and g_A are polynomials of degree at most $2r$, this constraint is actually a n-dimensional constraint on the coefficients of the polynomials. Abstractly, it can be written $R_n(f - c) = \tilde{R}_n(A)$ where R_n is the map defined above which maps a polynomial function to its coefficients, and $\tilde{R}_n$ is the linear map which maps $A \in \mathbb{R}^{p \times p}$ to the coefficients of the polynomial function $g_A(x)$ which are also in $\mathbb{R}^n$.

In the case of regular functions, we have access to f only through n function evaluations. What we do in chapter 8 is the following. Instead of enforcing the constraint $g_A = f - c$ everywhere, we enforce it only at the points x_i. The constraint is therefore approximated by the n equality constraints $g_A(x_i) = f(x_i) - c$ for $1 \leq i \leq n$. Once again, this can abstractly be written in the form $R_n(f - c) = \tilde{R}_n(A)$ where R_n is the evaluation function defined above, and $\tilde{R}_n$ is the linear map $A \mapsto (g_A(x_i))_{1 \leq i \leq n}$.

Combining these two approximations of the constraints, we build the following surrogate problem.

$$\begin{aligned} &\sup\ c - \Omega(A) \\ &\text{subject to } R_n(f - c) = \tilde{R}_n(A),\ A \succeq 0,\ A \in \mathbb{S}(\mathbb{R}^p),\ c \in \mathbb{R}, \end{aligned} \tag{1.26}$$

Linking the result of Eq. (1.26) with the result of the ideal problem Eq. (1.25) (that is the minimum f_* of f) can be challenging. In the context of polynomials, it can be shown to be a lower bound (Lasserre, 2010). In the context of functions, certain rates can be obtained when the penalty is the trace norm and under certain additional conditions (see chapter 8). However, from an optimization standpoint, the surrogate problem can be solved using an interior point method as described above for the primal or the dual problem. In this thesis, we will study global optimization through SDPs.

Optimization. As Eq. (1.26) is a semidefinite program which is often linear or quadratic, we can readily apply interior point methods to solve the SDP. However, interior point methods are conditioned by the existence of a feasible point, which is not always the case for this problem (in particular, in the polynomial setting). That is why we usually consider its dual problem, called the moment problem in the context of polynomials, where it is often easier to find a feasible point.

$$\begin{aligned} &\inf\ \lambda^\top R_n(f) + \Omega^*(B - \tilde{R}_n^* \lambda) \\ &\text{subject to } \lambda \in \mathbb{R}^n,\ B \succeq 0,\ \lambda^\top R_n \mathbf{1} = 1, \end{aligned} \tag{1.27}$$

where Ω^* is the Fenchel conjugate of Ω. Again, this problem can readily be solved with interior point methods as soon as there is a feasible point, using $K = \sqrt{p} \log \frac{1}{\epsilon}$ interior points obtained

via the damped Newton method applied to Eq. (1.27). For more details on the polynomial setting, we refer to chapter 5 of Lasserre (2010). For methods based on function evaluations, we refer to chapter 8. The cost of performing the Newton method depends in particular on the fenchel conjugate of the penalty Ω^*, and is detailed in both these particular cases.

1.2 Reproducing kernel Hilbert spaces

In Secs. 1.1.2 and 1.1.4, we have seen that an important choice when building a surrogate problem is the choice of model $\mathcal{H}$ (or the model G_+ in the case of global optimization). For this section, let us take the machine learning notations introduced in Secs. 1.1.1 to 1.1.3, and consider the task of learning a real-valued function $f : \mathcal{X} \to \mathbb{R}$ on a space $\mathcal{X}$. Recall that the ideal requirements for $\mathcal{H}$ are that *a)* it must be large enough to approximate the "ideal" problem well (for example, if the expected risk $\mathcal{R}$ reaches its minimum at a certain f_*, we could require that f_* be in $\mathcal{H}$), *b)* it must be adapted to the data at hand, being small or regularized enough not to overfit, and *c)* it must lead to a (rapidly) solvable optmization problem.

Linear models. One of the most widely used classes of functions in machine learning and applied mathemathics is the class of parametric linear models, mentioned in the beginning of Sec. 1.1.1. A linear model is characterized by a feature map $\phi : \mathcal{X} \to \mathbb{R}^p$ which represents the space $\mathcal{X}$ through p features $\phi_1, ..., \phi_p$. The ϕ_i's can be coordinates of x, moments, results of certain filters or convolutions applied to a signal. This feature map can itself be learnt (this is typically the role of encoders in deep learning), or designed according to the problem at hand. The linear model associated to the feature map is the set of functions defined as :

$$\mathcal{H} = \Big\{ f_\theta : \mathcal{X} \to \mathbb{R} \mid f_\theta(x) = \langle \theta, \phi(x) \rangle_{\mathbb{R}^p} \Big\}. \tag{1.28}$$

These models have many advantages from a modelling perspective. They are usually interpretable (the coefficient associated to a feature being a measure of its importance), they inherit the Euclidean structure of $\mathbb{R}^p$ (as well as any other normed structure defined on $\mathbb{R}^p$), and preserve the convexity of convex loss functions, allowing to leverage the entire set of optimization techniques in that setting as described in Sec. 1.1.3.

However, these models are parametric, and therefore are quite limited : the dimension p is finite, and they cannot adapt to any number of data points (indeed, one has to consider a larger model as the number of data points gets larger). Moreover, the choice of the feature map is sometimes not obvious, especially in cases where there is little *a priori* knowledge on the function f (take for example the problem of learning a function f where the only information we have is that it is regular).

In this section, we will present a type of model for functions called reproducing kernel Hilbert spaces (RKHS). This model generalizes linear models to potentially infinite dimensional spaces, leading to better approximation properties, and is focused not on the design of the feature map but rather on that of the kernel, which is a measure of similarity between points in $\mathcal{X}$. Moreover, the nice properties of linear models still hold, making this type of model a very interesting tool for machine learning and applied mathematics in general.

This section will be organized as follows. In Sec. 1.2.1, we will define what a RKHS is, and provide different points of view on that definition. In Sec. 1.2.2, we will provide examples of kernels and RKHS, some of which will be used in this thesis, and illustrate the strength of the kernel approach for approximation. In Sec. 1.2.3, we show how RKHSs are particularly adapted in the

empirical risk minimization setting, and how to design finite dimensional surrogate problems for the expected risk minimization problem even when the underlying space is infinite dimensional. In Sec. 1.2.4, we will discuss the statistical properties of kernel methods. Finally, in Sec. 1.2.5 we will briefly discuss the problems and perspectives of effectively solving the optimization problem (more details will be given in part I).

1.2.1 Definition and constructions of reproducing kernel Hilbert spaces

In this section, we define the different fundamental objects in relation to reproducing kernel Hilbert spaces (RKHS), through three different points of view. Let us formally define a reproducing kernel of a Hilbert space (Aronszajn, 1950; Scholkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004).

Definition 1.1 (reproducing kernel)**.** *Let $\mathcal{H}$ be a Hilbert separable space of real-valued functions defined on a set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **reproducing kernel** of $\mathcal{H}$ if :*

- *the functions $k_x : x' \in \mathcal{X} \mapsto k(x, x')$ all belong to $\mathcal{H}$;*
- *the **reproducing property** is satisfied with the k_x:*

$$\forall x \in \mathcal{X},\ \forall f \in \mathcal{H},\ f(x) = \langle f, k_x \rangle_{\mathcal{H}} \tag{1.29}$$

If a Hilbert space of functions has a reproducing kernel, then that reproducing kernel can be shown to be unique (given $\mathcal{H}$, we can talk about *its* reproducing kernel, if it has one). Conversely, if $\mathcal{H}$ and $\mathcal{H}'$ have the same reproducing kernel k, then they are equal. Thus, if $\mathcal{H}$ exists, $\mathcal{H}$ is called *the* reproducing kernel Hilbert space with kernel k. We now present three constructions of RKHS, starting from different perspectives. The most important in machine learning is arguably the second one.

Construction through a feature map. The first way to build a RKHS is to take the point of view of linear models and start from a feature map. Assume that we are given a feature map $\phi : \mathcal{X} \to H$ which represents $\mathcal{X}$ into a Hilbert space. As in the linear model case, we can define the associated space of functions :

$$\mathcal{H} := \Big\{ f : \mathcal{X} \to \mathbb{R} \ :\ \exists \theta \in H,\ f(x) = \langle \theta, \phi(x) \rangle_H \Big\}. \tag{1.30}$$

We are now going to show that $\mathcal{H}$ is indeed a reproducing kernel Hilbert space, with kernel $k(x, x') := \langle \phi(x), \phi(x') \rangle_H$. Note that this kernel is *positive definite* (p.d.), that is *a)* symmetric, *i.e.*, $k(x, x') = k(x', x)$, and *b)* for any $n \in \mathbb{N}$ and any $(x_i) \in \mathcal{X}^n$, the matrix $(k(x_i, x_j))_{1 \le i,j \le n}$ is positive semidefinite (this comes from the fact that this matrix is a Gram matrix). $k(x, x')$ is the natural measure of similarity between points implied by the feature map. We still denote with f_θ the functions $\langle \theta, \phi(x) \rangle_H$.

Let $H_0 := \{\theta \in H \ :\ \forall x \in \mathcal{X},\ \langle \theta, \phi(x) \rangle_H = 0\} = \{\theta \in H \ :\ f_\theta = 0\}$ be the set of vectors $\theta \in H$ which represent the function $f = 0$, so that $f_\theta = f_{\theta'}$ i.i.f. $\theta - \theta' \in H_0$ (in a sense, H_0 represent the useless θ in terms of function evaluations). The there is a natural linear isomorphism between the orthogonal of H_0 and $\mathcal{H}$. Note that H_0 can itself be seen as the orthogonal of the set $\{\phi(x) \ :\ x \in \mathcal{X}\}$ and that the orthogonal of H_0 is therefore $H_\phi := \overline{\text{span}(\{\phi(x) \ :\ x \in \mathcal{X}\})}$. Thus, there is a natural isomorphism between H_ϕ and $\mathcal{H}$, and $\mathcal{H}$ can therefore be naturally equipped with the Hilbert structure of H_ϕ. The fact that k is the reproducing kernel of $\mathcal{H}$ then follows.

Note that the isomorphism between $\mathcal{H}$ and H_ϕ directly shows that $\mathcal{H} = \overline{\text{span}(\{k_x \ : \ x \in \mathcal{X}\})}$. In fact, the map $x \in \mathcal{X} \mapsto k_x \in \mathcal{H}$ is a way of representing $\mathcal{X}$ in the Hilbert space $\mathcal{H}$ (it is a feature map). The reason for which we distinguish $\mathcal{H}$ and H is that the space $\mathcal{H}$ is a space of functions on $\mathcal{X}$ while H is not (it can be larger). Indeed, any Hilbert space on which an embedding $\phi : \mathcal{X} \to H$ exists contains a subspace isometric to $\mathcal{H}$, H_ϕ. Moreover, the inner product of $\mathcal{H}$, while inherited from that of H, is not exactly the same. Indeed, if $f_i(\cdot) = \langle v_i, \phi(\cdot)\rangle_H, \ i \in \{1, 2\}$, we have $\langle f_1, f_2\rangle_{\mathcal{H}} = \langle P_{H_\phi} v_1, P_{H_\phi} v_2\rangle_H$ and not $\langle f_1, f_2\rangle_{\mathcal{H}} = \langle v_1, v_2\rangle_H$.

In practice, it is usually hard to construct an infinite dimensional feature map. That is why, contrary to linear models, the point of view of reproducing kernel Hilbert spaces taken in machine learning is usually centered around the kernel function k, which can be used to construct the space $\mathcal{H}$.

Construction from the kernel function. In this case, we view the RKHS from the point of view of the *kernel*. A kernel function on a space $\mathcal{X}$ is simply a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Kernels play an essential role in many branches of mathematics, partial differential equations, harmonic analysis, probability theory. In most of these branches, kernels intervene through a *kernel operator* on a space of measures M on $\mathcal{X}$. This operator is usually defined as a convolution with the kernel k, that is $(T_k \mu)(x) = \int_{\mathcal{X}} k(x, x') \ \mu(dx')$ for measures μ in M. This point of view will be mentioned in Sec. 1.2.4. However, in machine learning, the setting is somewhat less abstract.

In machine learning, the kernel is seen as a similarity measure. This point of view gives natural ways to design kernels for certain specific tasks, in which a good similarity measure can be designed (which can be easier than designing a feature map). This is the case for biological sequences for instance, where kernels have been used to compare DNA and protein sequences (see Jaakkola, Diekhans, and Haussler (1999) and subsequent work). We will give examples of kernels in the next section Sec. 1.2.2.

Sine we would like the kernel k to be a similarity measure between two elements of $\mathcal{X}$, a good feature based linear model would be a model where the feature map $\phi : \mathcal{X} \to H$ (where H is a Hilbert space) reflects this similarity measure : $\langle \phi(x), \phi(x')\rangle_H = k(x, x')$. This is exactly the role of a RKHS $\mathcal{H}$ with kernel k as described above. The cornerstone of kernel methods is that we can indeed find such a space $\mathcal{H}$, given any p.d. kernel (see definition in Eq. (1.33)).

Basic functions. As we saw in the previous paragraph, if $\mathcal{H}$ is a RKHS with kernel k, it must necessarily be the closure of the space $\text{span}\{k_x \ : \ x \in \mathcal{X}\}$ for the right Hilbert structure (that is is must satisfy $\langle k_x, k_{x'}\rangle = k(x, x')$). Therefore, to build a RKHS associated to k, we start by considering the space of elementary functions defined by the kernel.

$$\begin{aligned} \mathcal{H}_0 &= \text{span}(\{k(x, \cdot) \ : \ x \in \mathcal{X}\}) \\ &= \left\{ f : \mathcal{X} \to \mathbb{R} \mid f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x), \ n \in \mathbb{N}, \ (x_i) \in \mathcal{X}^n, \ (\alpha_i) \in \mathbb{R}^n \right\}. \end{aligned} \tag{1.31}$$

Bilinear form This space can be equipped with a bilinear form defined by the kernel (in order to

satisfy $\langle k_x, k_{x'} \rangle_{\mathcal{H}_0} = k(x, x')$) :

$$\left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{m} \beta_j k(x'_j, \cdot) \right\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i k(x_i, x'_j) \beta_j, \tag{1.32}$$

for any $n, (x_i), (\alpha_i)$ and $m, (x'_j), (\beta_j)$. The condition for this bilinear form to be a scalar product is exactly that the kernel k be positive definite, as defined in the previous paragraph, *i.e.*, k is symmetric ($k(x, x') = k(x', x)$ for all x, x'), and

$$\forall n \in \mathbb{N},\ \forall (x_i) \in \mathcal{X}^n,\ \forall (\alpha_i)_{\in} \mathbb{R}^n,\ \sum_{i,j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j \geq 0. \tag{1.33}$$

Construction of $\mathcal{H}$. Under the assumption that k is p.d., the space $\mathcal{H}_0$ is called a pre-Hilbert space, that is a vector space equipped with a scalar product. From $\mathcal{H}_0$, we can construct the space $\mathcal{H}$ as the completion of $\mathcal{H}_0$ with respect to the norm defined by the scalar product Eq. (1.32). Since in that case, $\overline{\mathcal{H}_0} = \mathcal{H}$, it is exactly the reproducing kernel Hilbert space associated to k. This result is due to Aronszajn (1950) who attributes it to Moore (1916).

Theorem 1.1 (Aronszajn (1950))**.** *Let k be a p.d. kernel on a space $\mathcal{X}$. Then there exists a unique RKHS $\mathcal{H}$ with reproducing kernel k.*

Definition as a Hilbert space of functions with the reproducing property. The last way of defining a reproducing kernel Hilbert space is through the function space itself. Assume you are given a Hilbert space of functions $\mathcal{H}$. It is a reproducing kernel Hilbert space if, and only if the evalutations mappings $f \in \mathcal{H} \mapsto f(x)$ are continuous. Indeed, in that case, by the Riesz representation theorem, there exists an element $k_x \in \mathcal{H}$ such that $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Thus $\mathcal{H}$ is a reproducing kernel Hilbert space with kernel $k(x, x') := \langle k_x, k_{x'} \rangle_{\mathcal{H}}$.

This point of view allows to directly understand that certain classical Hilbert spaces of functions are RKHSs. This is the case of Sobolev spaces $W_2^s(\Omega)$ for $\Omega \subset \mathbb{R}^d$, which are reproducing kernel Hilbert spaces for $s > d/2$. However, finding an expression for the kernel is often complicated. In the case where $\Omega = [0, 1]$, it is known for $s = 1$ where $k(x, y) = \max(x, y)$. However, the expression for general s is much more complex (see Wahba (1990)). In Sec. 1.2.2, we will give examples of such kernels. Moreover, this point of view also emphasizes that RKHS are particularly adapted when the function f is known via function evaluations. In a sense, a reproducing kernel Hilbert space contains "diracs", *i.e.*, elements k_x with porperties which are similar to those of diracs in L^2 which satisfy $\langle f, \delta_x \rangle_{L^2(\mathbb{R}^d)} = f(x)$, but which lie in the space itself.

Kernels in machine learning. In this section, we have seen three ways of looking at RKHS and constructing them : one through a Hilbert valued feature map, one through a p.d. kernel on $\mathcal{X}$, and one by looking directly at the space of functions $\mathcal{H}$. The last two constitute the two main trends in the literature using RKHS.

The approach centered around function spaces uses the kernel k as a tool to describe and characterize the space of function $\mathcal{H}$, in particular through the kernel operator T_k defined above. This approach has been used to describe certain spaces in harmonic analysis (Zaremba, 1907).

On the other hand, the approach centered around the p.d. kernel itself has been developed successively by Mercer (1909), Moore (1916, 1935) and Aronszajn (1950).

In machine learning, we use both of these approaches. We are interested in the underlying function space from a modelling and statistics perspective. Indeed, the function space $\mathcal{H}$ is the space of test functions, that is the space where we look for a solution to our surrogate problem Eq. (1.7). Moreover, the natural Hilbert structure on $\mathcal{H}$ and in particular its norm will have a key role in regularization. We will therefore caracterize statistical errors with quantities which illustrate the relationship between our model space of functions $\mathcal{H}$ and the true target function f_* (if it exists). However, the kernel perspective is also of great importance to actually solve Eq. (1.7). Indeed, as will be shown in the next sections, there are large classes of algorithms and machine learning problems of the form Eq. (1.7) which can be solved using only dot products of feature maps in the case of linear models, and can therefore be adapted to the kernel setting by replacing dot products with the kernel **and otherwise ignore the underlying Hilbert space**. This is essentially because many problems depend on the function f only through function evaluations on the data $f(x_i)$ (see Sec. 1.2.3). Moreover, as the basic functions of the RKHS are linear combinations of kernel functions centered at different points, these can be used to approximate any function in the RKHS with given degree of precision. This will lead to compressions of functions in $\mathcal{H}$ in finite dimension, which will be of paramount importance to control the complexity of optimization algorithms (see Sec. 1.2.5 and chapters 6 and 8).

1.2.2 Examples of kernels

In this section, we briefly present and discuss examples of kernels, some of which we will use in the rest of this thesis. We start with a bit of nomenclature, by defining the following standard classes of kernels.

- *Translation invariant kernels* are defined on an abelian group $\mathcal{X}$ and are of the form $k(x, x') = v(x - x')$ for a certain function $v : \mathcal{X} \to \mathbb{R}$;
- *Radial basis function kernels* (RBF kernels) are defined on a space $\mathcal{X}$ equipped with a semi-distance $d_{\mathcal{X}}$ and are of the form $k(x, x') = v(d_{\mathcal{X}}(x, x'))$ for a certain function $v : \mathbb{R}_+ \to \mathbb{R}$.
- *Zonal kernels* are defined on a space $\mathcal{X}$ with a scalar product $\langle \cdot , \cdot \rangle_{\mathcal{X}}$ and are of the form $v(\langle x, x' \rangle_{\mathcal{X}})$ for a function $v : \mathbb{R} \to \mathbb{R}$.

In a sense, all these different types of kernels respect an underlying structure of $\mathcal{X}$: group structure, metric structure, Euclidean structure. One type of kernel which will be central to the different articles regrouped in this thesis are translation invariant kernels on $\mathcal{X} = \mathbb{R}^d$.

Translation invariant kernels on $\mathbb{R}^d$. Let $k(x, x') = v(x - x')$ be a translation invariant p.d. kernel on $\mathbb{R}^d$, and assume that $v \in L^1(\mathbb{R}^d)$. The RKHS defined by k can be conveniently defined through the Fourier transform.

Recall that given a function $f \in L^1(\mathbb{R}^d)$, the fourier transform of f is defined as

$$\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-i\langle \xi, x \rangle_{\mathbb{R}^d}} dx. \tag{1.34}$$

Note that since $v \in L^1(\mathbb{R}^d)$ and k is p.d., the fourier transform of v is nonnegative and bounded by $\|v\|_{L^1(\mathbb{R}^d)}$. Since the Fourier transform defines an isometry (up to a constant factor) from $L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ to $L^2(\mathbb{R}^d)$ equipped with the L^2 distance, it can actually be extended to the whole of $L^2(\mathbb{R}^d)$ and hence is defined on $L^2(\mathbb{R}^d)$ as well. As soon as $v \in L^1(\mathbb{R}^d)$ and defines a p.d. kernel k, the RKHS associated to k can be characterized using the fourier transform in the following way (see for example Scholkopf and Smola (2001)).

$$\begin{aligned}\mathcal{H} &= \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} \frac{|\widehat{f}(\xi)|^2}{\widehat{v}(\xi)} d\xi < \infty \right\}; \\ \forall f, g \in \mathcal{H},\ \langle f, g \rangle_{\mathcal{H}} &= \tfrac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\widehat{f}(\xi)\overline{\widehat{g}(\xi)}}{\widehat{v}(\xi)} d\xi.\end{aligned} \tag{1.35}$$

This clearly shows that the regularity of a function in $\mathcal{H}$ is linked to the regularity of the kernel function (as the regularity of f is related to the decay of $\widehat{f}$). If k and hence v is very regular, $\widehat{v}$ decays very fast, and hence $\widehat{f}$ must also decay very fast to be in $\mathcal{H}$. Note that if k is also radial (*i.e.*, $v(t) = \tilde{v}(\|t\|)$), then the Fourier transform of v is also radial. We will give examples of such kernels in the next paragraph.

Some generic kernels. In this paragraph, we present some widely used kernels, some of which we will use in parts II and III as they caracterize important regularity properties, or have nice computational properties.

Linear and polynomial kernels are defined as $k(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ and $k(x, x') = \langle x, x' \rangle^m,\ m \in \mathbb{N}$ respectively on any Euclidean space $\mathcal{X}$. The associated RKHS to the linear kernel is simply the set of linear functions on $\mathcal{X}$ while the set of functions associated to the polynomial kernel is simply the set of m homogeneous polynomial functions on $\mathcal{X}$. These RKHS are finite dimensional, and of dimension $\binom{m-1}{d-1}$, where d is the dimension of $\mathcal{X}$ and m the exponent of the polynomial kernel.

Gaussian kernels, or also Gaussian RBF kernels, are the radial, translation invariant kernels on $\mathbb{R}^d$, parametrized by a bandwidth σ, and defined as

$$\forall x, x' \in \mathbb{R}^d,\ k_\sigma(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2)). \tag{1.36}$$

It is one of the most used kernels if no *a priori* is known or when trying to approximate a regular functions, because it is easy to implement, and because of the freedom given by the choice of the parameter σ. We use it in almost all the applications we perform in this thesis, as well as in the developments in chapter 6, in which its computational properties play a key role. Using Eq. (1.35), the RKHS $\mathcal{H}_\sigma$ associated to the Gaussian kernel has norm

$$\|f\|^2_{\mathcal{H}_\sigma} = \left(\frac{\sigma}{\sqrt{\pi}}\right)^d \int_{\mathbb{R}^d} |\widehat{f}(\xi)|^2 \exp(\sigma^2 \|\xi\|^2/2) d\xi. \tag{1.37}$$

This shows that the functions belonging to $\mathcal{H}_\sigma$ are extremely regular, as their Fourier transform must decay exponentially fast. Moreover, the parameter σ controls the speed of this decay; the larger the σ, the more regular the functions. In chapter 6, we show that choosing the parameter σ in a good way (*i.e.*, in a way adapted to the number of samples) is statistically equivalent to using a Sobolev kernel of given regularity, defined below.

Sobolev kernels, also called *Matérn kernels*, are defined by Wendland (2004), and are radial, translation invariant kernels on open sets Ω of $\mathbb{R}^d$. They are defined for $s > d/2$ as :

$$\forall x, x' \in \Omega,\ k_s(x, x') = c_s \|x - x'\|^{s-d/2} \mathcal{K}_{s-d/2}(\|x - x'\|), \tag{1.38}$$

where $\mathcal{K}$ is the Bessel function of the second kind (see Wendland, 2004, chapter 5.10) and c_s is a normalizing constant (for more details, see chapter 8). *In the case where Ω has Lispchitz continuous boundary*, the associated RKHS $\mathcal{H}_s$ can be shown to be equivalent to the Sobolev space $W_2^s(\Omega)$ of functions whose derivatives up to order s are square integrable, and its norm is equivalent to the classical norm on these spaces, defined as :

$$\|f\|^2_{W_2^s(\Omega)} = \sum_{|\alpha|\leq s} \int_\Omega |\partial_\alpha f(x)|^2 \, dx. \tag{1.39}$$

Note that when $\Omega = \mathbb{R}^d$, k_s is of the form $k_s(x, x') = v_s(x - x')$, and the Fourier transform of v_s is given by $(1 + |\xi|^2)^{-s}$. This leads to the following expression for the RKHS norm of $\mathcal{H}_s$ associated to k_s :

$$\|f\|^2_{\mathcal{H}_s} = (2\pi)^{-d} \int_{\mathbb{R}^d} |\widehat{f}(\xi)|^2 (1 + |\xi|^2)^s \, d\xi, \qquad f \in \mathcal{H}_s. \tag{1.40}$$

One particular case of Sobolev kernels is the *exponential* or *Laplace* kernel, which is the Sobolev kernel for $s = d/2 + 1/2$ and is equal to $\exp(-\|x - x'\|)$. These kernels are very useful to characterize the regularity of a function in term of derivatives (indeed, the parameter s characterizes the degree of regularity of the functions in the RKHS). This is very useful from a theoretical point of view, to gain intuition on how regularity affects the properties of the learning problem. These kernels will be central examples to understand part I, and the cornerstone kernel in chapter 8, where we use it to leverage the regularity of a function to perform fast global optimization.

Advantages and disadvantages of using "off-the-shelf kernels". Using these "off-the-shelf" kernels has many advantages. From a practical perspective, they are easy to implement and use, and in the case of the Gaussian kernel, allow some tuning to the data by selecting the bandwidth σ, which is interpretable : it is the typical range of action of a data point. From a theoretical perspective, they are also very well known and characterized, and in the Gaussian and Sobolev kernel case, possess a very nice approximation property called *universality* : they can essentially approximate any function on $\mathbb{R}^d$ (for more details, see Sec. 1.2.4 and Micchelli, Xu, and Zhang (2006); Sriperumbudur, Fukumizu, and Lanckriet (2011)). However, their level of generality is also their weakness. These kernels are isotropic and therefore can fail to identify manifold structures in high dimensions (on which it is believed the data is usually concentrated). Moreover, as we see with the Gaussian kernel, there is a single scale of interaction given by σ. This can be a problem when multiple scales are needed to understand data (in image processing for instance; that is the role of convolutional layers).

Kernel engineering. In the previous paragraph, we have seen examples of generic kernels, defined mostly on $\mathcal{X} = \mathbb{R}^d$, which can be used in an "off-the-shelf" way. However, one of the main advantages of kernels is the fact that one can build a kernel suited to a specific task, even on data which is non vectorial (we count images in this setting as the vector does not reflect the geometrical structure of the image). Moreover, this can even be necessary to capture the structure of vectorial data, when it is anisotropic or multi-scale. In this paragraph, we thereofore briefly introduce different ways and examples of building kernels. The basic ways of constructing kernels are mentioned by Scholkopf and Smola (2001) and we therefore refer to it for the basic operations (product of kernels, compositions).

Non-vectorial data. It is impossible to list kernels for non-vectorial data as they are intrinsically taylored to the situation at hand. However, we can mention a few examples of such situations : they have been used on biological data (Jaakkola, Diekhans, and Haussler, 1999), on images (Harchaoui and Bach, 2007), on graphs in biology (Borgwardt, Ghisu, Llinares-Lopez, O'Bray, and Rieck, 2020) and many others.

Building kernels using multiple features. A classical way of designing kernels is, given a potentially infinite number of uniformly bounded features $\psi_i : \mathcal{X} \to \mathbb{R}$, to define the associated Mercer kernel, i.e.

$$k(x, x') = \sum_{i \in I} w_i \psi_i(x) \psi_i(x'). \tag{1.41}$$

where the w_i are non-negative weights which are summable. Note that this exactly corresponds to defining a feature map $\phi : x \in \mathcal{X} \mapsto (\sqrt{w_i}\psi_i(x))_{i \in I} \in \ell_2(I)$. The ψ_i are usually designed either as meaningful features or basis functions, which have been developed in many branches such as PDEs or signal processing. One can think for example of wavelet basis or cosine basis $\cos(\omega_i x)$. Generalizing this, one can consider kernels of the form

$$k(x, x') = \sum_{i \in I} w_i k_i(x, x'), \tag{1.42}$$

for bounded kernels k_i. For example, Opfer (2006) considers a basis function $\varphi(x)$ which is rescaled by a factor δ_i to form the kernel $k_i(x, x') = \delta_i^{-d} \varphi((x - x')/\delta_i)$; they correspond to different scales of interactions between data points. In general, these models can be used *a)* to model the interactions at different scales (as for images) using kernels k_i adapted to different scales δ_i, and *b)* to learn the kernel by learning the weights w_i (Bach, Lanckriet, and Jordan, 2004).

Note that Eq. (1.41) can be generalized (up to rescaling) to define random feature kernels (Rahimi and Recht, 2008) , *i.e.*, kernels of the form

$$k(x, x') = \mathbb{E}_{\omega \sim \mu} \left[\phi(x, \omega) \phi(x', \omega) \right] = \int_{\Omega} \phi(x, \omega) \phi(x', \omega) \mu(d\omega). \tag{1.43}$$

These form of kernels play a crucial role in dimension reduction techniques as we will see in Sec. 1.2.5 and chapter 2. For example, the Gaussian kernel can be seen as a random feature kernel :

$$k_\sigma(x, x') = \mathbb{E}_{\omega, b} \left[\cos(\omega x + b) \cos(\omega x' + b) \right] \qquad \omega \sim \mathcal{N}(0, 1/\sigma \mathbf{I}_d), b \sim \text{Uniform}(0, 2\pi). \tag{1.44}$$

Summary. In this section, we gave examples of kernels and RKHS used in practice and references to how to build them. In particular, we described the most commonly used "off-the-shelf" kernels as well as the induced RKHS, discussed their limitations, and gave certain directions as to how people in the community have tackled these problems and designed problem-specific kernels.

1.2.3 Kernels for empirical risk minimization

In this section, we describe how RKHS are adapted in the context of learning, using the framework from Sec. 1.1. Recall that our "ideal" problem is the expected risk minimization problem, of the form

$$\mathcal{R}_* = \inf_{f\in\mathcal{F}} \mathcal{R}(f) = \mathbb{E}\left[\ell_Z(f)\right], \tag{1.45}$$

where $\mathcal{F}$ is an ideal class of function (think of it as the largest class of functions on which the risk is defined), Z a random variable on a space $\mathcal{Z}$ with distribution ρ representing the data, and ℓ is a loss function. We denote with f_* any minimizer of Eq. (1.45) it it exists. In this section, we will assume that *a)* $\mathcal{F}$ is a class of functions from a space $\mathcal{X}$ to $\mathbb{R}$, and *b)* that we have a map $\pi : \mathcal{Z} \to \mathcal{X}$ such that $\ell_z(f)$ is of the form $\ell(z, f(\pi(z)))$, which we will sometimes denote with $\ell_z(f(\pi(z)))$. This rather abstract assumption simply states that the loss of f at data point z depends on f only through the *evalution* of f at a point $x = \pi(z)$ given by the data (indeed we assume π to be known). Define $X = \pi(Z)$; the expression of the risk becomes $\mathcal{R}(f) = \mathbb{E}\left[\ell_Z(f(X))\right]$.

Remark 2. *We want to emphasize the difference in notations which might be confusing. In the general case introduced in Sec. 1.1 and which we will again use in chapter 3, the loss function ℓ is a map $\ell : \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ and $\ell(z, f)$ evaluates the performance of f at z, while here, the map ℓ is a map $\ell : \mathcal{Z} \times \mathbb{R} \to \mathbb{R}$ which evaluates the performance of the function value of f at z which is $f(x) \in \mathbb{R}$ at $x = \pi(z)$. The fact that the loss only depends on f through function values is key for Theorem 1.2.*

To make things clearer, take the supervised learning setting, where the goal is to predict an output $Y \in \mathbb{R}$ from an input $X \in \mathcal{X}$. The data random variable is therefore $Z = (X, Y) \in \mathcal{Z} = \mathcal{X}\times\mathbb{R}$, and the risk is usually of the form $\mathbb{E}\left[\ell_{(X,Y)}(f(X))\right]$; in that case, the map π is simply the projection on the first coordinate. In the least-squares regression case, we had $\ell_z(f) = \|f(x) - y\|^2$, and hence this corresponds to $\ell_z(t) = \frac{1}{2}\|t - y\|^2$. For logistic regression, the loss function ℓ was defined as $\ell_z(f) = \log(1 + \exp(-yf(x)))$, which corresponds to $\ell_z(t) = \log(1 + \exp(-yt))$.

The fact that the loss function depends only on function evaluations allows us to reformulate *a)* the problem of minimizing the expected risk on $\mathcal{H}$, and *b)* the problem of minimizing the (regularized) empirical risk on $\mathcal{H}$ in a way that f appears only in a scalar product with an element k_x. Indeed, the expected risk minimization on $\mathcal{H}$ can be written

$$\mathcal{R}_{\mathcal{H}} = \inf_{f\in\mathcal{H}} \mathbb{E}\left[\ell_Z(\langle f, k_X\rangle_{\mathcal{H}})\right]. \tag{1.46}$$

We will denote with $f_{\mathcal{H}}$ a minimizer of Eq. (1.46) if it exists. As seen in Eq. (1.9) in Sec. 1.1.2, the surrogate problem which we will effectively solve is the so-called empirical risk minimization problem :

$$\widehat{f}_{n,\lambda} = \operatorname*{arg\,min}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell_{z_i}(\langle f, k_{x_i}\rangle_{\mathcal{H}}) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}, \qquad x_i = \pi(z_i), \tag{1.47}$$

where $z_1, ..., z_n$ are the data points, the $x_i = \pi(z_i)$ are the evaluation points of f associated to the z_i, and the regularization term $\frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$ is called the Tikhonov regularization. This regularization term is the most standard, but note that other regularization procedures exists, which we will not use in this thesis (Blanchard and Mücke, 2018). As RKHS are usually infinite dimensional, a regularization is usually necessary, as explained in Eq. (1.9) and detailed in Sec. 1.2.4. One of the cornerstone theorems in kernel methods is the following.

Theorem 1.2 (Representer theorem, Cucker and Smale (2002)). *Let $L : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous loss function which is bounded below and let $x_1, .., x_n \in \mathcal{X}$, and $\mathcal{H}$ a RKHS. If the problem*

$$\min_{f \in \mathcal{H}} L(f(x_1), ..., f(x_n)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{1.48}$$

has a solution, then there exists a solution of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x), \qquad \alpha \in \mathbb{R}^n. \tag{1.49}$$

Moreover such a solution always exists for $\lambda > 0$.

In other words, any problem of the form Eq. (1.48), that is which depends on f only through its evaluations at $x_1, ..., x_n$, has a minimizer in $\mathcal{H}_n = \text{span}(\{k_{x_1}, ..., k_{x_n}\})$. This is easily proved by decomposing f on $\mathcal{H}_n$ and its orthogonal. There are two valuable consequences of this theorem. The first is that the empirical risk minimization problem Eq. (1.47) can be formulated as a n-dimensional problem :

$$\widehat{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(e_i^\top \mathbf{K}\alpha) + \frac{\lambda}{2}\alpha^\top \mathbf{K}\alpha, \tag{1.50}$$

where $\mathbf{K} = (k(x_i, x_j))_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$ it the kernel matrix associated to the data points (x_i). The second consequence is that the resulting function $\widehat{f}_{n,\lambda}$ can be expressed only using the kernel : $\widehat{f}_{n,\lambda}(x) = \sum_{i=1}^n \widehat{\alpha}_i k(x_i, x)$.

To summarize, the main advantages of kernels in this setting is the fact that *a)* although RKHS can be infinite dimensional, the resulting empirical risk minimization problem is finite-dimensional, and *b)* the problems and resulting functions can be described using only the kernel function, omitting the abstract definition of the RKHS. However, note that the representer theorem (Theorem 1.2) only holds if the empirical risk depends on the function f only through evaluations. This does not include cases where constraints on f may be added. This can sometimes be useful, for example when learning probability densities (as in chapters 5 and 6), where a natural constraint can be that the function sums to one. However, we can show that under certain conditions, looking for f in $\mathcal{H}_n$ can be sufficient statistically (see chapter 6). In this thesis, we will try to move a bit aside from the representer theorem paradigm, and try to show that kernel methods perform well not because $\mathcal{H}_n$ contains the empirical risk minimizer, but because when one has n data points, $\mathcal{H}_n$ is a good enough approximation of $\mathcal{H}$. In the two next sections, we will briefly discuss the satistical properties of the empirical risk minimizer $\widehat{f}_{n,\lambda}$, and the optimization properties of solving the problem Eq. (1.50).

1.2.4 A statistics point of view on e.r.m. with kernels

Running example : least-squares. In the next two sections, and in chapter 2, we will use the kernel least-squares regression problem as a running example. Recall from Sec. 1.1 that in that case, the expected risk is $\mathcal{R}_{\texttt{ls}}(f) = \frac{1}{2}\mathbb{E}\left[\|f(X) - Y\|^2\right]$. Moreover, we will assume that $Y \in L^2$ which shows that there exists an optimal predictor f_* which is in $L^2(\mathcal{X}, \rho_{\mathcal{X}})$. Moreover, in this case, there is a closed form solution to the regularized empirical risk minimization problem Eq. (1.50) given data points (x_i, y_i) and $\lambda > 0$:

$$\widehat{f}_{n,\lambda} = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \qquad \alpha = (\mathbf{K} + \lambda n \mathbf{I})^{-1} y, \;\; y = (y_i) \in \mathbb{R}^n. \tag{1.51}$$

Handling the approximation error : universality. Recall from Sec. 1.1.2 that the standard way of decomposing the excess risk of the empirical risk mimizer $\widehat{f}_{n,\lambda}$ is to separate approximation and estimation error : $\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}_* = (\mathcal{R}_{\mathcal{H}} - \mathcal{R}_*) + (\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}_{\mathcal{H}})$. In kernel methods, as we will see in the next paragraph, one usually only looks at the estimation error.

Indeed, since RKHSs are usually infinite dimensional, the approximation error $\mathcal{R}_{\mathcal{H}} - \mathcal{R}_*$ can be shown to be zero for certain RKHSs and certain losses. These results are usually called *universality* results; they aim to show that under certain conditions on the RKHS $\mathcal{H}$ (or the kernel k), $\mathcal{H}$ can approximate any function $f \in \mathcal{F}$, i.e. is dense in $\mathcal{F}$ for a certain class of functions $\mathcal{F}$. Of course, "approximating any function" has to be defined, and many definitions can be given (Micchelli, Xu, and Zhang, 2006; Sriperumbudur, Fukumizu, and Lanckriet, 2011). If $\mathcal{X}$ is a locally compact Hausdorff space for example (this is the case of all spaces $\mathcal{X}$ we consider), one type of universality is the L^p-universality : $\mathcal{H}$ is dense in $L^p(\mathcal{X}, \rho_{\mathcal{X}})$ for any given Borel probability measure $\rho_{\mathcal{X}}$ (usually the law of X in the supervised setting), where $L^p(\mathcal{X}, \rho_{\mathcal{X}})$ is equipped with the L^p norm. Another type of universality, *cc* -universality, is the fact that $\mathcal{H}$ is dense in the set $C(\mathcal{X})$ of continuous functions on $\mathcal{X}$ for the compact convergence (*i.e.*, for any compact $K \subset \mathcal{X}$ and any function $f \in C(\mathcal{X})$, there exists a sequence $f_n \in \mathcal{H}$ of functions such that $\sup_{x \in K} \|f_n(x) - f(x)\| \underset{n \to \infty}{\longrightarrow} 0$). For instance, in the least-squares problem, the approximation error is zero as soon as the kernel is L^2-universal for the measure $\rho_{\mathcal{X}}$. The Gaussian and Sobolev kernels can be proved to be L^p-universal and *cc*-universal for instance (Micchelli, Xu, and Zhang, 2006; Sriperumbudur, Fukumizu, and Lanckriet, 2011).

Remark 3. *The approximation error being zero does not mean that f_*, if it exists, is in $\mathcal{H}$. Zero approximation error simply means that there exists a sequence f_n of functions in $\mathcal{H}$ such that $\mathcal{R}(f_n) \to \mathcal{R}(f_*)$. Indeed, if $\mathcal{H}$ is the Sobolev space or order s, it is of course not true that every function in $L^2(\mathbb{R}^d, \rho_{\mathcal{X}})$ for a given probability measure $\rho_{\mathcal{X}}$ is also in $W_2^s(\mathbb{R}^d)$.*

In the following discussion and throughout this thesis, we will only analyse the estimation error, and make hypothesis which always ensure that $\mathcal{R}_{\mathcal{H}} = \mathcal{R}_*$, therefore making the estimation error the actual excess risk.

Bias variance trade-offs in kernel methods. As explained in Sec. 1.1.2, when the space $\mathcal{H}$ is large, we modulate its capacity using some form of regularization. In this thesis we will use the Tikhonov regularization $\frac{\lambda}{2}\|f\|_{\mathcal{H}}$ as it depends only on the kernel norm, and has good optimization properties (it will make the objective strongly convex, see Sec. 1.1.3). We mention other regularizations in chapter 2.

In the case of kernel methods, the traditional approximation-estimation decomposition can be replaced by a so-called *bias-variance* decomposition of the estimation error, of the form

$$\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}_{\mathcal{H}} \leq \text{Bias}(\lambda) + \text{Variance}(\lambda, n), \tag{1.52}$$

where the bias term depends only on λ and can be seen as an analog of the approximation error for the space "$\mathcal{H}$ regularized with λ", and the variance term is an analog of the estimation error, and quantifies the statistical impact of having access to the true distribution only through the n data points.

More precisely, in the case of Tikhonov regularization, if f_λ is a minimizer of the regularized expected risk $\mathcal{R}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}$ (which exists as soon as $\mathcal{R}$ is lower semi-continuous and lower bounded as will be the case for all losses considered), then one can see the bias term as the excess risk of f_λ : $\text{Bias}(\lambda) \approx \mathcal{R}(f_\lambda) - \mathcal{R}_{\mathcal{H}}$. This bias term can be quantified by the regularity of the

optimal solution f_* with respect to the chosen RKHS $\mathcal{H}$. If f_* is in $\mathcal{H}$ for example, the bias term can be shown to behave as $\text{Bias}(\lambda) = O_0(\lambda)$. If f_* is not in $\mathcal{H}$, then the bias term will converge more slowly towards 0. This behavior is more precisely quantified in the context of least squares, briefly described below.

For the variance term, the typical form we will be looking for is $\text{Variance}(\lambda, n) = \frac{d_\lambda}{n}$, where d_λ is a notion of intrinsic dimension of the space $\mathcal{H}$ regularized with λ. This type of bounds is called "dimensionless" bounds, not because no notion of dimension appear, but because the dimension of the space $\mathcal{H}$ does not appear, which is necessary if one is to quantify the variance on an infinite dimensional Hilbert space.

In chapter 3, we establish such bias-variance decompositions in the case where the loss satisfies a certain property called generalized self-concordance. This decomposition has already been well studied in the literature in the least-squares setting.

The case of least squares. In the case of least-squares regression, such bias variance decompositions have been obtained in order to study optimal rates of convergence (Caponnetto and De Vito, 2007; Blanchard and Mücke, 2018). In the setting of Tikhonov regularization, we can prove the following bias variance decompositions in high probability (a more formal statement can be found in chapter 2) :

$$\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}_{\mathcal{H}} \leq C \left(\underbrace{b_\lambda}_{\text{bias}} + \underbrace{\frac{d_\lambda}{n}}_{\text{variance}} \right), \tag{1.53}$$

where b_λ is the bias term and d_λ is the so-called effective dimension of $\mathcal{H}$ (Caponnetto and De Vito, 2007) and C is a constant depending on the probability with which we want this statement to hold (as the sample points $z_1, ..., z_n$ are random). Informally, the effective dimension is a measure of the size of $\mathcal{H}$ with respect to $L^2(\mathcal{X}, \rho_{\mathcal{X}})$, where we know the optimal function f_ρ lives (we emphasize the link of the optimum with the measure in this section using this notation).

To understand the behavior of the bias term as well as the definition of the effective dimension, we need to introduce a bit more kernel machinery. Assume that the kernel k is continuous and bounded (*i.e.*, $k(x, x) \leq \kappa^2$), and that $\mathcal{X}$ is separable. Let ρ be the data distribution, $\rho_{\mathcal{X}}$ be its marginal on $\mathcal{X}$ and assume that $Y \in L^2(\mathcal{X} \times \mathbb{R}, \rho)$, that is is square integrable.

Understanding the effective dimension. By boundedness of the kernel, functions in $\mathcal{H}$ also belong to $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ ($\|f\|_{L^2} \leq \kappa \|f\|$) and that the $S : \mathcal{H} \to L^2(\mathcal{X}, \rho_{\mathcal{X}})$ is a linear continuous map between Hilbert spaces. Define the covariace operator :

$$\Sigma f = S^* S f = \int_{\mathcal{X}} f(x') k_{x'} \rho_{\mathcal{X}}(dx') \text{ such that } \langle f, \Sigma g \rangle = \langle f, g \rangle_{L^2(\mathcal{X}, \rho_{\mathcal{X}})}. \tag{1.54}$$

It is a trace-class, symmetric positive semidefinite operator (hence compact), and describes the relationship between the $\mathcal{H}$-norm and the L^2-norm, which is the one adapted to the least-squares problem. As Σ is a trace-class operator, it can be decomposed using the spectral theorem as $\sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i$ where the λ_i are positive values sorted in decreasing order, I is either of the form $\{1, ..., k\}$ if Σ is finite rank or $\mathbb{N}$ (which is the most common case), and

ϕ_i are corresponding renormalized eigenvectors. The space $\mathcal{H}$ can then be decomposed as $\mathcal{H} = \text{span}(\{\phi_i \ : \ i \in I\}) \overset{\perp}{\oplus} \text{Ker}(\Sigma)$, where $\text{Ker}(\Sigma)$ is the set of all functions in $\mathcal{H}$ such that $\|f\|_{L^2(\mathcal{X},\rho_\mathcal{X})} = 0$ (this is possible if the support of f is disjoint from the support of the measure $\rho_\mathcal{X}$). Note that the spectrum of Σ gives a lot of quantitative information about the size of $\mathcal{H}$ with respec to L^2. If the spectrum is finite, this means that only a finite dimensional subset of $\mathcal{H}$ is actually meaningful to approximate a function in L^2 norm. More generally, a large λ_i will be associated to an eigenvector capturing a lot of L^2 information, whereas eigenvectors associated to small λ_i will capture almost no information about the L^2 structure. The speed at which the eigenvalues converge to zero is therefore a good indicator of the size of the space $\mathcal{H}$ with respect to the ideal space $L^2(\mathcal{X}, \rho_\mathcal{X})$. This is exactly what the *effective dimension* d_λ measures, and is defined as

$$d_\lambda = \text{Tr}((\Sigma + \lambda \mathbf{I})^{-1}\Sigma) = \sum_{i \in I} \frac{\lambda_i}{\lambda + \lambda_i}, \qquad \lambda > 0. \tag{1.55}$$

The effective dimension d_λ is a quantity such that $d_\lambda \underset{\lambda \to 0}{\longrightarrow} +\infty$ as soon as the spectrum is infinite; what is interesting is the speed at wich this quantity goes to infinity. This is linked to the eigenvalue decay λ_i. For example, if $\mathcal{H}$ is a the RKHS associated to the Sobolev kernel k_s (see Sec. 1.2.2), and if the measure $\rho_\mathcal{X}$ is comparable to the Lebesgue measure (there exists a constant K such that $\rho_\mathcal{X} \leq K\Lambda$ where Λ is the Lebesgue measure), the eigenvalue decay is of the order $\lambda_i = O(i^{-2s/d})$, which implies that the $d_\lambda \leq C\lambda^{-d/(2s)}$. This confirms the intuition that the smaller the space (the larger the s) the smaller d_λ gets.

Understanding the bias term. The bias term is a term which quantifies the regularity of f_ρ with respect to the RKHS $\mathcal{H}$. We have $b_\lambda = o_{\lambda \to 0}(1)$ and the speed of convergence of b_λ towards zero is characterized by the regularity of f_ρ. If $f_\rho \in \mathcal{H}$, we have $b_\lambda = O(\lambda)$. If f_ρ is very regular, it goes even faster towards zero. To formally define what the regularity of f_ρ with respect to $\mathcal{H}$ means, define the kernel operator associated to k : $T_k : L^2(\mathcal{X}, \rho_\mathcal{X}) \to L^2(\mathcal{X}, \rho_\mathcal{X})$ such that

$$\forall g \in L^2(\mathcal{X}, \rho_\mathcal{X}), \ (T_k g)(x) = \int_\mathcal{X} k(x, x')g(x')\rho_\mathcal{X}(dx'). \tag{1.56}$$

The operator T_k is linked to the covariace operator through the map S which injects $\mathcal{H}$ in $L^2(\mathcal{X}, \rho_\mathcal{X})$; we have $\Sigma = S^*S$ and $T_k = SS^*$. In particular, $T_k = \sum_{i \in I} \lambda_i \ \psi_i \otimes_{L^2} \psi_i$ where the λ_i are the same eigenvalues as those of Σ and the ψ_i are the corresponding normalized eigenvectors, which are related to the eigenvectors of Σ, seen as elements of L^2, by the relation $\psi_i = \phi_i/\sqrt{\lambda_i}$. The ψ_i form an orthonormal basis of $\text{Ker}(T_k)^\perp \subset L^2(\mathcal{X}, \rho_\mathcal{X})$. Moreover, by the Mercer theorem (see Carmeli, De Vito, and Toigo (2005) and Dieuleveut and Bach (2016) for an extension), it can be shown that $\mathcal{H} = \text{range}(T_k^{1/2}) \oplus \mathcal{H}_0$ where $\mathcal{H}_0 = \{f \in \mathcal{H} \ : \ \|f\|^2_{L^2(\mathcal{X},\rho_\mathcal{X})} = 0\}$. Note that formally, $L^2(\mathcal{X}, \rho_\mathcal{X}) = \text{range}(T_k^0)$.

In general, we will say that $f \in L^2(\mathcal{X}, \rho_\mathcal{X})$ is r regular if it belongs to the set $\mathcal{H}^r := \text{range}(T_k^{r/2})$ for $r \geq 0$. We have $\mathcal{H}^0 = L^2(\mathcal{X}, \rho_\mathcal{X})$, $\mathcal{H}^1 \subset \mathcal{H}$, and in general $\mathcal{H}^r \subset \mathcal{H}$ for $r \geq 1$. As r increases, the regularity of the elements of $\mathcal{H}^r$ increases. Making the assumption that $f_\rho \in \mathcal{H}^r$ is a good way of quantifying the regularity of f_ρ (which is $r = 0$ *a priori*). The bias b_λ can be controlled by the regularity, and as expected, the more regular the function, the faster b_λ goes to zero :

$$b_\lambda \leq C \ \lambda^r \text{ if } f_\rho \in \mathcal{H}^r, \tag{1.57}$$

for a constant C depending on the norm of f_ρ in $\mathcal{H}^r$. Note that this condition can also be expressed as a condition on the coefficients f_i of the decomposition of f_ρ over the eigenbasis ψ_i with respect to the λ_i, *i.e.*, $f_\rho \in \mathcal{H}^r$ i.i.f $\sum_{i\in I} f_i^2/\lambda_i^r < +\infty$.

Example 1.1 (Uniform measure and Sobolev spaces)**.** *In the Sobolev kernel case, and if $\rho_\mathcal{X}$ is the uniform measure over $\mathcal{X}$, and if $\mathcal{H}_s$ denotes the RKHS associated to k_s, $f_\rho \in \mathcal{H}_s^r$ simply means that $f_\rho \in W_2^{sr}(\mathcal{X})$, which corresponds exactly to the interpolation between Sobolev spaces.*

Conclusion. In the least-squares case, we have a very fine way of quantifying the bias-variance trade-off, and a good understanding of the role of the regularity of f_ρ as well as that of the effective dimension of $\mathcal{H}$ in the performance of the empirical risk minimizer. This can help us gain some intuition on why a learning problem is hard or easy, or what to do to reduce its dimension (see part I).

1.2.5 An optimization point of view on e.r.m. with kernels

Kernel methods have both great potential advantages and drawbacks in terms of optimization to find the empirical risk minimizer $\widehat{f}_{n,\lambda}$. Recall that using the representer theorem, finding the e.r.m. minimizer is equivalent to solving Eq. (1.50) recalled here for clarity :

$$\widehat{\alpha} \in \arg\min_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(e_i^\top \mathbf{K}\alpha) + \frac{\lambda}{2}\alpha^\top \mathbf{K}\alpha.$$

The main advantages are that the problem inherits convexity from the loss functions $l_z(\cdot)$ as well as the smoothness of l_z as soon as the kernel is bounded. Moreover, if we reparametrize the problem setting $\beta = \mathbf{K}^{1/2}\alpha$, the problem is λ-strongly convex in β, which allows to apply fast first order methods. The disadvantages are that the kernel matrix can be huge (of size $n \times n$), which is prohibitive for large data sets both in terms of memory and computational cost (as computing one stochastic gradient would cost $O(n)$ and a full gradient $O(n^2)$). Moreover, reparametrizing with β such that $\alpha = \mathbf{K}^{1/2}\beta$ is both necessary (because the kernel matrix is very ill-conditioned) and costly ($O(n^3)$ due to the matrix inversion and cannot be fully parallelized on GPUs). Even then, the λ which appear to be optimal after evaluation of the model are very small in many cases (see the case of the Higgs data set in chapter 4). This means that the problem will be very ill-conditioned even after reparametrization, and that the classical first order methods will be slow.

Thankfully, in recent years, ways have been developed to overcome these challenges mainly in the context of least-squares regression. The first main area of improvement has been dimension reduction; i.e. reducing the dimension of the optimization problem to a certain $m \ll n$. There are two main ways of doing so. One method is to sub-samples points $\tilde{x}_1, ..., \tilde{x}_m$ from $x_1, ..., x_n$ and looks for a solution to the e.r.m. problem restricted to the set of functions of the form $f(x) = \sum_{j=1}^m \tilde{\alpha}_j k(\tilde{x}_j, x)$ (this is called Nyström or column sub-sampling, see Rudi, Camoriano, and Rosasco (2015)). The other is to have a kernel whose expression is of the form $k(x, x') = \mathbb{E}_{\omega\sim p}[\phi(x,\omega)\phi(x',\omega)]$, that as an expectation of random features (see Eq. (1.43) and the following discussion on the Gaussian kernel); in that case, one approximates the kernel matrix with a low rank approximation of the form $\mathbf{K} = RR^\top$ where $R = \frac{1}{\sqrt{m}}(\phi(x_i,\omega_j))_{\substack{1\le i\le n\\ 1\le j\le m}}$ where the ω_j are drawn from p. We will detail the properties of these methods in chapter 2; they allow to reduce the dimension from n to m of order d_λ, without losing any statistical performance for the empirical risk minimizer. The second area of improvement has been algorithms. Using the conjugate gradient method Eq. (1.19) in proposition 1.1 along with a pre-conditioning technique

(for more details, see chapter 2), Rudi, Carratino, and Rosasco (2017) have sped up the solving of the dimension-reduced e.r.m. to a complexity of order roughly $O(nd_\lambda)$, without paying the price of ill-conditioning. These methods have been implemented in a library by Meanti, Carratino, Rosasco, and Rudi (2020) and allow to handle billions of points, using in particular subroutines developed by Charlier, Feydy, Glaunes, Collin, and Durif (2021) to compute kernel evaluations $k(x, x')$ in a fast way.

In part I and more specifically in chapter 4, we extend these techniques for other losses than the square loss. These techniques have since been implemented in a library by Meanti, Carratino, Rosasco, and Rudi (2020), along with the fast least-squares techniques.

1.3 Outline of the thesis

In this section, we give a very broad outline of the thesis and its contributions; these will be made precise in the introduction of each part. The thesis is organised in three main parts, which each corresponds to two published or submitted articles on a specific topic. Note that this outline may have some redundencies with the introductions of these different works.

1.3.1 Part I : fast rates and algorithms for logistic regression

This first part is based on the two articles by Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019); Marteau-Ferey, Bach, and Rudi (2019). In these two works, we study the classical problem of expected risk minimization Eq. (1.4) through the solving of the regularized empirical risk minimization problem Eq. (1.7), on a reproducing kernel Hilbert space $\mathcal{H}$ with kernel k (see Sec. 1.2). The goal of these two works was to extend the precise bias-variance trade-offs, fast rates and fast algorithms already known for least-squares to generalized self-concordant losses. This is an interesting class of losses as it contains the logistic regression loss $\ell_{x,y}(f) = \log(1 + \exp(-yf(x)))$, and the multiclass logistic regression loss (see Sec. 1.1.1).

Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) start by studying the excess risk of the regularized empirical risk minimizer $\widehat{f}_{n,\lambda}$. We show that as in the least-squares case, there is a bias-variance trade-off of the form

$$\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}_{\mathcal{H}} \leq C \log \tfrac{1}{\delta} \left(b_\lambda + \tfrac{d_\lambda}{n} \right), \qquad \text{with probability at least } 1 - \delta, \tag{1.58}$$

where the bias term and the effective dimension are defined as meaningful quantities which match those defined in the least-squares case (where essentially, one replaces the covariace operator with the Hessian matrix of the expected risk at the optimum). In particular, this allows us to essentially derive upper rates of convergence.

Marteau-Ferey, Bach, and Rudi (2019) develop a practical algorithm to solve large-scale kernel logistic regression, which is now used in the library created by Meanti, Carratino, Rosasco, and Rudi (2020). It is based on three contributions.

- The first is to show that as in the least-squares case, the e.r.m. problem can be restricted to a much smaller dimensional problem through Nyström sampling (we reduce the dimension from n to $m \approx d_\lambda$).

- The second is to design a globally convergent second-order algorithm for generalized self-concordant losses, whose complexity is of order $O(nd_\lambda)$ up to log factors. Similarly to interior point methods, this method uses a decreasing λ scheme $(\lambda_0, ..., \lambda_T)$ to iteratively optimize the regularized empirical risk for λ_{t+1} starting at the previous estimation $\widetilde{f}_{n,\lambda_t}$. The optimization at each step is performed using an approximate Newton method, which allows to approximately compute Newton steps using Hessian sketching.
- The third is to prove that the result of the algorithm achieves the same statistical error as that of the full e.r.m. estimator, *i.e.*, it satisfies a bound of the form Eq. (1.58).

Crucially, this method is independent of the conditioning of the problem (up to log factors). We illustrate this fact on real-life data sets where the optimal regularization parameter λ has to be taken very small.

1.3.2 Part II : introducing non-parametric models for non-negative functions, and applications to sampling

This part is based on the two articles by Marteau-Ferey, Bach, and Rudi (2020, 2022a). In this part, we develop a model for non-negative functions based on reproducing kernel Hilbert spaces, and apply this model to sample from probability distributions given their un-normalized density function, breaking the curse of dimensionality with regularity.

In chapter 5, we describe the work by Marteau-Ferey, Bach, and Rudi (2020). We consider a class of models with non-negative outputs, which exhibit the same properties as linear models and kernel methods. This model is to consider functions parametrized by a PSD operator on a Hilbert space $\mathcal{H}$ on which the input space $\mathcal{X}$ is mapped through $\phi : \mathcal{X} \to \mathcal{H}$ with linear models notations and $x \in \mathcal{X} \mapsto k_x \in \mathcal{H}$ in RKHS notation :

$$\{f_A \ : \ A \succeq 0\} \qquad f_A(x) = \phi(x)^\top A \phi(x) \text{ or } f_A(x) = \langle k_x, A k_x \rangle_{\mathcal{H}}, \ x \in \mathcal{X}. \tag{1.59}$$

We call these models *PSD models*. As this model is itself linear, it can directly be used in the same way to solve e.r.m. problems Eqs. (1.47) and (1.50). We derive a representer theorem similar to Theorem 1.2 for our models in the context of empirical risk minimization, and provide a convex finite-dimensional dual formulation of the learning problem, depending only on the training examples. We also prove that this model has good approximation properties, and apply the method to the problems of density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression. We derive the corresponding learning algorithms for convex dual formulation, and compare it with standard techniques used for the specific problems on a few reference simulations.

In chapter 6, we describe the work by Marteau-Ferey, Bach, and Rudi (2022a). This work applies the modelling framework above to the problem of sampling n i.i.d. samples from a distribution whose density is known up to a constant through function evaluations. Contrary to most of the existing methods in the literature such as rejection sampling or Monte-Carlo Markov chain methods (Gelman, Carlin, Stern, and Rubin, 2004; Liu, 2008; Lelièvre, Rousset, and Stoltz, 2010; Robert and Casella, 2013), we propose a two-step procedure by first modelling this density using a positive semidefinite model, and then sampling from this PSD model using an adapted algorithm. In particular, we use PSD models with the Gaussian kernel defined in Sec. 1.2.2 (we

will call these models Gaussian PSD models), that is we approximate the target density with a function of the form

$$f_A(x) = \sum_{i,j=1}^{n} A_{ij} k_\sigma(x_i, x) k_\sigma(x_j, x), \qquad A \succeq 0,\ k_\sigma(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2)). \tag{1.60}$$

modelling probability distributions with Gaussian PSD models has been developed by Rudi and Ciliberto (2021). In this work, we continue to explore the good properties of this model for probability distributions, and derive an algorithm that is easy to implement and which can generate an arbitrary number of i.i.d. samples from a given Gaussian PSD model, with any given precision.

We then show that we can sample an arbitrary number of i.i.d. samples from a target probability distribution that is regular enough, with any given precision. The algorithm consists in (a) approximating the un-normalized density p via a PSD model, using evaluations of p, and (b) extracting i.i.d. samples from the PSD model. We show that for sufficiently regular densities the resulting PSD model is concise and avoids the curse of dimensionality : to achieve error ε, the Gaussian PSD model requires a number of parameters and a number of evaluations of p that are in the order $\varepsilon^{-2-d/\beta}$, where d is the dimension of the space and β is the order of differentiability of the density.

1.3.3 Part III : making a bridge with sum of squares : global optimization with PSD models

This part is based on the work by Rudi, Marteau-Ferey, and Bach (2020); Marteau-Ferey, Bach, and Rudi (2022b).

Rudi, Marteau-Ferey, and Bach (2020) describe a framework and algorithm to perform global optimization of functions f of class C^r on an open subset $\mathcal{X}$ of $\mathbb{R}^d$. We have described this framework in Sec. 1.1.4. Recall that we formulate the ideal problem $\min_{x\in\mathcal{X}} f(x)$ as a convex problem

$$\begin{aligned} &\sup\ c \\ &\text{subject to } c \in \mathbb{R},\ g = f - c,\ g \geq 0, \end{aligned} \tag{1.25}$$

where the functional constraints $f - c = g$ and $g \geq 0$ have to hold pointwise for all $x \in \mathcal{X}$. Similarly to the poynomial sum of squares setting (Lasserre, 2010), we tighten the constraint $g \geq 0$ by asking g to be a PSD model of the form :

$$g_A(x) = \langle k_x, A k_x \rangle_{\mathcal{H}_s},\ A \succeq 0,\ \mathrm{Tr}(A) < \infty, \tag{1.61}$$

where $\mathcal{H}_s$ is the Sobolev space on $\mathcal{X}$ equipped with the Sobolev kernel k_s. Eq. (1.25) is thus tightened as

$$\begin{aligned} &\sup\ c \\ &\text{subject to } c \in \mathbb{R},\ f(x) - c = \langle k_x, A k_x \rangle_{\mathcal{H}_s},\ x \in \mathcal{X},\ A \succeq 0,\ \mathrm{Tr}(A) < \infty. \end{aligned} \tag{1.62}$$

In order to solve a finite dimensional problem, and assuming we have access to f through function values, we use the following surrogate problem which replaces the constraint $f - c = g_A$ by the

same constraint evaluated at n sample points $x_1, ..., x_n$:

$$\begin{aligned} &\sup\ c - \lambda \operatorname{Tr}(A) \\ &\text{subject to } c \in \mathbb{R},\ f(x_i) - c = \langle k_{x_i}, A k_{x_i} \rangle_{\mathcal{H}_s},\ x \in \mathcal{X},\ A \succeq 0, \end{aligned} \tag{1.63}$$

where we have added an extra regularization term $\lambda \operatorname{Tr}(A)$ to avoid overfitting due to the finite amount of data. Note that this problem is neither a tightening or a relaxation of the initial problem, which is one of the main differences with polynomial optimization.

Rudi, Marteau-Ferey, and Bach (2020) show that the SDP in Eq. (1.63) can be solved with ε error using a Newton method in time $O(n^{3.5} \log \frac{1}{\varepsilon})$. Moreover, we prove that using the Sobolev kernel k_s with $s = r - 3$, the obtained estimation $\widehat{c}_{n,\lambda}$ of the minimium f_* satisfies roughly

$$\|\widehat{c}_{n,\lambda} - f_*\| \leq C_{r,d} n^{-r/d+1/2+3/d}, \qquad \lambda = n^{-r/d+1/2}, \tag{1.64}$$

where r is the regularity of f, $C_{r,d}$ is a constant which depends on the dimension d, on f, and on its regularity r, but not on n. The point of this result is to show that this algorithm indeed breaks the curse of dimensionality in the rates for global optimization, leveraging the regularity to obtain a convergence of order $n^{-r/d}$ (as opposed to typical slow rates of order $n^{-1/d}$). Note however that the constants may still depend exponentially in d. Rudi, Marteau-Ferey, and Bach (2020) also provide ways of computing the global minimizer.

The main assumption in order for Eq. (1.64) to hold is that Eq. (1.62) has a solution, i.e. that there exists a trace-class operator A such that $f(x) - f_* = \langle k_x, A k_x \rangle_{\mathcal{H}_s}$ for all $x \in \mathcal{X}$. A sufficient condition for this to hold is the fact that $f - f_*$ is a finite sum of squares of functions $f_i \in C^s(\mathcal{X})$. In Rudi, Marteau-Ferey, and Bach (2020), we assume that f has only strict minima (i.e. the Hessian at the minima is positive definite), and that f stays far away from its minimum at the boundary of $\mathcal{X}$, *i.e.*, that $\{f - f_* \leq \delta\}$ is compact in $\mathcal{X}$ for a certain $\delta > 0$.

Marteau-Ferey, Bach, and Rudi (2022b) provide an assumption for a function $f \in C^{s+2}$ to be decomposed as a sum of squares of functions of class C^s which generalizes the one above, and allows to deal with functions which are defined on manifolds, and which have manifolds of zeros. The condition we find is that the Hessian of f has to be positive definite orthogonally to the manifolds of zeros. This is an important case for situations where functions f have "large" sets of zeros, such as in optimal transport (see Vacher, Muzellec, Rudi, Bach, and Vialard (2021)). This is linked in spirit to the different *Positivstellensätzen* in polynomial optimization, to guarantee convergence of algorithms.

Part I

Fast rates and algorithms for generalized self-concordant losses

Table of Contents

2 Backround and main results 55

3 Beyond least squares 93

4 Second order strikes back 137

Chapter 2

Backround and main results

Contents

2.1 Fast rates for least-squares 57
2.2 Generalized self-concordance 73
2.3 Main results and contributions of this part 80

In this chapter, we present work by Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019); Marteau-Ferey, Bach, and Rudi (2019). These two articles are then reported as verbatim, without modification from their original version, in chapter 3 and chapter 4 (these articles have been peer-reviewed and published). As explained in Sec. 1.3.1, our ideal goal is to study the classical problem of expected risk minimization Eq. (1.4) through the solving of the regularized empirical risk minimization problem Eq. (1.7), on a reproducing kernel Hilbert space $\mathcal{H}$ with kernel k (see Sec. 1.2). The goal of these two works is to extend the precise bias-variance trade-offs, fast rates and fast algorithms already known for least-squares to a broader class of loss functions including logistic regression and multiclass logistic regression (see Sec. 1.1.1).

We will slightly simplify and reformulate the different problems we consider from the original works from Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019); Marteau-Ferey, Bach, and Rudi (2019) in order to keep uniform notations between introductions, and to make the statements clearer. More formally, we will consider the setting where we are learning a function $f : \mathcal{X} \to \mathbb{R}$ (note that the results hold for $\mathbb{R}^p$ valued functions and kernels in the reference articles), where $\mathcal{X}$ is a Borel space, through i.i.d. samples $z_1, ..., z_n \sim Z \in \mathcal{Z}$. The points $z_1, ..., z_n$ define n evaluation points $x_i = \pi(z_i)$ which follow the law of $X = \pi(Z)$ (we assume π to be known). We also assume the loss function is of the form $\ell_z(t)$, $z \in \mathcal{Z}$, $t \in \mathbb{R}$ ($\ell_z(f(x))$ measures the cost of predicting $f(x)$ at z). In that case, the expected risk minimization problem takes the form :

$$\begin{aligned} &\text{find } f_\rho \\ &\text{such that } \mathcal{R}(f_\rho) = \min_{f \in \mathscr{M}(\mathcal{X}, \mathbb{R})} \mathcal{R}(f), \qquad \mathcal{R}(f) = \mathbb{E}_{Z \sim \rho} \left[\ell_Z(f(X)) \right], \end{aligned} \tag{2.1}$$

where $\mathscr{M}(\mathcal{X}, \mathbb{R})$ is the set of measurable functions from $\mathcal{X}$ to $\mathbb{R}$, and ρ is the probability distribution of Z. We highlight the relation of f_ρ (if it exists) with the distribution ρ. As in the

introduction, $\rho_{\mathcal{X}}$ will denote the density of $X = \pi(Z)$ if Z has density ρ. As seen in Eq. (1.47) in Sec. 1.2.3, the regularized empirical risk minimizer can be defined as

$$\widehat{f}_{n,\lambda} = \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(\langle f, k_{x_i} \rangle_{\mathcal{H}}) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2, \qquad x_i = \pi(z_i). \tag{1.47}$$

To start with, we will make the following three assumptions.

Assumption 2.1 (well-specified)**.** *For the data distribution ρ, f_ρ exists and belongs to the space $\mathcal{H}$.*

This assumption implies in particular that the approximation error is zero. This is a strong assumption, but makes the whole statistical analysis much easier. It is commonly made in the reference works on statistics for least-squares, such as the work by Blanchard and Mücke (2018). Caponnetto and De Vito (2007) make a slightly different assumption, which is weaker in the case where the set $\mathcal{H}$ is not L^2 universal (it asks only that the projection of f_ρ on the closure of $\mathcal{H}$ in $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ be in $\mathcal{H}$), but which, in the case of the Gaussian or Sobolev kernels, corresponds exactly to Assumption 2.1. Moreover, this does not change the analysis much, and we believe that in general, analysis using Assumption 2.1 can easily be adapted to this projected setting (with a slightly different control of the noise term in Assumption 2.4).

Assumption 2.2 (convexity)**.** *The loss function ℓ is convex, i.e., $\ell_z(\cdot)$ is convex for all $z \in \mathcal{Z}$.*

This assumption is a weaker version than the one we will actually make later, the generalized self-concordance assumption. However, it already allows to guarantee that for any $\lambda > 0$, the empirical risk minimizer exists and is unique; $\widehat{f}_{n,\lambda}$ is therefore well-defined.

Assumption 2.3 (bounded kernel)**.** *The kernel k is bounded by 1, i.e., $k(x, x) \leq 1$ for all $x \in \mathcal{X}$.*

This last assumption could easily be replaced by the condition that $k(x, x) \leq \kappa^2$ for some positive constant κ, by considering the modified kernel $k/(\kappa^2)$ which defines the same RKHS with norm $\frac{1}{\kappa}\|\cdot\|_{\mathcal{H}}$. For simplicity, we therefore assume that $\kappa = 1$. It is satisfied by the Gaussian and Sobolev kernels (see Sec. 1.2.2), which will be the ones we will use the most in this thesis.

In this chapter, we start by presenting the backround which leads to the statistical and optimization results of this part. In particular, in Sec. 2.1, we continue the work started in Secs. 1.2.4 and 1.2.5 and recall the main results for least-squares regression. These results concern bias-variance trade-offs, optimal rates of convergence in terms of statistics, dimension reduction (which is also statistical), and optimization. After presenting this vast literature on square-loss, Sec. 2.2 introduces the main tool which has allowed us to go from the square-loss setting to a broader class of functions : the notion of generalized self-concordance (Bach, 2010). In Sec. 2.3, we present the main results of this part, using the same notations as that of the introduction.

Note on slow rates

If the loss $\ell_z(t)$ is convex and uniformly B-Lipschitz (*i.e.* $\ell_z(\cdot)$ is B-lipschitz for all $z \in \mathcal{Z}$), which is the case for logistic regression or the support vector machines, there are very general *non-asymptotic* bounds for the excess risk by Sridharan, Shalev-Shwartz, and Srebro (2009), of the form :

$$\mathcal{R}(\widehat{f}_{n,\lambda}) \leq C \left(\frac{B^2 \log \frac{2}{\delta}}{\lambda n} + \lambda \|f_\rho\|_{\mathcal{H}}^2 \right), \tag{2.2}$$

with probability at least $1-\delta$, where C is an explicit constant (it does not depend on any parameters). The bound above is a bias-variance decomposition. Indeed, $B^2/(\lambda n)$ is a *variance term* which depends on the sample size n but not on the optimal predictor f_ρ, and $\lambda\|f_\rho\|_{\mathcal{H}}^2$ is a *bias term*, which depends on the optimal predictor but not on the sample size n. All our bounds will have this form but with smaller quantities (but asking for more assumptions). Without further assumptions, in Eq. (2.2), λ is taken proportional to $1/\sqrt{n}$, and we get the usual optimal slow rate in excess risk of $O(1/\sqrt{n})$ associated with such a general set-up (see, e.g., Cesa-Bianchi, Mansour, and Shamir, 2015).

2.1 Fast rates and algorithms for least-squares regression

In this section, we present a brief overview of the results which exists in the case of least-squares regression. In that case, $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and that is the case where $\ell_{(x,y)}(t) = \frac{1}{2}|t-y|^2$, as in Sec. 1.2.4.

In Sec. 2.1.1, we present the statistical aspects of least-squares regression, the bias variance trade offs for regularized empirical risk minimization, classical classes of distributions and associated upper and lower rates. We will also try to give hints as to the proof techniques and different quantities to control, as a similar methodology will be applied to prove oour results.

These bias-variance trade-offs are important not only for the purely statistical part, but will diffuse throughout the section as they provide a precision target for the optimziation routine (which does not need to go beyond the statistical precision), and also, through a notion of effective dimension, the dimension we must aim for when reducing the dimension of the empirical risk minimization problem.

In Sec. 2.1.2, we will describe the two main ways of performing dimension reduction on the empirical risk minimization problem Eq. (1.47) developped for least-squares.

Finally, in Sec. 2.1.3, we will present work by Rudi, Carratino, and Rosasco (2017), which develops a fast algorithm to solve this dimension reduced-problem, and which will be a sub-routine in our main work.

2.1.1 Statistics and rates for least squares

Let us first introduce some key linear operators in the context of least-squares regression. Some of them have already been introduced in Sec. 1.2.4.

Covariance operator. Recall from the definition of the operator $S : f \in \mathcal{H} \mapsto f \in L^2(\mathcal{X}, \rho)$ where S is just the injection from $\mathcal{H}$ to $L^2(\mathcal{X}, \rho_{\mathcal{X}})$, and the definition of the covariance operator

$$\Sigma = S^*S = \int_{\mathcal{X}} k_x \otimes k_x \, d\rho_{\mathcal{X}}, \qquad \Sigma f = \int_{\mathcal{X}} f(x) k_x \, d\rho_{\mathcal{X}}, \ f \in \mathcal{H}, \tag{1.54}$$

where for any u, v in a Hilbert space H, $u \otimes v$ denotes the operator defined by $(u \otimes v)w = \langle v, w \rangle_H u$. We can express the excess risk of a function f using only this operator in the case of the square loss (which is the quantity we wish to minimize) :

$$\mathcal{R}(f) - \mathcal{R}(f_\rho) = \tfrac{1}{2}\|\Sigma^{1/2}(f - f_\rho)\|_{\mathcal{H}}^2. \tag{2.3}$$

Empirical operators. Given the n samples $z_1, ..., z_n \sim Z$, we can define the analog finite dimensional operators $\widehat{S}_n : f \in \mathcal{H} \mapsto (f(x_i))_{1 \leq i \leq n} \in \mathbb{R}^n$, where $\mathbb{R}^n$ is equipped with the scalar product associated to the mean : $(\alpha|\beta) = \frac{1}{n}\sum_{i=1}^n \alpha_i \beta_i$. When equipped with that structure, $\widehat{S}_n^* : \alpha \mapsto \frac{1}{n}\sum_{i=1}^n \alpha_i k_{x_i}$ (recall that the adjoint satifies $(\alpha|\widehat{S}_n f) = \langle \widehat{S}_n^* \alpha, f \rangle_{\mathcal{H}}$). The empirical covariance is defined as

$$\widehat{\Sigma}_n = \widehat{S}_n^* \widehat{S}_n = \frac{1}{n}\sum_{i=1}^n k_{x_i} \otimes k_{x_i}, \tag{2.4}$$

and the regularized empirical risk can be written as $\frac{1}{2}\langle f, (\widehat{\Sigma}_n + \lambda \mathbf{I}) f \rangle_{\mathcal{H}} - \langle f, \widehat{S}_n^* y \rangle$ plus a constant term. The regularized empirical risk minimizer can therefore be expressed as $\widehat{f}_{n,\lambda} = (\widehat{\Sigma}_n + \lambda \mathbf{I})^{-1} \widehat{S}_n^* y$ where $y = (y_i) \in \mathbb{R}^n$.

Note that one of the great advantages of the least-squares setting is that we have closed form of the solutions using these linear operators. One of the main tools in order to generalize those results will be to be able to use such operators near the optimal solution.

Finer bias-variance trade offs

One of the quantities which has to be controlled is the noise, i.e. the quantity $Y - f_\rho(X)$. One way to do so is that the noise, conditionally on X, be uniformly sub-gaussian (where by uniformly, we mean that $\varepsilon|\{X = x\}$ is sub gaussian with parameters which are uniform on $x \in \mathcal{X}$). This assumption can be found as Hypothesis 2. in the work by Caponnetto and De Vito (2007), and as Assumption 2.9 in the work by Blanchard and Mücke (2018).

Assumption 2.4 (noise distribution)**.** *We make the following assumption on the noise* $\varepsilon = Y - f_\rho(X)$ *:*

$$\mathbb{E}\left[|\varepsilon|^m | X\right] \leq \frac{1}{2} m! \sigma^2 M^{m-2} \qquad \rho_{\mathcal{X}} \text{ a.s. }, \ m \geq 2. \tag{2.5}$$

As explained in Sec. 1.2.4, the bias-variance trade off in the least-square setting is expressed as a function of two main "kernel" quantities.

Definition 2.1 (Definition of the bias and the effective dimension)**.** *We define the two following key quantities :*

- *the effective dimension, defined as*

$$d_\lambda = \mathrm{Tr}((\Sigma + \lambda \mathbf{I})^{-1}\Sigma); \tag{2.6}$$

- *the bias, defined as*

$$b_\lambda = \lambda^2 \|(\Sigma + \lambda)^{-1/2} f_\rho\|_{\mathcal{H}}^2 = \|\Sigma^{1/2}(f_\lambda - f_\rho)\|_{\mathcal{H}}^2, \tag{2.7}$$

  where $f_\lambda = (\Sigma + \lambda)^{-1}\Sigma f_\rho$ *is the minimizer of the regularized expected risk.*

The following non-asymptotic bias-variance result can be obtained with these quantities (see Blanchard and Mücke (2018), Proposition 5.8 or Caponnetto and De Vito (2007), Theorem 4).

Theorem 2.1 (bias-variance trade-off for least squares)**.** *There exists two explicit constants C_1, C_2 such that for any $\delta \in (0,1]$ and $\lambda \in (0,1]$, if $n \geq C_1 \frac{\max(1,d_\lambda)}{\lambda} \log^2 \frac{4}{\delta}$, then the following bound holds with probability at least $1-\delta$:*

$$\|\Sigma^{1/2}(\widehat{f}_{n,\lambda} - f_\rho)\|_{\mathcal{H}} \leq C_2 \log(4/\delta) \left(\sqrt{b_\lambda} + \frac{M}{\lambda^{1/2} n} + \sqrt{\frac{\sigma^2 \, d_\lambda}{n}} \right). \tag{2.8}$$

This trade off can be obtained by using the following main ingredients. Recall that using the closed forms, it holds

$$\|\Sigma^{1/2}(\widehat{f}_{n,\lambda} - f_\rho)\|_{\mathcal{H}} = \|\Sigma^{1/2}((\widehat{\Sigma}_n + \lambda \mathbf{I})^{-1} \widehat{S}_n^* y - f_\rho)\|_{\mathcal{H}}. \tag{2.9}$$

Equivalence of the empirical and expected covariance. The first step is to guarantee that the regularized covariance and empirical covariance are equivalent, *i.e.*, $\frac{1}{2}\left(\widehat{\Sigma}_n + \lambda \mathbf{I}\right) \preceq \Sigma + \lambda \mathbf{I} \preceq 2\left(\widehat{\Sigma}_n + \lambda \mathbf{I}\right)$. This can be obtained with classical "dimensionless" concentration bounds (Tropp, 2012) with probability $1-\delta$ as soon as $n \geq C_1 \frac{\max(1,d_\lambda)}{\lambda} \log^2 \frac{2}{\delta}$ (hence this condition in the theorem). This equivalence allows to change from $(\Sigma + \lambda \mathbf{I})$ factors to $(\widehat{\Sigma}_n + \lambda \mathbf{I})$ factors by only multiplying by constants : in particular $\|(\Sigma + \lambda \mathbf{I})^{-1/2}(\widehat{\Sigma}_n + \lambda \mathbf{I})^{1/2}\| \leq \sqrt{2}$ and $\|(\Sigma + \lambda \mathbf{I})^{1/2}(\widehat{\Sigma}_n + \lambda \mathbf{I})^{-1/2}\| \leq \sqrt{2}$. With this, we can decompose the error Eq. (2.9) into two main terms (here C denotes an explicit constant which does not depend on any problem parameters) :

$$\|\Sigma^{1/2}(\widehat{f}_{n,\lambda} - f_\rho)\|_{\mathcal{H}} \leq C \left(\sqrt{b_\lambda} + \|(\Sigma + \lambda \mathbf{I})^{-1/2}(\widehat{S}_n^* y - \Sigma f_\rho)\| \right) \tag{2.10}$$

Concentration of the second term. The second main step is to concentrate the second term. For example, this is done by Blanchard and Mücke (2018). The way of doing this is simply to note that the second term can be written as

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \zeta_i - \mathbb{E}\left[\zeta\right] \right\|_{\mathcal{H}}, \qquad \zeta = (\Sigma + \lambda \mathbf{I})^{-1} Y k_X. \tag{2.11}$$

One can then use classical concentration bounds on sub-gaussian variables, by showing that ζ is sub-gaussian with parameters $\sigma^2 d_\lambda$ and $M/\sqrt{\lambda}$ using Eq. (2.5) (see Blanchard and Mücke (2018), proposition 5.2 for more details). This shows that with probability $1-\delta$, the right hand side of Eq. (2.10) is bounded by a term of the form $C' \log \frac{2}{\delta} \left(\frac{M}{\lambda^{1/2} n} + \sqrt{\frac{\sigma^2 \, d_\lambda}{n}} \right)$. Performing a union bound on the probabilistic results yields Theorem 2.1.

In chapter 3, we extend Theorem 2.1 using slightly different assumptions in the context of generalized self-concordant losses (in particular, for simplicity, we assume bounded noise as the noise is not as easily defined). These results are reported in Sec. 2.3.

Minimax upper bounds for least-squares. Recall from Sec. 1.1.2 that in order to define minimax rates, we have to define a class of test measures $\mathcal{M}$ for which we get the minimax rate for all measures $\rho \in \mathcal{M}$. To get upper minimax rates, we define the class $\mathcal{M}^<(R, r, \beta, b, \sigma, M)$ to be the set of Borel probability measures ρ on $\mathcal{X} \times \mathbb{R}$ such that Eq. (2.5) holds with σ and M, and the following assumptions hold for R, r, β, b.

Assumption 2.5 (Source condition)**.** *f_ρ satisfies a source condition of the type seen in Sec. 1.2.4 (see definitions before Eq. (1.57)) with $r \geq 1/2$, i.e., there exists $h \in \mathcal{H}$ such that $f_\rho = \Sigma^{r-1/2} h$ and $\|h\|_{\mathcal{H}} \leq R$. This corresponds to the hypothesis of f_ρ being in $\mathcal{H}^{2r} = \mathrm{Im}(T_k^r)$ mentioned in Sec. 1.2.4.*

Assumption 2.6 (Capacity condition)**.** *The eigenvalues λ_i of Σ satisfy $\lambda_i \leq \frac{\beta}{i^b}$, that is they decrease at least polynomially with order b. This shows that the effective dimension will explode as $d_\lambda \leq Q_{b,\beta}^2 \lambda^{-1/b}$, where $Q_{b,\beta}$ depends on b, β. In the literature (Rudi and Rosasco, 2017; Marteau-Ferey, Ostrovskii, Bach, and Rudi, 2019; Marteau-Ferey, Bach, and Rudi, 2019), the capacity condition is sometimes directly formulated as $d_\lambda \leq Q^2 \lambda^{-\gamma}$, expressed in terms of Q, γ instead of β, b.*

Let $\Theta = \{(R, \sigma, M) \in \mathbb{R}_{++}^3\}$ where $\mathbb{R}_{++}$ is the set of positive real numbers. For fixed values of r, β, b, we denote with $\mathcal{M}_\theta^<$ the class $\mathcal{M}^<(R, r, \beta, b, \sigma, M)$. Indeed, the minimax upper rates which can be obtained here are uniform over θ when the other parameters are fixed

We can now informally formulate the minimax rates obtained in this setting (it will be made formal in Theorem 2.2) : if we set

$$\lambda_{n,\theta} = \min\left(\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{b}{2br+1}}, 1\right), \qquad a_{n,\theta} = R^2 \left(\frac{\sigma^2}{R^2 n}\right)^{\frac{2br}{2br+1}}, \tag{2.12}$$

then it holds :

$$\mathcal{R}(\widehat{f}_{n,\lambda_{n,\theta}}) - \mathcal{R}(f_\rho) \leq a_{n,\theta} \text{ in the minimax sense on } \mathcal{M}_\theta. \tag{2.13}$$

To get an intuition of why that is the case, note that if we substitute the value of λ by that of $\lambda_{n,\theta}$ given in Eq. (2.12), we obtain a bound of the following form.

Corollary 2.1 (Blanchard and Mücke (2018), Cor. 5.9)**.** *Let $M > 0$, $\beta > 0$, $R > 0$, $r \in [1/2, 1]$, $b > 1$ and $\sigma > 0$. There exists constants $C_{\beta,b,r}$, $C_{\beta,b,r,\sigma,R}$ and $n_{\beta,b,r,\sigma,R,M}$ depending on the subscripted parameters such that for any $n \geq n_{\beta,b,r,\sigma,R,M}$, any $\delta \in (0, 1]$, it holds*

$$\frac{\mathcal{R}(\widehat{f}_{n,\lambda_{n,\theta}}) - \mathcal{R}(f_\rho)}{a_{n,\theta}} \leq C_{\beta,b,r} \log^2 \tfrac{4}{\delta} \tag{2.14}$$

provided $\log \delta^{-1} \leq C_{\beta,b,r,\sigma,R}\, n^{\frac{b(r-1/2)}{2br+1}}$

In the works by Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019); Marteau-Ferey, Bach, and Rudi (2019) which can be found in chapters 3 and 4, results of the type Cor. 2.1 are presented as fast rates. While using the same methodology as that of Caponnetto and De Vito (2007); Blanchard and Mücke (2018), we could get minimax upper bounds, this is not formally done. Let us now state such upper bounds and sketch very roughly the technique to go from Cor. 2.1 to minimax rates. We cite here a particular case of the result by Blanchard and Mücke (2018), Theorem 3.4.

Theorem 2.2 (Minimax upper rates)**.** *Let $r \in [1/2, 1], \beta > 0, b > 1$ be fixed, and assume $\lambda_{n,\theta}$ and $a_{n,\theta}$ are defined by Eq. (2.12) . For all $\psi(\cdot) = |\cdot|^p$ for $p > 0$, the following minimax rates of convergence hold with rate $a_{n,\theta}$:*

$$\sup_{\theta \in \Theta} \limsup_{n \to +\infty} \sup_{\rho \in \mathcal{M}_\theta^<} \mathbb{E}_{\rho^{\otimes n}} \left[\psi\left(\frac{\mathcal{R}(\widehat{f}_{n,\lambda_{n,\theta}}) - \mathcal{R}(f_\rho)}{r_{n,\theta}}\right)\right] < \infty \tag{2.15}$$

This shows that $a_{n,\theta}$ is a minimax upper rate of convergence for all L^p norms and uniformly in θ.

The rates obtained by Caponnetto and De Vito (2007) are a bit different, and quantify the probability that $\frac{\mathcal{R}(\widehat{f}_{n,\lambda_{n,\theta}})-\mathcal{R}(f_\rho)}{a_{n,\theta}} > \tau$ for all τ.

High level ideas to obtain minimax rates from Cor. 2.1. Reformulate Cor. 2.1 as the following : there exists constants $C, C_{\beta,b,r}, C_{\beta,b,r,\sigma,R}$ such that for $n \geq n_{\beta,b,r,\sigma,R,M}$ and all $\tau \in [0, \tau_0(n)]$,

$$\mathbb{P}\left(\frac{\mathcal{R}(\widehat{f}_{n,\lambda_{n,\theta}}) - \mathcal{R}(f_\rho)}{a_{n,\theta}} > \tau\right) \leq C \ \exp(-C_{\beta,b,r}\sqrt{\tau}), \text{where } \tau_0(n) = C_{\beta,b,r,\sigma,R} n^{\frac{2br-b}{2br+1}}. \tag{2.16}$$

This almost gives the desired bound : indeed, without the condition that $\tau \leq \tau_0(n)$, we would simply integrate the bound using the fact that for any differentiable ψ on $\mathbb{R}_+$ and Borel measurable random variable T, it holds $\mathbb{E}[\psi(T)] = \int_{\tau=0}^{+\infty} \psi'(\tau)\mathbb{P}(T > \tau)\, d\tau$. To handle the condition that $\tau \leq \tau_0(n)$, Blanchard and Mücke (2018) combine this bound which holds for all small enough τ with another bound which is rougher but which works for all $\tau > 0$. For more details, we refer directly to the original work.

Saturation effect and spectral regularization. Note that in Cor. 2.1 and Theorem 2.2, we have imposed the condition $r \leq 1$. This is due to the fact that there is a *saturation effect* with the Tikhonov regularization. Indeed, when using Tikhonov regularization, we cannot obtain better bias decrease than $b_\lambda \leq C\lambda^{\max(r,1)}$, and hence we do not leverage regularity beyond $r = 1$. It is possible to do so, using techniques from spectral filtering Gerfo et al. (2008); Blanchard and Mücke (2018). Spectral filtering consists in regularizing with a spectral regularization function $g : (t, \lambda) \in [0, 1] \times \mathbb{R} \to \mathbb{R}$, which defines a regularization $\widehat{f}_{n,\lambda} = g(\widehat{\Sigma}_n, \lambda)\widehat{S}_n^* y$ where $g(\cdot, \lambda)$ is applied spectrally. Tikhonov regularization corresponds to $g(t, \lambda) = \frac{1}{t+\lambda}$, but other regularizations exists (such as iterated Tikhonov, King and Chillingworth (1979)) which have a better *qualification* (for a precise definition see Gerfo, Rosasco, Odone, Vito, and Verri (2008); Blanchard and Mücke (2018); Beugnot, Mairal, and Rudi (2021)). Under the assumption that the spectral filter g has better qualification than r, one can go beyond the previous setting and bound the bias term by $C\lambda^r$, and obtain the same minimax upper bounds as the ones stated above.

Note that the work by Blanchard and Mücke (2018) goes beyond the setting presented above, taking into account many spectral regularizations, as well as obtaining minimax rates for other metrics that the excess risk, and taking into account the inverse problem setting.

Minimax lower bounds. The results obtained in this thesis do not concern minimax lower bounds : indeed, the methodology for obtaining such bounds is very different from the one to obtain upper bounds, and does not involve constructing an estimator (like the regularized e.r.m.). In this paragraph, we give a few elements in order to understand the definitions and high level strategies in order to obtain lower bounds for least squares. Throughout this paragraph, we refer to the works by Blanchard and Mücke (2018); Caponnetto and De Vito (2007) who obtain minimax lower bounds which essentially match those of Theorem 2.2, and hence shows that the rate $a_{\theta,n}$ defined above is optimal.

Classes. Define the class $\mathcal{M}(R, r, \nu, \sigma, M)$ (or $\mathcal{M}(\nu)$ for short) to be the set of Borel probability measures ρ on $\mathcal{X} \times \mathbb{R}$ such that the following conditions hold.

(i) *Marginal* $\rho_{\mathcal{X}} = \nu$.

(ii) *Source condition* f_ρ satisfies a source condition of the type seen in Sec. 1.2.4 (see definitions before Eq. (1.57)) with $r \geq 1/2$, *i.e.*, there exists $h \in \mathcal{H}$ such that $f_\rho = \Sigma^{r-1/2}h$ and $\|h\|_{\mathcal{H}} \leq R$.

(iii) *Assumption on the noise.* Eq. (2.5) holds with σ and M.

Note that $\mathcal{M}^<(R, r, b, \beta, \sigma, M)$ defined for the minimax upper bounds is just the union of all $\mathcal{M}(R, r, \nu, \sigma, M)$ over the $\nu \in \mathcal{P}^<(b, \beta)$, *i.e.*, such that the capacity condition holds : the eigenvalues λ_i of Σ (whis is defined by ν) satisfy $\lambda_i \leq \frac{\beta}{i^b}$.

To obtain minimax lower bounds, Blanchard and Mücke (2018) instead consider the set $\mathcal{P}^>(b, \alpha)$ of probability distributions on $\mathcal{X}$ such that the capacity is lower bounded : the eigenvalues λ_i of Σ satisfy $\lambda_i \geq \frac{\alpha}{i^b}$. Moreover, they add a technical condition in order to obtain *strong minimax lower rates* (see the definition of $\mathcal{P}^>_{strong}$ theirein). This technical assumption is satisfied as soon as $\nu \in \mathcal{P}^<(b, \beta)$ for some $\beta > \alpha$. The class studied for minimax lower bounds is the union $\mathcal{M}^>(R, r, b, \alpha, \sigma, M) = \bigcup_{\nu \in \mathcal{P}^>(b,\alpha)} \mathcal{M}(R, r, \nu, \sigma, M)$. In the same way as in the upper bound case, we consider the parameter $\theta \in \Theta = \{(R, \sigma, M) \in \mathbb{R}^3_{++} \ : \ \sigma \leq M\}$ as being free, i.e. $\mathcal{M}_\theta(\nu)$ resp. $\mathcal{M}^>_\theta$ denote the previous models for fixed ν, r resp. r, α, b. As explained in Sec. 1.1.2, in order to obtain minimax lower bounds, we must lower bound the following quantity asymptotically :

$$\mathrm{MinMax}(\psi, \mathcal{H}, n, \theta, \mathcal{M}^>_\theta, a_{n,\theta}) := \inf_{\widehat{f}_n} \sup_{\rho \in \mathcal{M}^>_\theta} \mathbb{E}_{\rho^{\otimes n}} \left[\psi \left(\frac{\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_\rho)}{a_{n,\theta}} \right) \right], \tag{2.17}$$

where the $\widehat{f}_n$ are minimized on the set of estimators $\widehat{f}_n : \mathcal{Z}^n \to \mathcal{H}$. Blanchard and Mücke (2018), prove the following theorem on minimax lower rates for the L^p norm.

Theorem 2.3 (Blanchard and Mücke (2018))**.** *Let $r \geq 0, \alpha > 0, b > 1$ be fixed, and assume $a_{n,\theta}$ is defined by Eq. (2.12) . For all $\psi(\cdot) = |\cdot|^p$ for $p > 0$, the following minimax rates of convergence hold with rate $a_{n,\theta}$:*

$$\inf_{\theta \in \Theta} \liminf_{n \to +\infty} \mathrm{MinMax}(\psi, \mathcal{H}, n, \theta, \mathcal{M}^>_\theta, a_{n,\theta}) > 0 \tag{2.18}$$

This shows that $a_{n,\theta}$ is a minimax lower rate of convergence for all L^p norms and uniformly in θ.

We now very briefly describe the proof techniques in order to get such rates. Fix α, b, r.

Reduction to a single ν The first step is to note that since

$$\mathcal{M}^>(R, r, b, \alpha, \sigma, M) = \bigcup_{\nu \in \mathcal{P}^>(b,\alpha)} \mathcal{M}(R, r, \nu, \sigma, M), \tag{2.19}$$

one can decompose the minimax along the $\nu \in \mathcal{P}^>(b, \alpha)$ to get a lower bound :

$$\mathrm{MinMax}(\psi, \mathcal{H}, n, \theta, \mathcal{M}^>_\theta, a_{n,\theta}) \geq \sup_{\nu \in \mathcal{P}^>(b,\alpha)} \mathrm{MinMax}(\psi, \mathcal{H}, n, \theta, \mathcal{M}_\theta(\nu), a_{n,\theta}). \tag{2.20}$$

Reduction to a bound in probability over a finite set. Note that for a fixed $\nu \in \mathcal{P}^>(b, \alpha)$, the risk $\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_\rho)$ can be expressed as a squared distance on $\mathcal{H}$ defined by ν :

$$\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_\rho) = d_\nu(\widehat{f}_n, f_\rho)^2 \qquad d_\nu(f, f') := \|\Sigma_\nu^{1/2}(f - f')\|_{\mathcal{H}}, \tag{2.21}$$

where we have highlighted by Σ_ν that the covariance operator Σ depends only on ν. Hence, using the Markov property, for any $\varepsilon > 0$, it holds

$$\begin{aligned}\mathrm{MinMax}(\psi, \mathcal{H}, n, \theta, \mathcal{M}_\theta(\nu), a_{n,\theta}) &\geq \psi(\varepsilon^2/a_{n,\theta}) \inf_{\widehat{f}_n} \sup_{\rho \in \mathcal{M}_\theta(\nu)} \mathbb{P}_{\rho^{\otimes n}} \left(\left(\mathcal{R}(\widehat{f}_n) - \mathcal{R}(f_\rho) \right) > \varepsilon \right), \\ &\geq \psi(\varepsilon^2/a_{n,\theta}) \inf_{\widehat{f}_n} \sup_{\rho \in \mathcal{M}_\theta(\nu)} \mathbb{P}_{\rho^{\otimes n}} \left(d_\nu(\widehat{f}_n, f_\rho) > \varepsilon \right).\end{aligned} \tag{2.22}$$

Applying the Tskybakov method. We now concentrate on bounding the term

$$\inf_{\widehat{f}_n} \sup_{\rho \in \mathcal{M}_\theta(\nu)} \mathbb{P}_{\rho^{\otimes n}} \left(d_\nu(\widehat{f}_n, f_\rho) > \varepsilon \right).$$

This term can be bounded using a generic method presented by Tsybakov (2008) in Chapter 2. Applied to this setting, the idea is that one can lower bound

$$\inf_{\widehat{f}_n} \sup_{\rho \in \mathcal{M}_\theta(\nu)} \mathbb{P}_{\rho^{\otimes n}} \left(d_\nu(\widehat{f}_n, f_\rho) > \varepsilon \right) \geq \inf_{\widehat{f}_n} \sup_{1 \leq i \leq N} \mathbb{P}_{\rho_i^{\otimes n}} \left(d_\nu(\widehat{f}_n, f_{\rho_i}) > \varepsilon \right) \tag{2.23}$$

for well-chosen N and $\rho_0, ..., \rho_N$ (which have to be adapted to θ, ε and n in this setting). These measures $\rho_0, ..., \rho_N$ have to be designed as being 2ε-separated for the distance d_ν, and have "small" KL-divergence $\mathrm{KL}(\rho|\rho') = \mathbb{E}_\rho \left[\log \frac{d\rho}{d\rho'} \right]$ between them. This basically says that the ρ_i will be statistically indistinguishable while being far from each other in d_ν distance, and hence the best estimator will not be able to choose between them with precision better than ε. More formally, we can set the following requirements, which are adapted from Thm. 2.5 by Tsybakov (2008) wihch is itself based on Fano's lemma, in the present setting.

Proposition 2.1 (Tsybakov method)**.** *Fix $\theta \in \Theta$. Let $\varepsilon > 0$. Assume that there exists $N_\varepsilon > 2$, $\omega_\varepsilon \geq 0$ and $\rho_0, ..., \rho_{N_\varepsilon} \in \mathcal{M}_\theta(\nu)$ such that*

(i) For any $0 \leq i < j \leq N_\varepsilon$, $d_\nu(f_{\rho_i} f_{\rho_j}) \geq 2\varepsilon$;

(ii) For any $j = 1, ..., N_\varepsilon$, ρ_j is absolutely continuous with respect to ρ_0, and

$$\frac{1}{N_\varepsilon} \sum_{i=1}^{N} \mathrm{KL}(\rho_j|\rho_0) \leq \omega_\varepsilon \log(N_\varepsilon). \tag{2.24}$$

Then for all n such that $0 \leq n\omega_\varepsilon \leq 1/8$, it holds

$$\begin{aligned}\inf_{\widehat{f}_n} \sup_{\rho \in \mathcal{M}_\theta(\nu)} \mathbb{P}_{\rho^{\otimes n}} \left(d_\nu(\widehat{f}_n, f_\rho) > \varepsilon \right) &\geq \inf_{\widehat{f}_n} \sup_{0 \leq i \leq N_\varepsilon} \mathbb{P}_{\rho_i^{\otimes n}} \left(d_\nu(\widehat{f}_n, f_{\rho_i}) > \varepsilon \right) \\ &\geq \frac{\sqrt{N_\varepsilon}}{1 + \sqrt{N_\varepsilon}} \left(1 - 2n\omega_\varepsilon - \sqrt{\frac{2n\omega_\varepsilon}{\log(N_\varepsilon)}} \right) \\ &\geq \underbrace{\frac{1}{4} \left(\frac{3}{2} - \sqrt{\frac{1}{\log(2)}} \right)}_{:=c} > 0.\end{aligned} \tag{2.25}$$

Constructing the ρ_i to apply Tsybakov's method. The crux of the problem is therefore to design the $\rho_0, ..., \rho_{N_\varepsilon}$ in order to apply the above proposition. Both Caponnetto and De Vito (2007); Blanchard and Mücke (2018) proceed in essentially the same manner. They design the ρ_i in the form $\rho_i(dx, dy) = \rho_{f_i}(dy|x)\nu(dx)$ where $\rho_{f_i}(dy|x) = \mathcal{N}(f_i(x), \sigma^2)$, that is ρ_i is gaussian gaussian conditionally on x. It is clear the $f_{\rho_i} = f_i$, and a bound on the KL-divergence can easily be obtained in therms of the distance $d_\nu(f_i, f_j)$ and σ. Based on a technical lemma from Caponnetto and De Vito (2007), Proposition 6.4 by Blanchard and Mücke (2018) give a constant $C_{\alpha,b,r}$ and ε_0 depending on $\alpha, b, r, \sigma, M, R$ such that for all $\varepsilon < \varepsilon_0$, one can construct ρ_i, $1 \leq i \leq N_\varepsilon$ satisfying conditions *(i)* and *(ii)* with $\omega_\varepsilon = C_{\alpha,b,r}\sigma^{-2}R^2(\varepsilon R^{-1})^{2+1/br}$.

Bringing everything together. The final step is setting $\varepsilon_n = D_{\alpha,b,r}R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{br}{2br+1}}$, where $D_{\alpha,b,r}$ is set to satisfy $n\omega_\varepsilon \leq 1/8$, and such that $\varepsilon_n \leq \varepsilon_0$, as soon as $n \geq n_0$ where n_0 depends on $\sigma, R, M, r, \beta, b$. We can then go back up the chain of inequalities : combining Eq. (2.25) and Eq. (2.22) as well as Eq. (2.12) for the definition of $a_{n,\theta}$, it holds that for all $n \geq n_0$,

$$\mathrm{MinMax}(\psi, \mathcal{H}, n, \theta, \mathcal{M}_\theta(\nu), a_{n,\theta}) \geq \psi(D^2_{\alpha,b,r})c.$$

Theorem 2.3 then immediately follows from Eq. (2.20) and going to the limit.

Note that the work by Blanchard and Mücke (2018) goes beyond the setting presented above, dealing with other distances d_ν as well as problems coming from inverse problems.

Brief recap. In Sec. 2.1.1, we give the main statistical trade-offs, as well as upper and lower minimax rates for least squares, as well as the high level methodology to prove these results. We will use similar a very similar methodology in Sec. 2.3 to prove our bias variance trade offs and upper rates (not exactly minimax but which could be made into minimax upper rates). These upper rates, in the least squares setting, are actually optimal for a well-chosen class of functions, as corresponding lower bounds can be proven in the well specified setting (that is when $f_\rho \in \mathcal{H}$).

2.1.2 Dimension reduction with the effective dimension

In Sec. 2.1.1, we have presented the classical statistical rates, and bias variance trade offs for the regularized empirical risk minimizer $\widehat{f}_\lambda$. Recall that *a priori*, for $\lambda > 0$, $\widehat{f}_\lambda$ exists and is the unique solution to

$$\widehat{f}_{n,\lambda} = \operatorname*{arg\,min}_{f \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - y_i|^2 + \lambda \|f\|^2_{\mathcal{H}}. \tag{1.47}$$

The representer theorem (Theorem 1.2) allows to guarantee that $\widehat{f}_{n,\lambda}$ automatically belongs to the set $\mathcal{H}_n = \mathrm{span}(\{k_{x_i} \ : \ 1 \leq i \leq n\})$ which is n dimensional and can be parametrized by $\alpha \in \mathbb{R}^n \longmapsto \sum_{i=1}^n \alpha_i k_{x_i} \in \mathcal{H}_n$, leading to the n-dimensional problem

$$\widehat{\alpha}_{n,\lambda} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n}\|\mathbf{K}\alpha - y\|^2 + \lambda \alpha^\top \mathbf{K}\alpha. \tag{1.50}$$

where $\mathbf{K} = (k(x_i, x_j))_{1 \leq i,j \leq n}$ is called the kernel matrix. The fact that the problem Eq. (1.50) is n-dimensional is computationally limiting : computing the full matrix $\mathbf{K}$ already has a complexity

of order $O(n^2)$, and computing gradients of the above functional would cost $O(n^2)$ per gradient. Moreover, solving the linear system would take $O(n^3)$ (indeed, $\widehat{\alpha}_{n,\lambda} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}y$). Either way, we seem not to be able to escape a cost of $O(n^2)$ if not $O(n^3)$, which is prohibitive in the settings we wish to consider (i.e. n in the order of millions or billions of points).

The goal of dimension reduction is to find a low dimensional set of functions $\mathcal{H}_m$ of dimension m, equipped with a RKHS structure, such that replacing $\mathcal{H}$ by $\mathcal{H}_m$ in Eq. (1.47) does not cost anything statistically. If $\widehat{f}_{n,\lambda,m}$ is the solution to

$$\widehat{f}_{n,\lambda,m} = \operatorname*{arg\,min}_{f\in\mathcal{H}_m} \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - y_i|^2 + \lambda\|f\|_{\mathcal{H}}^2, \tag{2.26}$$

where we index the result with the dimension m, we want to guarantee that $\mathcal{R}(\widehat{f}_{n,\lambda,m}) - \mathcal{R}(f_\rho) \leq C(\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}(f_\rho))$, ideally with $m \ll n$. Such results are usually proved in two forms. The first, form is to ask that as soon as $m \geq m_\lambda$, the bias-variance decomposition of Theorem 2.1 holds not only for $\widehat{f}_{n,\lambda}$ but also for $\widehat{f}_{n,\lambda,m}$. We therefore have to design m_λ in order for the following decomposition to hold.

Theorem 2.4 (Prototypical dimension reduced bias-variance trade off)**.** *There exists two explicit constants C_1, C_2 such that for any $\delta \in (0,1]$ and $\lambda \in (0,1]$, if $n \geq C_1 \frac{\max(1,d_\lambda)}{\lambda} \log^2 \frac{4}{\delta}$ and $m \geq m_\lambda$ then the following bound holds with probability at least $1-\delta$:*

$$\|\Sigma^{1/2}(\widehat{f}_{n,\lambda,m} - f_\rho)\|_{\mathcal{H}} \leq C_2 \log(4/\delta)\left(\sqrt{b_\lambda} + \frac{M}{\lambda^{1/2}n} + \sqrt{\frac{\sigma^2\, d_\lambda}{n}}\right). \tag{2.27}$$

As in the previous section, such a bias-variance decomposition can also be used to derive upper rates of convergence for certain classes of function. Recall from Sec. 2.1.1 the definition of the class $\mathcal{M}^<(R,r,\beta,b,\sigma,M)$ of distributions ρ, which we shorthand with $\mathcal{M}_\theta^<$ when r,β,b are fixed and $\theta \in \Theta = \{(R,\sigma,M) \in \mathbb{R}^3_{++} \ : \ \sigma \leq M\}$. Combining Cor. 2.1 with Theorem 2.4, setting $m_{\theta,n} = m_{\lambda_{\theta,n}}$, it is possible to show an upper rate of convergence for the dimension reduced problem for the class $\mathcal{M}_\theta^<$.

Theorem 2.5 (Prototypical upper rates with dimension reduction)**.** *Fix $r \in [1/2,1], \beta, b > 0$. There exists a constant C such that for all $\theta \in \Theta$, there exist n_0, n_1 depending on θ such that for any $n \geq n_0$, any $\delta \in (0,1]$, and any $m \geq m_{\theta,n}$, it holds*

$$\mathcal{R}(\widehat{f}_{n,\lambda_{n,\theta},m_{n,\theta}}) - \mathcal{R}(f_\rho) \leq C \log^2\tfrac{4}{\delta}\ R^2\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{2br}{2br+1}}, \qquad \rho \in \mathcal{M}_\theta^<, \tag{2.28}$$

provided $\log \delta^{-1} \leq (n/n_1)^{\frac{b(r-1/2)}{2br+1}}$ and where $\lambda_{n,\theta}$ is given by Eq. (2.12).

This second form of result is usually very useful to get an idea on how hard it is to approximate $\mathcal{H}$ with $\mathcal{H}_m$ depending on the number of samples n and the hardness of the problem (parametrized by the class $\mathcal{M}_\theta^<$. Typically, the ideal dimesnion would be $m_{n,\theta} = d_{\lambda_{n,\theta}} = C_{\beta,b,r}\left(\frac{\sigma^2}{R^2 n}\right)^{-1/(2br+1)}$, which is, in a sense, the true dimension of the space $\mathcal{H}$ regularized with $\lambda_{n,\theta}$. However, this is not always possible, depending on the way we construct the space $\mathcal{H}_m$.

We will detail two ways of constructing spaces $\mathcal{H}_m$ in the literature : random features and Nyström projections.

Random features

As explained in Sec. 1.2.5 random features are based on approximating the RKHS $\mathcal{H}$ defined by a kernel whose expression is of the form $k(x,x') = \mathbb{E}_{\omega\sim\pi}[\phi(x,\omega)\phi(x',\omega)]$, that as an expectation of random features (see Eq. (1.43) and the following discussion on the Gaussian kernel). For the sake of simplicity, we will assume in this section that the features are bounded by 1; in particular, this implies that the kernel is bounded by 1.

The idea of using random features to approximate the RKHS $\mathcal{H}$ is to approximate it with the set

$$\mathcal{H}_m := \left\{ x \mapsto \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \alpha_j \phi(x,\omega_j) \ : \ \alpha \in \mathbb{R}^m \right\}, \tag{2.29}$$

where the ω_j are m i.i.d. samples from π. $\mathcal{H}_m$ can be equipped with the RKHS structure given by $\langle \frac{1}{\sqrt{,m}} \sum_{j=1}^m \alpha_j \phi(\cdot,\omega_j), \frac{1}{\sqrt{m}} \sum_{j=1}^m \beta_j \phi(\cdot,\omega_j)\rangle_{\mathcal{H}_m} = \sum_{i=1}^m \alpha_j\beta_j$, associated to the kernel $k_m(x,x') = \frac{1}{m}\sum_{j=1}^m \phi(x,\omega_j)\phi(x',\omega_j)$.

This rather abstract formulation can also simply be seen as replacing the kernel matrix $\mathbf{K}$ in Eq. (1.50) with a low rank approximation $\mathbf{K} \approx \Phi^\top\Phi$ where $\Phi^\top = \frac{1}{\sqrt{m}}(\phi(x_i,\omega_j))_{i,j} \in \mathbb{R}^{n\times m}$. Indeed, using the parametrization in Eq. (2.29), Eq. (2.26) becomes

$$\min_{\alpha\in\mathbb{R}^m} \tfrac{1}{n}\|\Phi^\top\alpha - y\|_{\mathbb{R}^n}^2 + \lambda\|\alpha\|^2, \tag{2.30}$$

whose solution is $(\Phi\Phi^\top + n\lambda\mathbf{I})^{-1}\Phi y$.

In Rudi and Rosasco (2017), theorems of the form Theorems 2.4 and 2.5 are proved and give lower bounds on m in order to obtain the same performance as that of the classical empirical risk minimizer. For example, Theorem 2. from Rudi and Rosasco (2017) can be adapted to show the following.

Theorem 2.6 (Rates for random features)**.** *Fix $r \in [1/2,1], \beta > 0, b > 1$. There exists C and m_0 such that for all $\theta \in \Theta$, there exists n_0, n_1 depending on θ such that for any $n \geq n_0$, for any $\delta \in (0,1]$ if $m \geq m_{\theta,n} \log\frac{108n}{\delta}$ where*

$$m_{\theta,n} = m_0 \left(\frac{\sigma^2 n}{R^2}\right)^{\frac{b+2r-1}{2br+1}} \tag{2.31}$$

and $\log\delta^{-1} \leq (n/n_1)^{\frac{b(r-1/2)}{2br+1}}$, it holds with probability at least $1-\delta$:

$$\mathcal{R}(\widehat{f}_{n,\lambda_{\theta,n},m}) - \mathcal{R}(f_\rho) \leq CR^2 \left(\frac{\sigma^2 n}{R^2}\right)^{\frac{2br}{2br+1}}. \tag{2.32}$$

Note that in the situation where $b \approx 1$ and $r \approx 1/2$ (that is the loosest assumptions), typically $m \approx \sqrt{n}$ features are needed, reducing the dimension by a significant exponent. Note that this result is proved from a more complex bias-variance decomposition such as that of Theorem 2.4, in Theorem 5 by Rudi and Rosasco (2017).

Note that this result does not match the "ideal" dimension m w would get as $m_{n,\theta} = d_{\lambda_{n,\theta}}$. This will be explained in part in the following discussion, which details the high level techniques and principles to obtain Theorem 2.6.

Sketch of the techniques involved. In order to compare the space $\mathcal{H}_m$ with the space $\mathcal{H}$, one compares the associated kernel operators on $L^2(\mathcal{X}\rho_{\mathcal{X}})$, defined for a generic kernel k by Eq. (1.56) as $(T_k g)(x) = \int_{\mathcal{X}} k(x,x')g(x')\rho(dx')$. Let T_m be a short-hand for T_{k_m} and T for T_k.

There are main two terms depending on m and which have to be controlled in order to get the same precision with $\mathcal{H}_m$ as with $\mathcal{H}$. They are linked to two guarantees which must be satisfied.

(i) The fact that the metric of $\mathcal{H}_m$ and $\mathcal{H}$ are essentially the same, at fixed λ. This can be formalized as the operators $T + \lambda\mathbf{I}$ and $T_m + \lambda\mathbf{I}$ being equivalent.

(ii) The fact that we are approximating a projected problem leads to a satisfactory solution, i.e. controlling the term $\|(T_m + \lambda\mathbf{I})^{-1}T_m f_\rho - (T + \lambda\mathbf{I})^{-1}Tf_\rho\|$, which is the difference in performance between the regularized expected minimizer on $\mathcal{H}_m$ and $\mathcal{H}$.

These terms can be controlled using the key quantity

$$\mathcal{F}_\infty(\lambda) = \sup_{\omega\in\mathrm{supp}(\pi)} \|(T + \lambda\mathbf{I})^{-1/2}\phi(\cdot,\omega)\|^2_{L^2(\mathcal{X},\rho_{\mathcal{X}})}. \tag{2.33}$$

Using concentration bounds, it is possible to show that (i) holds as soon as $m \geq \mathcal{F}_\infty(\lambda)$ up to logarithmic terms. Contronlling (ii) is a bit more involved, and the lower bound for m to guarantee that a bias-variance decomposition of the form Theorem 2.4 holds depends on $\mathcal{F}_\infty(\lambda)$ but also on the effective dimension d_λ as well as the bias term.

Reducing the dimension by adapting the sampling to the problem. Note that $s_\lambda(\omega) = \|(T + \lambda\mathbf{I})^{-1/2}\phi(\cdot,\omega)\|^2_{L^2(\mathcal{X},\rho_{\mathcal{X}})}$ measures the *score* of ω, that is the contribution of ω to the approximation of T by T_m. $\mathcal{F}_\infty$ is therefore the highest possible score, and imposing $m \geq \mathcal{F}_\infty(\lambda)$ essentially says that to lower bound m, we pay the price of having to see the ω with the highest score. To obtain a lower bound on m which is less conservative, it is natural to sample not from π, but from a density which favors ω with high scores, and therefore is adapted to the problem.

Ideally, we would like to use reweighted features and sample ω from the distribution

$$\pi_\lambda(d\omega) = \frac{s_\lambda(\omega)\pi(d\omega)}{d_\lambda}, \tag{2.34}$$

which is a probability density (d_λ is the renormalization term). If we use this measure to sample the ω, it is possible to show that $\mathcal{F}_\infty = d_\lambda$, and hence that we can better adapt to the effective dimension, *i.e.*, we would obtain $m_{\theta,n} = m_0\left(\sigma^2 n/R^2\right)^{\frac{2br-b+1}{2br+1}}$, which is lower than the one obtained before as $b > 1$, but still not as good as the ideal $d_{\lambda_{\theta,n}}$.

Of course, we do not have access to this ideal measure π_λ, as it is defined using problem dependent quantities. However, Rudi and Rosasco (2017) believe it is possible to approximate this measure well enough in order to obtain the same rates, even though no formal result has been published on the subject yet.

Nyström sampling

In the case of Nyström sampling, the set $\mathcal{H}$ is approximated with a set of the form

$$\mathcal{H}_m = \left\{ x \mapsto \sum_{j=1}^{m} \alpha_j k(x, \widetilde{x}_j) \ : \ \alpha \in \mathbb{R}^m \right\}, \tag{2.35}$$

where $\widetilde{x}_1, ..., \widetilde{x}_m \in \mathcal{X}$. Usually, the Nyström points are sampled from from the $x_1, ..., x_n$, as we will see below. This set can simply be seen as a m dimensional subspace of $\mathcal{H}$ with the induced norm : $\langle \sum_{j=1}^m \alpha_j k(\cdot, \widetilde{x}_j), \sum_{j=1}^m \beta_j k(x, \widetilde{x}_j) \rangle_{\mathcal{H}_m} = \alpha^\top \mathbf{K}_{mm} \beta$ where $\mathbf{K}_{mm}$ is the kernel matrix associated to the $\widetilde{x}_j$. The kernel associated to the space $\mathcal{H}$ can be $k_m(x, x') = \mathbf{k}_m(x)^\top \mathbf{K}_{mm}^\dagger \mathbf{k}_m(x)$ where $\mathbf{k}_m(x) = (k(\widetilde{x}_j, x))_{1 \leq i \leq m} \in \mathbb{R}^m$ and $\dagger$ denotes the Moore-Penrose pseudoinverse. Using the parametrization of $\mathcal{H}_m$ in Eq. (2.35), the problem Eq. (2.26) becomes:

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \|\mathbf{K}_{n,m}\alpha - y\| + \lambda \alpha^\top \mathbf{K}_{mm} \alpha, \qquad \mathbf{K}_{n,m} = (k(x_i, \widetilde{x}_j))_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}. \tag{2.36}$$

Eq. (2.36) shows quite clearly why Nyström is also sometimes called column subsampling. Indeed, if the $\widetilde{x}_1, ..., \widetilde{x}_m$ are sampled from the $x_1, ..., x_n$, the matrix $\mathbf{K}_{n,m}$ consists in subsampled columns of the full matrix $\mathbf{K}$, and therefore Eq. (2.36) is a subsampled version of Eq. (1.50). Note in particular that if we take the Nyström points to be $x_1, .., x_n$, we recover the standard empirical risk minimization problem. The following reparametrization is therefore also valid in that setting.

A much better way parametrizing is to write $\mathbf{K}_{mm} = \mathbf{T}^\top \mathbf{T}$ using a cholesky decomposition, where $\mathbf{T}$ is upper triangular matrix. One then notess that defining $\phi(x) = \mathbf{T}^{-\top} \mathbf{k}_m(x) \in \mathbb{R}^m$, ϕ is a feature map which defines the kernel k_m : $k_m(x, x') = \phi(x)^\top \phi(x')$. Hence, $\mathcal{H}_m = \{\alpha^\top \phi(\cdot) \ : \ \alpha \in \mathbb{R}^m\}$. If Φ is the matrix whose columns are the $\phi(x_i)$, that is $\Phi^\top = \mathbf{K}_{n,m} \mathbf{T}^{-1}$, the problem Eq. (2.26) can be formulated as

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{n} \|\Phi^\top \alpha - y\|_{\mathbb{R}^n}^2 + \lambda \|\alpha\|^2. \tag{2.37}$$

This is exactly the form of Eq. (2.30). Of course, this comes at the price of computing and saving a cholesky factor $\mathbf{T}$, which is $O(m^3)$ in time and $O(m^2)$ in memory.

In Rudi, Camoriano, and Rosasco (2015), theorems of the form Theorems 2.4 and 2.5 are proved and give lower bounds on m in order to obtain the same performance as that of the classical empirical risk minimizer. For example, Theorem 1. from Rudi, Camoriano, and Rosasco (2015) can be adapted to show the following in the case where the points $\widetilde{x}_1, ..., \widetilde{x}_m$ are sampled uniformly from $\rho_{\mathcal{X}}$ (to obtain such points, we sample m points uniformly without replacement from $x_1, ..., x_n$).

Theorem 2.7 (Rates for uniform Nyström)**.** *Fix $r \in [1/2, 1], \beta > 0, b > 1$. There exists C and m_0 such that for all $\theta \in \Theta$, there exists n_0, n_1 depending on θ such that for any $n \geq n_0$, for any $\delta \in (0, 1]$ if $m \geq m_{\theta,n} \log \frac{108n}{\delta}$ where*

$$m_{\theta,n} = m_0 \left(\frac{\sigma^2 n}{R^2} \right)^{\frac{b}{2br+1}} \tag{2.38}$$

and $\log \delta^{-1} \leq (n/n_1)^{\frac{b(r-1/2)}{2br+1}}$, *if the* m *Nyström points are sampled uniformly from* $\rho_{\mathcal{X}}(dx)$, *it holds with probability at least* $1-\delta$ *:*

$$\mathcal{R}(\widehat{f}_{n,\lambda_{\theta,n},m}) - \mathcal{R}(f_\rho) \leq CR^2 \left(\frac{\sigma^2 n}{R^2}\right)^{\frac{2br}{2br+1}}. \tag{2.39}$$

Note that in the situation where $b \approx 1$ and $r \approx 1/2$ (that is the loosest assumptions), typically $m \approx \sqrt{n}$ features are needed, reducing the dimension by a significant exponent.

Note that as in the random features case, this result does not match the "ideal" dimension m w would get as $m_{n,\theta} = d_{\lambda_{n,\theta}}$. In the case of Nyström points, this can be corrected. This will be explained in part in the following discussion.

High level techniques. The analysis is somewhat simpler in the setting of Nyström points, as the $\mathcal{H}_m$ are sub-spaces of $\mathcal{H}$ can can be characterized by the associated orthogonal projection P_m from $\mathcal{H}$ onto $\mathcal{H}_m$. Therefore, there is no need to go to $L^2(\mathcal{X}, \rho_{\mathcal{X}})$, and operators are therefore only compared as operators on $\mathcal{H}$. Let $\Sigma_m = P_m \Sigma P_m$ be the covariance operator of $\mathcal{H}_m$ (but defined on the whole of $\mathcal{H}$ for convenience through the projection).

There are main two terms depending on m and which have to be controlled in order to get the same precision with $\mathcal{H}_m$ as with $\mathcal{H}$. They are linked to two guarantees which must be satisfied.

(i) The fact that the metric of $\mathcal{H}_m$ and $\mathcal{H}$ are essentially the same, at fixed λ. This can be formalized as the operators $\Sigma_m + \lambda \mathbf{I}$ and $\Sigma + \lambda \mathbf{I}$ being equivalent.

(ii) The fact that we are approximating a projected problem leads to a satisfactory solution, i.e. controlling the term $\|\Sigma^{1/2}(\Sigma_m + \lambda \mathbf{I})^{-1}\Sigma_m f_\rho - (\Sigma + \lambda \mathbf{I})^{-1}\Sigma f_\rho\|$, which is the difference in performance between the retularized expected minmimaxer on $\mathcal{H}_m$ and $\mathcal{H}$.

Note that both these terms can actually be controlled using only the projection, through the quantity $\|\Sigma^{1/2}(I - P_m)\|$, which quantifies the information Σ contained that we omit when projecting on $\mathcal{H}_m$.

In the same way as for random features, these terms can be controlled using the key quantity

$$\mathcal{F}_\infty(\lambda) = \sup_{x \in \mathcal{X}} \frac{d\rho(x)}{d\pi(x)} \|(\Sigma + \lambda \mathbf{I})^{-1} k_x\|_{\mathcal{H}}^2, \tag{2.40}$$

where π denotes the measure from which we sample the points x. Using concentration bounds, it is possible to show that (i) and (ii) are well behaved as soon as $m \geq \mathcal{F}_\infty(\lambda)$ up to logarithmic terms, and that in that case, we recover a bound of the form Theorem 2.4.

Reducing the dimension by adapting the sampling to the problem. In this setting, we can also define the score or *leverage score* s_λ of a sample : $s_\lambda(x) = \|(\Sigma + \lambda \mathbf{I})^{-1/2} k_x\|^2$. It the contribution of x to the approximation of $\Sigma + \lambda \mathbf{I}$ by $\Sigma_m + \lambda \mathbf{I}$. $\mathcal{F}_\infty$ is therefore the highest possible score, and imposing $m \geq \mathcal{F}_\infty(\lambda)$ essentially says that to lower bound m, we pay the price of having to see the x with the highest score. To obtain a lower bound on m which is less conservative, it is natural to sample not from $\rho_{\mathcal{X}}$, but from a density which favors x with high scores, and therefore is adapted to the problem.

Ideally, we would like to sample x from the distribution

$$\pi_\lambda(dx) = \frac{s_\lambda(x)\rho_{\mathcal{X}}(dx)}{d_\lambda}, \tag{2.41}$$

which is a probability density (d_λ is the renormalization term). If we use this measure to sample the x, it is possible to show that $\mathcal{F}_\infty = d_\lambda$, and hence that we can better adapt to the effective dimension, *i.e.*, we would obtain $m_{\theta,n} = m_0 \left(\frac{\sigma^2 n}{R^2}\right)^{\frac{1}{2br+1}}$, which is lower than the one obtained before and **matches the ideal** $d_{\lambda_{\theta,n}}$.

Of course, we do not have access to this ideal measure π_λ, as it is defined using problem dependent quantities. However, Rudi, Camoriano, and Rosasco (2015); Rudi, Calandriello, Carratino, and Rosasco (2018) show that it is possible to approximate samples from that ideal distribution by computing approximate leverage scores. The strategy is to sub-sample indices $i_1, ..., i_m$ from $\{1, ..., n\}$ according to a probability vector p_i in order to obtain Nyström centers $\widetilde{x}_j = x_{i_j}$. The probability p_i of sampling index i is set to be proportional to the leverage score :

$$\widehat{l}_i(t) = e_i^\top (\mathbf{K} + tn\mathbf{I})^{-1}\mathbf{K}e_i = \|(\widehat{\Sigma}_n + t\mathbf{I})^{-1/2}k_{x_i}\|^2, \; p_i = \frac{\widehat{l}_i(t)}{\widehat{d}_{n,t}}, \; \widehat{d}_{n,t} = \sum_{i=1}^{n} \widehat{l}_i(t). \tag{2.42}$$

and x is therfore sampled from the distribution

$$\widehat{\pi}_{n,t}(dx) = \frac{\widehat{s}_{n,t}(x)}{\widehat{d}_{n,t}}\widehat{\rho}(dx), \qquad \widehat{s}_{n,t}(x) = \|(\widehat{\Sigma}_n + t\mathbf{I})^{-1/2}k_x\|^2, \widehat{\rho} = \sum_{i=1}^{n} \delta_{x_i}. \tag{2.43}$$

It is therefore possible to show that under assumptions on n and λ, sampling points from this distributionfor $t \approx \lambda$ actually yields the same performance as sampling points from the ideal distribution in Eq. (2.41). Of course, as the point of reducing dimension is not to compute these leverage scores exactly (this would cost $O(n^3)$), Rudi, Calandriello, Carratino, and Rosasco (2018) develop a technique to compute good approximation of these scores in just $\min(nd_\lambda^2, \; d_\lambda^2/\lambda)$. A rapid introduction to the use of leverage scores can be found in Sec. 4.D .3.

Summary. To summarize using either random feature techniques or Nyström points, it is possible to reduce the dimension to $m \ll n$ while keeping the same statistical properties. In the case of Nyström sub-sampling, we can even reduce that dimension to d_λ, the "true" dimension of the space $\mathcal{H}$ regularized with λ, and this with a good computational cost. One last step remains to be done in order to actually compute a good estimator : the actual solving of problems Eqs. (2.30) and (2.37).

2.1.3 Fast algorithms

As we have seen in Eqs. (2.30) and (2.37), the dimension-reduced empiricical risk minimization problem are equivalent to the solving of a problem of the form :

$$\widehat{\alpha}_{n,\lambda,m} = \min_{\alpha \in \mathbb{R}^m} \tfrac{1}{n}\|\Phi^\top \alpha - y\|^2 + \lambda\|\alpha\|_{\mathbb{R}^m}^2, \qquad \widehat{f}_{n,\lambda,n}(x) = \widehat{\alpha}_{n,\lambda,m}^\top \phi(x), \tag{2.44}$$

where $\Phi \in \mathbb{R}^{m \times n}$ and whose solution is $(\Phi\Phi^\top + n\lambda\mathbf{I})^{-1}\Phi y$. The definition of Φ and ϕ from the have been given in Sec. 2.1.2 for the Nyström approach and the random features approach. Recall that the Nyström approach includes the standard e.r.m., where we just take the Nyström points

to be $x_1, ..., x_n$. Note that $\Phi = 1/\sqrt{m}(\phi(x_i, \omega_j))$ in the random features case, and $\Phi = \mathbf{T}^{-\top}\mathbf{K}_{n,m}^{\top}$ in the Nystrom case, where $\mathbf{T}$ is an upper triangular matrix.

Computing $\widehat{\alpha}_{n,\lambda,m}$ by directly solving the system $(\Phi\Phi^{\top} + n\lambda\mathbf{I})\alpha = \Phi y$ and inverting the matrix $(\Phi\Phi^{\top} + n\lambda\mathbf{I})$ is prohibitive. Suprisingly, this is not because of the matrix inversion, whose cost is of order $O(m^3)$, but because computing the matrix $(\Phi\Phi^{\top} + n\lambda\mathbf{I})$ costs $O(nm^2)$ because of the matrix product (and the inversion of the triangular matrix in the Nystrom case).

However, computing matrix vector products of the form Φa for $a \in \mathbb{R}^n$ and $\Phi^{\top}\alpha$ for $\alpha \in \mathbb{R}^m$ has a cost of $O(nm + m^2) = O(nm)$ since $m \leq n$. This suggest the use of an iterative method, such as gradient descent or conjugate gradient descent (see proposition 1.1), the second one being specifically adapted to the setting of solving linear systems, and which consist in solving iteratively the associated optimization problem

$$\widehat{\alpha}_{n,\lambda,m} = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^m} \; F(\alpha) = \tfrac{1}{2}\alpha^{\top}(\Phi\Phi^{\top} + n\lambda\mathbf{I})\alpha - \alpha^{\top}B^{-1}\Phi y, \tag{2.45}$$

However, as we have seen in proposition 1.1, the number of gradient steps needed depends on the conditioning of the system, which will be of order $\frac{1}{\lambda}$ in this case. This can be solved by applying a pre-condtitionning method, which we describe below.

Preconditioning. The idea of preconditioning is to find a matrix B called preconditioner such that $\kappa^{-1/2}BB^{\top} \preceq (\Phi\Phi^{\top} + n\lambda\mathbf{I}) \preceq \kappa^{1/2}BB^{\top}$ for some fixed $\kappa \geq 1$, to define $\alpha = B^{\top}\beta$ and to solve the problem in β, which becomes well conditioned. The condition to do this is that B be fast to compute, and that solving linear systems with B be easy to compute as well. We can then solve the system

$$B^{-1}(\Phi\Phi^{\top} + n\lambda\mathbf{I})B^{-\top}\widehat{\beta}_{n,\lambda,m} = B^{-1}\Phi y, \tag{2.46}$$

which can also be expressed as the convex minimization problem

$$\widehat{\beta}_{n,\lambda,m} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^m} \; \tfrac{1}{2}\beta^{\top}B^{-1}(\Phi\Phi^{\top} + n\lambda\mathbf{I})B^{-\top}\beta - \beta^{\top}B^{-1}\Phi y, \tag{2.47}$$

which is well conditioned, with conditioning κ (it is $\sqrt{\kappa}$ smooth and $1/\sqrt{\kappa}$ strongly-convex). One can therefore readily apply either the conjugate gradient or gradient descent algorithm (in order for the method to be iterative). The number of gradient or conjugate gradient steps needed is of order $\log\frac{1}{\varepsilon}$ in order to achieve precision ε (as the condition number is now a fixed κ which disappears from the bound). The precision which must be achieved is of the order of the statistical error.

Pre-conditioned gradient descent seen as an approximate Newton step. An interesting remark is that pre-conditioned gradient descent is actually closely related to a second order method. Indeed, note that if we apply standard gradient descent, *i.e.*,

$$\beta_{t+1} = \beta_t - \tfrac{1}{\sqrt{\kappa}}\left(B^{-1}(\Phi\Phi^{\top} + n\lambda\mathbf{I})B^{-\top}\beta_t - B^{-1}\Phi y\right),$$

then the corresponding equation for the variable α is just $\alpha_{t+1} = \alpha_t - \widetilde{\Delta}_t$ where $\widetilde{\Delta}_t = \widetilde{\mathbf{H}}^{-1}((\Phi\Phi^{\top} + n\lambda\mathbf{I})\alpha_t - \Phi y)$ and $\widetilde{\mathbf{H}} = \sqrt{\kappa}BB^{\top}$. As $\widetilde{\Delta}_t = \widetilde{\mathbf{H}}^{-1}\nabla F(\alpha_t)$ and $\nabla^2 F(\alpha_t) \preceq \widetilde{\mathbf{H}} \preceq \kappa\nabla^2 F(\alpha_t)$, this is an approximate Newton step. In Sec. 2.3 and in particular in Sec. 2.3.3, we will use these methods in order to minimize a broader class of convex functions.

Ways of finding pre-conditioners

In order to find a good pre-conditioner for $(\Phi\Phi^\top + n\lambda\mathbf{I})$ exploiting the particular structure of that matrix, it is common to proceed in two steps.

(i) Find a matrix $\widetilde{\Phi} \in \mathbb{R}^{m\times q}$ with $q \leq m$ such that :

$$1/\sqrt{\kappa}(\Phi\Phi^\top + n\lambda\mathbf{I}) \preceq \widetilde{\Phi}\widetilde{\Phi}^\top + n\lambda\mathbf{I} \preceq \sqrt{\kappa}(\Phi\Phi^\top + n\lambda\mathbf{I}). \tag{2.48}$$

(ii) Compute B as a lower triangular cholesky factor : $BB^\top = \widetilde{\Phi}\widetilde{\Phi}^\top + n\lambda\mathbf{I}$. Finding this cholesky factor has cost $O(\min(m,q)^3)$ in time and $O(m^2)$ in memory.

Such a matrix $\widetilde{\Phi}$ can be obtained in two main ways, and the complexity of obtaining such a matrix depends on the effective dimesion $d_\lambda(\Phi) = \mathrm{Tr}((\Phi\Phi^\top + n\lambda\mathbf{I})^{-1}\Phi\Phi^\top) \leq \min(m,n)$. Note that if n and m satisfy the assumptions needed for the non asymptotic bias variance trade offs of the type Theorem 2.4, the effective dimension $d_\lambda(\Phi)$ actually matches the effective dimension of the statistical problem d_λ with high probability.

Column subsampling. This is exactly the same thing as Nyström sampling, i.e. we approximate $\Phi = (\phi_1|...|\phi_n)$ with $\widetilde{\Phi} = (\phi_{i_1}|...|\phi_{i_q})\,\mathrm{diag}(1/\sqrt{p_{i_k}}) \in \mathbb{R}^{m\times q}$ where the selected columns $(i_k)_{1\leq k\leq q}$ are q i.i.d. samples from $\{1,..,,n\}$ with distribution given by the probability vector $(p_i)_{1\leq i\leq n}$.

In the literature, two main types of column subsampling exist, and satisfy the Eq. (2.48) with probability at least $1-\delta$ for $\delta \in (0,1]$.

(a) Subsampling uniformly from the colums (Roosta-Khorasani and Mahoney, 2019; Rudi, Camoriano, and Rosasco, 2015), where $q = O(1/(\lambda\kappa)\log\frac{1}{\lambda\delta})$. This leads to a total complexity of $O(m^3 + 1/\lambda^2)$ in time to compute the precondioner, up to logarithmic terms and considering κ is fixed.

(b) Subsampling with approximate leverage scores (Roosta-Khorasani and Mahoney, 2019; Alaoui and Mahoney, 2015; Rudi, Camoriano, and Rosasco, 2015), where the the samples are sampled using a probability vector (p_i) which are an approximation of the probability vector proportional to the the true leverage scores $(e_i^\top(\Phi^\top\Phi + n\lambda\mathbf{I})^{-1}\Phi^\top\Phi e_i)_{1\leq i\leq n}$. Such an approximation can be computed in time $O(\min(n,1/\lambda)d_\lambda(\Phi)^2)$ (Rudi, Calandriello, Carratino, and Rosasco, 2018). In that case, taking $q = O(\kappa^{-1}d_\lambda(\Phi)\log\frac{1}{\lambda\delta})$ is sufficient to guarantee Eq. (2.48).

Sketching. In this case, we approximate Φ by $\widetilde{\Phi} = \Phi\Omega$ where $\Omega \in \mathbb{R}^{n\times q}$ is a random matrix, and where the product $\Phi\Omega$ can be computed in time $O(nm\log n)$ using fast Fourier or Hadamard transforms as is done by Pilanci and Wainwright (2017). In the work by Pilanci and Wainwright (2017), it is shown that it is sufficient to take $q = O(\kappa^{-1}d_\lambda(\Phi)\log\frac{1}{\lambda\delta})$ in order to guarantee Eq. (2.48) with probability at least $1-\delta$.

To summarize, it is possible to find a good preconditioner in time $O(nm + m^3 + md_\lambda(\Phi)^2)$ up to logarithmic terms if we use sketching, or $O(m^3 + md_\lambda(\Phi)^2 + \min(n,1/\lambda)d_\lambda(\Phi)^2)$ up to logaritmic terms if we use column subsampling with approximate leverage scores.

An algorithm with optimal statistical guarantees and a low complexity

In the work by Rudi, Carratino, and Rosasco (2017), a complete methodology to obtain an estimator with optimal guarantees but with low computational complexity is presented. This

methodology uses the following ingredients.

- The dimension reduction phase is done with m Nyström points, which are subsampled from the (x_i) using a probability vector $(p_i)_{1\leq i\leq n}$ computed from approximate leverage scores.
- Solving Eq. (2.37) is done with t steps of a conjugate gradient descent, and with a preconditioner which is computed using column subsampling using the same indices as that of the m Nyström points (and hence associated with the same probability vector (p_i)).

Denote with $\widetilde{f}_{n,\lambda,m,t}$ the function in $\mathcal{H}_m$ obtained after t iterations of conjugate gradient descent using the preconditioner described above. This estimator has been dubbed the "FALKON" estimator by Rudi, Carratino, and Rosasco (2017). The following non-asymptotic upper rates can be obtained for this estimator,which *a)* matches the rates obtained for the regularized empirical risk minimizer, and *b)* is computable in finite time with a complexity of order $O(nd_\lambda + d_\lambda^3)$. Theorem 2.8 is a relatively informal rewriting of Theorem 5. by Rudi, Carratino, and Rosasco (2017) for the probability classes described in Sec. 2.1.1.

Theorem 2.8 (Informal rates for FALKON)**.** *Fix $r \in [1/2, 1], \beta > 0, b > 1$. There exists C and m_0 such that for all $\theta \in \Theta$, there exists $n_0, n_1, t_0 \in \mathbb{N}$ depending on θ such that for any $n \geq n_0$, for any $\delta \in (0, 1]$ if $m \geq m_{\theta,n} \log \frac{n}{\delta}$ where*

$$m_{\theta,n} = m_0 \left(\frac{\sigma^2 n}{R^2}\right)^{\frac{1}{2br+1}}, \tag{2.49}$$

if $\log \delta^{-1} \leq (n/n_1)^{\frac{b(r-1/2)}{2br+1}}$, and if $t \geq t_0 \log(n)$, it holds with probability at least $1-\delta$:

$$\mathcal{R}(\widetilde{f}_{n,\lambda_{\theta,n},m,t}) - \mathcal{R}(f_\rho) \leq CR^2 \left(\frac{\sigma^2 n}{R^2}\right)^{\frac{2br}{2br+1}}. \tag{2.50}$$

Moreover, if $m = m_{\theta,n}$ and $t = t_0 \log(n)$, the time complexity of computing $\widetilde{f}_{n,\lambda_{\theta,n},m,t}$ is of order $O(n^{\frac{2br+2}{2br+1}})$, where the O notations hides constants depending on the class parameters, and logarithmic terms in n and $1/\delta$. This corresponds to a complexity of order $O(nd_{\lambda_{\theta,n}})$ up to logarithmic terms.

2.2 Going beyond the quadratic case with (generalized) self-concordance

In order to extend the result of Sec. 2.1, we will need tools which allow to go from quadratic loss functions to more general convex losses. The core idea is to be able to locally approximate the losses by their second-order taylor expansion, that is if F is defined on a Hilbert pace H and is twice differentialble,

$$F(w) = F(w_0) + \nabla F(w_0)^\top (w - w_0) + \tfrac{1}{2}\|w - w_0\|^2_{\mathbf{H}(w_0)} + o_{\|w-w_0\|\to 0}(\|w - w_0\|^2), \tag{2.51}$$

where $\mathbf{H}(w)$ is the Hessian of F at w. However, in general, such an approximation is not strong enough. In order to generalize the results from least-squares, a precise control on the evolution of the Hessians is needed (since in the least-squares case, the Hessian stays constant). In particular,

this control has to encompass the entire spectrum of the Hessian : one must find regions where the relative evolution of the Hessian is small. We have two main types of results to generalize from the least-squares setting : *a)* optimization results, which are independent from the condition number, and *b)* satistical results, to show that a bias-variance decomposition is still possible. The main difficulty here is that there is no closed-form solution for the estimators $\widehat{f}_{n,\lambda}$; they are only known as minimizers of a certain problem. Moreover, in order to have optimization algorithms which are independent from the condition number, one has to resort to second-order type methods.

Extending the optimization properties. To extend the optimization algorithms deployed in the least-square case, it is natural to think of a Newton method, as the full Newton methods formally minimizes the quadratic approximation of F at each step. Moreover, as it is invariant by reparametrization, it is independent of the conditioning of the problem.

Recall that a Newton method is of the form $w_{t+1} = w_t - \alpha_t \Delta(F, w_t)$ with $\Delta(F, w_t) = \mathbf{H}(w_t)^\dagger \nabla F(w_t)$ is the Newton step and α_t is a step size and is set to 1 in the case of a full Newton method. In the least-squares case, the Hessian never varies, and this makes the Newton method actually just a one step method. But if the Hessian changes rapidly between two close points, (even in terms of small eigenvalues), the Newton steps between two close points may be significantly different. In general, there is no fast convergence rate for Newton methods even in the L-lipschitz μ-strongly convex case.

In the literature, another assumption is made, which allows to control the evolution of the entire Hessian spectrum locally, and which leads to very fast convergence of the Newton method, which is called self-concordance (Nesterov and Nemirovskii, 1994).

Extending the statistical rates. Recall that we consider the problem of minimizing the expected risk $\mathcal{R}(f) = \mathbb{E}\left[\ell_Z(f(X))\right]$ for a loss function $\ell_z(t)$, and that we would like to study the performance of the regularized empirical risk minimizer $\widehat{f}_{n,\lambda}$. Let $\mathcal{R}_\lambda$ be the regularized expected risk and f_λ its minimizer. In the case where ℓ_z is a log-likelihood, as explained by Ostrovskii and Bach (2018), the local asymptotic normality theory for maximum likelihood estimation can show for instance that for any $\lambda > 0$,

$$\sqrt{n}\mathbf{H}_\lambda^{1/2}(\widehat{f}_{n,\lambda} - f_\lambda) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \mathbf{H}_\lambda^{-1/2}\mathbf{G}_\lambda\mathbf{H}_\lambda^{-1/2}) \tag{2.52}$$

where $\mathbf{G}_\lambda = \mathbb{E}\left[(\ell'_Z(f_\lambda(X))k_X + \lambda f_\lambda) \otimes \ell'_Z(f_\lambda(X))k_X + \lambda f_\lambda)\right]$, $\mathbf{H}_\lambda$ is the Hessian of $\mathcal{R}_\lambda$ at f_λ, and the convergence happens in distribution.

These results are usually proven using the second order delta method, which relies on the second order development of the function $\mathcal{R}_\lambda$ as in Eq. (2.51). However, they are asymptotic. Since in the non-parametric (kernel) setting, there is a trade off between the regularization parameter and n, a non-asymptotic result is necessary to understand the interaction between the two and derive upper rates of convergence.

Note that while the result Eq. (2.52) is not sufficient, it gives a good intuition on what the variance term in a bias-variance decomposition should look like. In order to establish finer non-asymptotic bounds, we need to be able to measure upper bound the distance between $\widehat{f}_{n,\lambda}$ and f_λ with an expression which can be computed in closed form, knowing that the minimizer themselves are not accessible as such. This will be the purpose of the localization lemmas, introduced by Bach (2010); Ostrovskii and Bach (2018) and which we generalize to the random

design, infinite dimensional setting in the work by Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019). We will further discuss these lemmas in Sec. 2.2.3.

2.2.1 (Generalized)-self concordance and control of the Hessian

In this section, we present two ways of controlling the variations of the Hessians which can be found in the literature. While the first, called *self-concordance*, is arguably the best in terms of optimization, it is not obvious to show that our ideal loss function $\mathcal{R}$ satisfies this assumption, based on assumptions on the loss $\ell_z(f(x))$; in a sense, it is *a priori* more adapted to optimization than statistical learning. The second, called *generalized self-concordance*, is more adapted to the statistical learning setting, and is satisfied by classical losses in machine learning. However, it is less performant from an optimization point of view, as the evolution of the Hessian is bounded less tightly.

In this section, we will denote with $F : H \to \mathbb{R}$ functions defined on a Hilbert space H. We will use the notation $\|\cdot\|_{\mathbf{M}}$ to denote $\|\mathbf{M}^{1/2}\cdot\|$ for a positive semi-definite operator $\mathbf{M}$ on H. We will denote with $\mathbf{H}(w)$ the Hessian of F at point $w \in H$.

Self concordance

The notion of self-concordance has been defined and used by Nesterov and Nemirovskii (1994) in order to design Newton based second-order methods. Standard barrier functions, such as the log-barrier or the log determinant in the context of semidefinite programming, satisfy this assumption, making it the key to the success of interior point methods (see works by Boyd and Vandenberghe (2004); Nesterov and Nemirovskii (1994); Nesterov (2018) as well as the description of interior point methods in Sec. 1.1.3).

Definition 2.2 (self-concordance)**.** *A function F defined on a domain of H is said to be self-concordant if it is thrice differentiable and*

$$\forall h \in H,\ D^3F(w)[h,h,h] \leq 2\left(D^2F(w)[h,h]\right)^{3/2}, \tag{2.53}$$

where D^kF is simply the k-th diffential of F and is a symmetric k form on H.

This assumption implies the following control on the second derivatives which can be found as Eq. (9.46) by Boyd and Vandenberghe (2004) :

$$\forall h \in H,\ \frac{1}{(1+t\|h\|_{\mathbf{H}(w_0)})^2}\|h\|^2_{\mathbf{H}(w_0)} \leq \|h\|^2_{\mathbf{H}(w_0+th)} \leq \frac{1}{(1-t\|h\|_{\mathbf{H}(w_0)})^2}\|h\|^2_{\mathbf{H}(w_0)}, \tag{2.54}$$

this being valid for all t, h, w_0 such that $t < \|h\|^{-1}_{\mathbf{H}(w_0)}$. This bound allows to control the Hessian values close to w_0 only using the metric $\|\cdot\|_{\mathbf{H}(w_0)}$, and can be used to derive the different localization results we will see in Sec. 2.2.2.

However, this self-concordance notion is less adapted to the statistical setting. Indeed, if $F_1, ..., F_K$ are self concordant, then $K\left(\frac{1}{K}\sum_{i=1}^K F_i\right)$ is self-concordant (and not $\left(\frac{1}{K}\sum_{i=1}^K F_i\right)$). This is a problem in the setting we consider since in general, even if the functions $f \in \mathcal{H} \mapsto \ell_z(f(x))$ are all self-concordant, the expected risk $\mathcal{R}(f) = \mathbb{E}\left[\ell_Z(f(X))\right]$ is not necessarily self-concordant; it is hard to know if a problem is self-concordant based only on the loss function. Note that more involved assumptions can be made to guarantee this property, as in the work by Ostrovskii and Bach (2018).

Generalized self-concordance

The fact that self-concordant functions do not interact well with expectations motivated the following *generalized self-concordance* or *pseudo self-concordance* assumption (GSC), made by Bach (2010) in order to analyse empirical risk minimization for logistic regression. Note that the logistic loss is not self-concordant, but satisfies a different property controlling the variations of its Hessians.

Definition 2.3 (generalized self-concordance)**.** *A function F defined on a domain of H is said to be $\mathfrak{R}$-generalized self-concordant if for any w in the domain of H, it holds*

$$\forall h_1, h_2 \in H,\ D^3F(w)[h_1, h_1, h_2] \leq \mathfrak{R}\|h_2\|_H D^2F(w)[h_1, h_1]. \tag{2.55}$$

Moreover, we define $\mathsf{r}(F) = \inf_{w\in\mathrm{dom}(F)} \sqrt{\lambda(w)}/\mathfrak{R}$, where $\lambda(w)$ is the smallest eigenvalue of $\mathbf{H}(w)$.

Note that in order for $\mathsf{r}(F)$ to be positive, the function F must be strongly convex, and $\mathsf{r}(F) = \sqrt{\lambda}/\mathfrak{R}$ where λ is the highest strong convexity constant of F. The quantity $\mathsf{r}(F)$ will be refered to as the Dikin radius of the function F.

Remark 4. *The term "generalized self-concordance" can be a bit confusing as a self-concordant function is not necessarily generalized self-concordant. However, all self-concordant functions with bounded Hessians are generalized self-concordant.*

Note that a more involved definition of generalized self-concordance is given in chapters 3 and 4. In this introduction, we will keep to this definition for simplicity. Note that if $F_1, ..., F_K$ are $\mathfrak{R}$ generalized self concordant, then $\left(\frac{1}{K}\sum_{i=1}^K F_i\right)$ is also $\mathfrak{R}$ generalized self-concordant. More generally, expectations of GSC functions are GSC, which is adapted to our setting. Example 3.1 provide examples of functions in machine learning which are generalized self-concordant, among which the logistic regression function for outputs in $\{-1, 1\}$ with a RKHS associated to the kernel k: $f \in \mathcal{H} \mapsto \ell_{x,y}(f(x))$ is GSC with $\mathfrak{R} = 2\sqrt{k(x,x)}$.

In the case of generalized self-concordance, the following bound can be obtained on the Hessians and are proved by Bach (2010).

$$e^{-t\mathfrak{R}\|h\|_H}\|\cdot\|_{\mathbf{H}(w_0)} \preceq \|\cdot\|_{\mathbf{H}(w_0+th)} \preceq e^{t\mathfrak{R}\|h\|_H}\|\cdot\|_{\mathbf{H}(w_0)}. \tag{2.56}$$

This bound is not as good as the one in Eq. (2.54) as it depends on the norm of h in H which, in a sense, is not the right norm (the right one being the norm $\|h\|_{\mathbf{H}(w_0)}$).

In chapters 3 and 4, we make an analysis analoguous to that of least squares but for generalized self-concordance losses, *i.e.*, when the $\ell_z(\cdot)$ are self-concordant with the same constant.

In what follows, we will recall the main properties of self-concordant and generalized self-concordant functions in terms of optimization and statistics, and relate them to the problem we will handle in Sec. 2.3.

2.2.2 Optimization and Newton methods

In this section, we will show how the Newton method behaves under the *self-concordant* (SC) and the *generalized self-concordant* (GSC) assumption. Define the *Newton decrement* $\nu(F, w)$ of F at w :

$$\nu(F,w)^2 = \|\Delta(F,w)\|^2_{\mathbf{H}(w)} = \nabla F(w)^\top \mathbf{H}(w)^\dagger \nabla F(w) = \|\nabla F(w)\|^2_{\mathbf{H}(w)^{-1}}, \tag{2.57}$$

with a slight abuse of notation for the last case (when the Hessian is not invertible). The Newton decrement is a key quantity to analyse the Newton method for SC and GSC functions, and was used for these analysis by Nesterov and Nemirovskii (1994). Indeed, it has the nice property of depending only on the function F at w. Moreover, it is possible to show that when the Newton decrement is small enough, then it is equivalent to the quantities $\|w - w_*\|_{\mathbf{H}(w_*)}$ and $\sqrt{F(w) - F(w_*)}$, which are the quantities we would like to minimize but to which we have no access since w_* is unknown. More formally, we define the *Dikin ellipsoids* of F as the sets :

- $\mathsf{D}(F, \mathsf{c}) = \{w \in H \ : \ \nu(F, w) \leq \mathsf{c}\}$ if F is self-concordant;
- $\mathsf{D}(F, \mathsf{c}) = \{w \in H \ : \ \nu(F, w) \leq \mathsf{c}\ \mathsf{r}(F)\}$ if F is generalized self concordant.

These sets have been introduced by Ostrovskii and Bach (2018); Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019); Marteau-Ferey, Bach, and Rudi (2019) in different ways which are essentially equivalent to this one. The following proposition shows that as soon as the Newton decrement is small enough, *i.e.*,, as soon as w belongs to a certain Dikin ellipsoid, the Newton decrement is actually equivalent to the distance to the optimum. This proposition can easily be deduced from the standard Taylor expansion bounds for SC and GSC functions, as can be found in the work by Nesterov and Nemirovskii (1994) for SC functions and propositions 3.4 and 3.5 in chapter 3 for the GSC case (note that these last results are themselves inspired from results by Ostrovskii and Bach (2018); Bach (2010)).

Proposition 2.2 (localization)**.** *There exists positive constants c_1, c_2, c_3, c_4 such that if there exists $w_0 \in \mathsf{D}(F, \mathsf{c}_0)$ where $\mathsf{c}_0 = 1/2$ in the GSC case and $\mathsf{c}_0 = 1/4$ in the SC case, then a minimizer w_* of F exists, and for any $w \in \mathsf{D}(F, \mathsf{c}_0)$, it holds*

$$\nu(F, w)^2 \leq c_1 \|w - w_*\|^2_{\mathbf{H}(w)} \leq c_2 \|w - w_*\|^2_{\mathbf{H}(w_*)} \leq c_3 (F(w) - F(w_*)) \leq c_4 \nu(F, w)^2. \tag{2.58}$$

The result stated in proposition 2.2 is called a *localization* result since it shows that one can localize the minimizer w_* with only local information at w (*i.e.*, its Newton decrement) as soon as the Newton decrement is small enough.

The previous results shows that if we are able to control the decrease of the Newton decrement, then we are able to control the distance to the optimum. The analysis of the full Newton method relies precisely on this fact, and shows that as soon as w is in the right Dikin ellipsoid, then the Newton decrement of $w - \Delta(F, w)$ can be controlled by the one of w. This result can be found as Proposition 2 in the work by Bach (2010) for the GSC case. The formulas for the SC case are well known from the work by Nesterov and Nemirovskii (1994) and are also recalled in section 1. of the work by Bach (2010).

Theorem 2.9 (behavior of the full Newton method)**.** *If F is self concordant, then as soon as $w \in \mathsf{D}(F, \mathsf{c}_0)$ for $\mathsf{c}_0 = 1/4$, F has a global minimizer w_* and it holds :*

$$\nu(F, w - \Delta(F, w)) \leq \nu(F, w)^2. \tag{2.59}$$

If F is generalized self concordant, then as soon as $w \in \mathsf{D}(F, \mathsf{c}_0)$ for $\mathsf{c}_0 = 1/2$, F has a global minimizer w_* and it holds :

$$\frac{\nu(F, w - \Delta(F, w))}{\mathsf{r}(F)} \leq \left(\frac{\nu(F, w)}{\mathsf{r}(F)} \right)^2. \tag{2.60}$$

The results of Theorem 2.9 directly imply that a full Newton method $w_{t+1} = w_t - \Delta(F, w_t)$ starting at $w_0 \in \mathsf{D}(F, t_0)$ converges quadratically, *i.e.*, $w_t \in \mathsf{D}(F, t_0^{2^t})$.

Note that this fast convergence only happens "near" the optimum, that is for small Newton decrements. If one wishes to apply a Newton method starting from any point $w_0 \in H$, one cannot apply a full Newton method directly.

This is not an issue in the context of SC functions. Indeed, one can apply a damped Newton method $w_{t+1} = w_t - \alpha_t \Delta(F, w_t)$, where the stepsize α_t is suitably chosen (either by a line search or defined by the Newton decremement, see the original method by Nesterov and Nemirovskii (1994) or the book by Boyd and Vandenberghe (2004) for the line-search version). This first phase can be shown to converge in a finite number of steps which is independent from the function F to minimize and is proporitional to the quantity $F(w_0) - F(w_*)$ (in particular, it does not depend on the conditioning of the Hessians of F). In a sense, for SC functions, the Dikin ellipsoid is large, and reaching it is not complicated.

In the context of GSC functions however, this becomes an issue. First, choosing the stepsize α_t is more involved, and does not lead to a constant decrease of $F(w_t) - F(w_*)$ as in the SC case (see the results by Sun and Tran-Dinh (2017)). Moreover, the Dikin ellipsoid is smaller, since $\mathsf{r}(F)$ depends on the conditioning of the problem through the strong convexity constant of F. Roughly, one has to reach a precision of order $F(w_t) - F(w_*) \approx \frac{\lambda}{\mathfrak{R}^2}$ in order to be in the Dikin ellipsoid, where λ is the strong convexity constant (this is to be compared to $F(w_t) - F(w_*) \approx \frac{1}{16}$ in the SC case). That is why the Newton method is said to be a local method for GSC functions, *i.e.*, it converges fast locally. In Sec. 2.3.3 and chapter 4, we will present a global scheme to minimize GSC functions.

Summary. To summarize, we have defined classes of functions, SC and GSC functions, which allow a precise local control of the Hessians. We have seen that SC and GSC functions can be locally optimized using a full Newton method. These results are based on two main ingredients : *a)* as soon as a point w_0 is in the Dikin ellipsoid of F with certain radius t_0, the Newton decrement characterizes the distance of w_0 to the optimum, and *b)* in this same Dikin ellipsoid, the full Newton methods induces a quadratic decrease of the Newton decrement and hence of the distance to the optimum in function values. However, while a global scheme can be derived for SC functions, it is not the case for GSC functions, and reaching the Dikin ellipsoid is *a priori* a challenge.

Remarks. Note that a unifying framework to analyze Newton methods for SC and GSC functions (and other classes of functions) has been provided by Sun and Tran-Dinh (2017). In particular, they provide a choice of stepsize in the first phase of the Newton method for the GSC case. However, this does not lead to a constant decrease in function values, as is the case for SC functions.

Moreover, Sun and Tran-Dinh (2017) also consider the case where the Newton method is not applied exactly, but using a *relative approximation* of the Newton step. More formally, they consider the case where the Newton step $\Delta(F, w_t)$ is approximated by $\widetilde{\Delta}_t$ such that $\|\Delta(F, w_t) - \widetilde{\Delta}_t\|_{\mathbf{H}(w_t)} \leq \rho\nu(F, w_t)$ for a fixed $\rho < 1$. This can be quite useful in practice, as it allows to avoid solving the linear system to compute the Newton step. Using this approximate Newton step instead of the full Newton step still leads to a fast convergence of the Newton method in the Dikin ellipsoid, but at the cost of losing the quadratic convergence for a linear

convergence. We will detail this in chapter 4 as well as in Sec. 2.3.3, where we derive a globally convergent algorithm for GSC functions based on approximate Newton methods.

2.2.3 Statistics

In this section, we show how the control brought by generalized self-concordance can help derive statistical bounds, by allowing to localize the empirical risk minimizer using a Newton decrement, on which wecan then apply standard concentration bounds. More specifically, in this section, we will show how the reguarlized empirical risk minimizer can be localized in proposition 2.3, paving the way to the use of standard concentration bounds to bound the variance term in a bias variance decomposition. We will the use these bounds in Sec. 2.3.1 to prove a bias-variance decomposition in the setting of GSC functions.

In spirit, the same type of bounds have been developed by Bach (2010); Ostrovskii and Bach (2018). However, they are made for a different setting than ours, and therefore slightly differ from the method we will expose here. In particular, Bach (2010) considers the fixed design setting, while the work by Ostrovskii and Bach (2018) does not include the regularization parameter, which is crucial in the non-parametric setting.

Let us once again consider the problem of minimizing the expected riskt $\mathcal{R}(f) = \mathbb{E}\left[\ell_Z(f(X))\right]$ over the set $\mathcal{H}$, whose Hessian at f we denote with $\mathbf{H}(f)$. We will make the following assumption on the loss function.

Assumption 2.7 (GSC assumption)**.** *For all $z \in \mathcal{Z}$, the function $t \mapsto \ell_z(t)$ is $\mathfrak{R}$ generalized self-concordant.*

As in the previous sections, we will also assume that the kernel is bounded by 1 (see Assumption 2.3). We adopt the following notations for the different regularized functions and Hessian of functions.

- We denote with $\mathcal{R}_\lambda = \mathcal{R} + \frac{\lambda}{2}\|\cdot\|_{\mathcal{H}}^2$ the regularized expected risk, and denote with $\mathbf{H}_\lambda(f)$ its Hessian at f, and with f_λ its minimizer (which exists by strong convexity as soon as $\lambda > 0$). $\mathcal{R}_\lambda$ is $\mathfrak{R}$ GSC and λ strongly convex.
- We also denote with $\widehat{\mathcal{R}}_{n,\lambda}$ the regularized empirical risk, and denote with $\widehat{\mathbf{H}}_{n,\lambda}(f)$ its Hessian at point f. Recall that as soon as $\lambda > 0$, it $\widehat{\mathcal{R}}_{n,\lambda}$ has a unique minimizer $\widehat{f}_{n,\lambda}$. $\widehat{\mathcal{R}}_{n,\lambda}$ is also $\mathfrak{R}$ GSC and λ strongly convex.
- Finally, we denote with r_λ the quantity $\sqrt{\lambda}/\mathfrak{R}$. It is a lower bound for the Dikin radius of $\mathcal{R}_\lambda$ and $\widehat{\mathcal{R}}_{n,\lambda}$: we have $\mathsf{r}(\mathcal{R}_\lambda), \mathsf{r}(\widehat{\mathcal{R}}_{n,\lambda}) \geq \mathsf{r}_\lambda$.

Bounding the Newton decrement of a function f

$$\nu(\widehat{\mathcal{R}}_{n,\lambda}, f) = \|\nabla\widehat{\mathcal{R}}_{n,\lambda}(f)\|_{\widehat{\mathbf{H}}_{n,\lambda}^{-1}(f)} \leq \|\widehat{\mathbf{H}}_{n,\lambda}^{-1/2}(f)\mathbf{H}_\lambda{}^{1/2}(f)\|\ \|\nabla\widehat{\mathcal{R}}_{n,\lambda}(f)\|_{\mathbf{H}_\lambda^{-1}(f)}, \tag{2.61}$$

we can show the following localization bound for the empirical risk minimizer by using proposition 2.2.

Proposition 2.3 (localization of the the regularized e.r.m.)**.** *There exists an explicit constant c such that for any $f \in \mathcal{H}$, if*

$$\|\nabla\widehat{\mathcal{R}}_{n,\lambda}(f)\|_{\mathbf{H}_\lambda{}^{-1}(f)} \leq \tfrac{\mathsf{r}_\lambda}{4}, \qquad \|\widehat{\mathbf{H}}_{n,\lambda}^{-1/2}(f)\mathbf{H}_\lambda{}^{1/2}(f)\| \leq 2, \tag{2.62}$$

then the following holds :

$$\|f - \widehat{f}_{n,\lambda}\|_{\mathbf{H}_\lambda(f)} \leq c\|\nabla\widehat{\mathcal{R}}_{n,\lambda}(f)\|_{\mathbf{H}_\lambda{}^{-1}(f)}^2 \tag{2.63}$$

To obtain a bound on the variance term in the bias variance decomposition for GSC functions, we will apply proposition 2.3 to f_λ, which paves the way to bound $\|f_\lambda - \widehat{f}_{n,\lambda}\|_{\mathbf{H}_\lambda(f_\lambda)}$. This term will account for the variance term in our bias-variance trade off (see Sec. 2.3.1 as well as the sketch of proof in Sec. 3.6). Indeed, it is possible to guarantee $\|\widehat{\mathbf{H}}_{n,\lambda}^{-1/2}(f)\mathbf{H}_\lambda{}^{1/2}(f)\|^2 \leq 2$ using standard concentration bounds for operators, and to bound $\|\nabla\widehat{\mathcal{R}}_{n,\lambda}(f_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(f_\lambda)}$ as we can see it as $\|\frac{1}{n}\sum_{i=1}^n \xi_i\|$ where ξ_i are i.i.d. samples from $\xi = \mathbf{H}_\lambda(f_\lambda)^{-1/2}\left(\ell'_Z(f_\lambda(X))k_X + \lambda f_\lambda\right)$, which has zero mean. This term can therefore also be concentrated using standard concentration bounds for random vectors in a Hilbert space.

2.3 Main results and contributions of this part

In this section, we present the main results of our contributions on extending the fast rates and algorithms existing for the least squares problem to a broader class of convex functions, the class of GSC functions.

The goal of this section is to present our results in a coherent, unified framework, as well as an idea of the technical contributions. In the original articles, whose verbatim can be found in chapters 3 and 4, the notations are slightly different, and the assumptions are often weaker. These articles contain all the proofs and precise statements, and have been peer-reviewed. We will refer to the results we invoke from these articles in the introduction, even though the notations may vary.

In order to present the results, the outline will be the following. We will start in Sec. 2.3.1 by presenting the statistical results obtained in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) on the non-asymptotic performance of the regularized e.r.m. estimator, which mirrors the results in the least squares setting, presented in Sec. 2.1.1. We will continue in Sec. 2.3.2 to show how it is possible to reduce the dimension of the resulting finite dimensional problem, in the same way as in the least-squares case (although the proofs are a bit more involved). These results have been proved by Marteau-Ferey, Bach, and Rudi (2019). In Sec. 2.3.3, we present one of the main contributions by Marteau-Ferey, Bach, and Rudi (2019) which is a globally convergent second order scheme for GSC functions which is roughly independent of the conditioning of the problem. Finally, in Sec. 2.3.4, we show that, as the FALKON algorithm in the least-squares case, it is possible to effectively compute an estimator, using the previous scheme, with low time complexity and optimal statistical precision. This has also been derived by Marteau-Ferey, Bach, and Rudi (2019).

Assumptions

We use all notations for functions and Hessians introduced in Sec. 2.2.3. The goal is to minimize the expected risk $\mathcal{R}(f) = \mathbb{E}\left[\ell_Z(f(X))\right]$ from samples z_i from Z. Let ρ be the distribution of Z. As in the least squares case, we make the assumption that the samples $z_1, ..., z_n$ are i.i.d. samples from Z (see Assumption 1.1). Moreover, we assume there exists $f_\rho \in \mathcal{H}$ such that f_ρ minimizes the expected risk (see Assumption 2.1). We will assume that the kernel is bounded by 1, although the results hold as soon as the kernel is bounded (see Assumption 2.3). We will also suppose that Assumption 2.7 is satisfied, that is that the functions $\ell_z(\cdot)$ are $\mathfrak{R}$ GSC for all $z \in \mathcal{Z}$.

We also need a more technical assumption. First, for any distribution ρ of Z, define

$$\mathsf{b}_1(f) = |\ell'_Z(f(X))|_{L^\infty(\rho)}, \qquad \mathsf{b}_2(f) = |\ell''_Z(f(X))|_{L^\infty(\rho)}. \tag{2.64}$$

We will assume that these quantities is bounded above at the optimum f_ρ.

Assumption 2.8 (Bounded gradients and Hessians)**.** *There exists finite constants* $\mathsf{b}_1, \mathsf{b}_2$ *such that*

$$\mathsf{b}_1(f_\rho) \leq \mathsf{b}_1, \qquad \mathsf{b}_2(f_\rho) \leq \mathsf{b}_2. \tag{2.65}$$

The assumption on the gradient replaces the noise assumption in the least squares setting. Indeed, if ℓ is the loss associated to the least-squares loss, the assumption on the gradient is just $|Y - f_\rho(X)| \leq \mathsf{b}_1$, which in particular implies the noise assumption with $\sigma, M = \mathsf{b}_1$.

Note that we will use the shorthand notations $\mathbf{H}, \mathbf{H}_\lambda, \widehat{\mathbf{H}}_{n,\lambda}$ to denote the quantities $\mathbf{H}(f_\rho)$, $\mathbf{H}_\lambda(f_\rho)$, and $\widehat{\mathbf{H}}_{n,\lambda}(f_\rho)$.

Examples

We recall two examples where the GSC assumption Assumption 2.7 is satisfied : logistic regression, and the square loss. Other examples can be found in Example 3.1.

Example 2.1. *For logistic regression, we have* $\ell_z(t) = \log(1 + \exp(-yt)$*. If we assume that* $\mathcal{Z} = \mathcal{X} \times [-1, 1]$ *(or more generally that the* y *are bounded), Assumption 2.7 is satisfied with* $\mathfrak{R} = 2$*. Moreover, the* ℓ_z *are all* 1 *lipschitz, and we have :*

$$\forall f \in \mathcal{H},\ \mathsf{b}_1(f) \leq 1,\ \mathsf{b}_2(f) \leq 1/4. \tag{2.66}$$

Example 2.2. *For least-squares regression, we have* $\ell_z(t) = |t-y|^2$*. In this case, Assumption 2.7 is satisfied for any positive* $\mathfrak{R}$*. Moreover, we have, for any distribution* ρ *for* Z*,*

$$\forall f \in \mathcal{H},\ \mathsf{b}_1(f) = |Y - f(X)|_{L^\infty(\rho)},\ \mathsf{b}_2(f) \leq 1. \tag{2.67}$$

2.3.1 Statistics of the regularized empirical risk minimizer for GSC functions

In this section, we present the results obtained by Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) on regularized empirical risk minimization for GSC functions. We start by showing that, as in the least-squares case, there is a bias-variance trade-off of the form

$$\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}_{\mathcal{H}} \leq C \log^2 \tfrac{1}{\delta} \left(b_\lambda + \tfrac{d_\lambda}{n} \right), \qquad \text{with probability at least } 1 - \delta, \tag{1.58}$$

where the bias term and the effective dimension are defined as meaningful quantities which match those defined in the least-squares case, described below. We then present rates of convergence for classes of distributions ρ, as in the least-squares case, before presenting a high level description of the proof of the bias variance trade off.

In Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), great attention has been put to obtain bounds with explicit constants. While we do not report them here, these can be found in the referenced theorems.

Bias-variance decomposition

In this context, the descritption of the bias and variance tade off is done through the introduction of two key quantities, which are the analogs of the quantities b_λ and $\sigma^2 d_\lambda$ in the least-squares setting.

- For any $\lambda > 0$, the bias b_λ is defined as

$$b_\lambda := \nu(\mathcal{R}_\lambda, f_\rho)^2 = \|\nabla \mathcal{R}_\lambda(f_\rho)\|^2_{\mathbf{H}_\lambda^{-1}} = \lambda^2 \|f_\rho\|^2_{\mathbf{H}_\lambda^{-1}}. \tag{2.68}$$

- For any $\lambda > 0$, the effective dimension df_λ is defined as

$$\begin{aligned} \mathsf{df}_\lambda &:= \mathbb{E}\left[\|\mathbf{H}_\lambda^{-1/2} \ell'_Z(f_\rho(X)) k_X\|^2\right] \\ &= \mathrm{Tr}(\mathbf{H}_\lambda^{-1/2} \mathbf{G} \mathbf{H}_\lambda^{-1/2}), \quad \mathbf{G} = \mathbb{E}\left[\ell'_Z(f_\rho(X))^2 \; k_X \otimes k_X\right] \end{aligned} \tag{2.69}$$

In the least squares setting, we have $\mathbf{H} = \Sigma$ and

$$\mathbf{G} = \mathbb{E}\left[(Y - f_\rho(X))^2 k_X \otimes k_X\right]. \tag{2.70}$$

The definition of the bias b_λ in Eq. (2.68) therefore exactly matches the definition given in Sec. 2.1.1 in the least-squares case. The definition of the effective dimension is slightly different, and can be related to the effective dimension d_λ introduced in the context of least-squares as follows :

$$\mathsf{df}_\lambda \leq \sigma^2 d_\lambda, \tag{2.71}$$

where σ^2 is an upper bound on the variance of the noise $\mathbb{E}\left[(Y - f_\rho(X))^2 | X\right]$ (see Assumption 2.4). Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) prove the following bias variance decomposition (we denote with $a \vee b$ the maximum between a and b), which is the main result of their paper, and is reported here in Theorem 3.4.

Main theorem 1: Marteau-Ferey et al. (2019), Theorem 4.

There exists explicit constants n_0, n_1 and $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ such that for any $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{b}_2$, whenever

$$n \geq n_0 \; \frac{\mathsf{b}_2}{\lambda} \log \frac{\mathsf{b}_2}{\lambda \delta}, \;\; n \geq n_1 \; \frac{\mathsf{df}_\lambda \vee \mathsf{q}^2}{\mathsf{r}_\lambda^2} \; \log \frac{2}{\delta}, \;\; \sqrt{b_\lambda} \leq \mathsf{r}_\lambda / 2,$$

with $\mathsf{q}^2 = \mathsf{b}_1^2 / \mathsf{b}_2$, then with probability at least $1 - 2\delta$, it holds

$$\mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}(f_\rho) \leq \mathsf{C}_{\text{bias}} \; b_\lambda + \mathsf{C}_{\text{var}} \; \frac{\mathsf{df}_\lambda \vee \mathsf{q}^2}{n} \; \log \frac{2}{\delta}. \tag{2.72}$$

Upper rates of convergence

Following what is done in Sec. 2.1.1, we define classes of test measures $\mathcal{M}$ for which we get precise statistical rates. Formally, for $R, \mathsf{Q}, \mathsf{b}_1, \mathsf{b}_2 > 0$, $r \in [1/2, 1]$ and $b \geq 1$, we will define the class of test measures $\mathcal{M}(R, r, \mathsf{Q}, b, \mathsf{b}_1, \mathsf{b}_2)$ as the set of measures ρ such that the following hold.

(i) *Source condition* : there exists $h \in \mathcal{H}$ such that $f_\rho = \mathbf{H}^{r-1/2} h$ and $\|h\|_{\mathcal{H}} \leq R$. This is exactly the source condition in the least-squares setting, but adapted to the case where the metric is not the L^2 metric anymore, but the metric associated to the loss function ℓ.

(ii) *Capacity condition* : the generalized eigenvalues λ_i of the pair $(\mathbf{G}, \mathbf{H}_\lambda)$ (that is the eigenvalues of $\mathbf{H}_\lambda^{-1/2} \mathbf{G} \mathbf{H}_\lambda^{-1/2}$) satisfy

$$\sum_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \lambda} = \mathrm{Tr}(\mathbf{H}_\lambda^{-1/2} \mathbf{G} \mathbf{H}_\lambda^{-1/2}) \leq Q^2 \lambda^{-1/b}. \tag{2.73}$$

In particular, this is satisfied as soon as the $\lambda_i \leq \frac{\beta}{i^b}$ for some $b > 1$.

(iii) *Boundedness of gradients and Hessians* : Assumption 2.8 is satisfied with $\mathsf{b}_1, \mathsf{b}_2$.

On this class of functions, the following non-asymptotic upper rates of convergence are derived from Main theorem 1, which was originally proved as Cor. 3.4.

Main corollary 1: Marteau-Ferey et al. (2019), Corollary 3

Fix $\mathsf{b}_1, \mathsf{b}_2 > 0$, $1 \geq r > 1/2$, $Q > 0$,$R > 0$, and $b \geq 1$ and let $\mathcal{M} = \mathcal{M}(R, r, \mathsf{Q}, b, \mathsf{b}_1, \mathsf{b}_2)$. If we set

$$\lambda = \left(\frac{256\mathsf{Q}^2}{R^2 n}\right)^{\frac{1}{2br+1}}, \tag{2.74}$$

then for any $\delta \in (0, 1/2]$, with probability at least $1 - 2\delta$,

$$\forall \rho \in \mathcal{M},\ \mathcal{R}(\widehat{f}_{n,\lambda}) - \mathcal{R}(f_\rho) \leq 8\ R^2 \left(\frac{256\mathsf{Q}^2}{R^2 n}\right)^{\frac{2br}{2br+1}} \log \frac{2}{\delta}, \tag{2.75}$$

provided $n \geq N$ and where N is defined in Eq. (3.46), depends on all the parameters of the model $\mathcal{M}$ as well as δ, (in particular, it depends polynomially on $\log \frac{1}{\delta}$).

This result can be adapted to resemble Cor. 2.1 (in particular, where the relationship between n and $\log \frac{1}{\delta}$ is more explicit), which we assume could then lead to minimax upper rates. However, this is not done in the work by Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019). Note that we have assmumed a stronger source condition than the simple well-specified assumption, *i.e.*, we have assumed $r > 1/2$. In the case where $r = 1/2$, some rates can be obtained, but have a worst depedence on the problem parameters, through an exponential constant $e^{\mathfrak{R}\|f_\rho\|}$. This is due to the fact that we cannot guarantee the condition $b_\lambda \leq \mathsf{r}_\lambda/2$ in that setting. For more details, we refer to chapter 3 for more elements on this question.

Proof techniques

We refer to Sec. 3.6 for a sketch of the proof.

2.3.2 Reducing the dimension while keeping optimal rates

As explained in the least-squares regression case, solving the regularized empirical risk minimization problem can be computationally expensive. The first step to have a better algorithm is to reduce dimension of the problem through one of the dimension reduction techniques presented in Sec. 2.1.2.

Marteau-Ferey, Bach, and Rudi (2019) consider the case where the dimension is reduced by sampling m Nyström points $\widetilde{x}_1, ..., \widetilde{x}_m$ either uniformly or using approximate leverage scores associated to the kernel matrix $\mathbf{K} = (k(x_i, x_j)) \in \mathbb{R}^{n \times n}$. Recall that in the last case, the technique is to sub-sample indices $i_1, ..., i_m$ from $\{1, ..., n\}$ using a good approximation of the probability vector $p_i = \widehat{l}_i(t) / \sum_{i'} \widehat{l}_{i'}(t)$, where $\widehat{l}_i(t) = e_i^\top \mathbf{K}(\mathbf{K} + nt\mathbf{I})^{-1} e_i$ (for more details on leverage scores, see Eq. (2.42) or Sec. 4.D .3). The Nyström points are then defined as $\widetilde{x}_j = x_{i_j}$. When performing Nyström sampling, $\mathcal{H}$ is approximated by

$$\mathcal{H}_m = \left\{ x \mapsto \sum_{j=1}^{m} \alpha_j k(x, \widetilde{x}_j) \ : \ \alpha \in \mathbb{R}^m \right\}, \tag{2.35}$$

This set can simply be seen as a m dimensional subspace of $\mathcal{H}$ (it is simply $\text{span}(k_{\widetilde{x}_j} \ : \ 1 \leq j \leq m)$) with the induced norm : $\langle \sum_{j=1}^m \alpha_j k(\cdot, \widetilde{x}_j), \sum_{j=1}^m \beta_j k(x, \widetilde{x}_j) \rangle_{\mathcal{H}_m} = \alpha^\top \mathbf{K}_{mm} \beta$ where $\mathbf{K}_{mm}$ is the kernel matrix associated to the $\widetilde{x}_j$. The kernel associated to the space $\mathcal{H}$ can be expressed as $k_m(x, x') = \mathbf{k}_m(x)^\top \mathbf{K}_{mm}^\dagger \mathbf{k}_m(x)$ where $\mathbf{k}_m(x) = (k(\widetilde{x}_j, x))_{1 \leq i \leq m} \in \mathbb{R}^m$ and $\dagger$ denotes the Moore-Penrose pseudoinverse. The dimension reduced estimator is then computed as

$$\widehat{f}_{n,\lambda,m} = \underset{f \in \mathcal{H}_m}{\arg\min} \, \tfrac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(f(x_i)) + \tfrac{\lambda}{2} \|f\|_{\mathcal{H}_m}^2 \tag{2.76}$$

Using the parametrization of $\mathcal{H}_m$ in Eq. (2.35), the problem Eq. (2.76) becomes

$$\widehat{\alpha}_{n,\lambda,m} = \underset{\alpha \in \mathbb{R}^m}{\arg\min} \, \tfrac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(e_i^\top \mathbf{K}_{n,m} \alpha) + \tfrac{\lambda}{2} \alpha^\top \mathbf{K}_{mm} \alpha, \qquad \mathbf{K}_{n,m} = (k(x_i, \widetilde{x}_j))_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}, \tag{2.77}$$

and the minimizer of Eq. (2.76) can be expressed as $\widehat{f}_{n,\lambda,m} = \mathbf{k}(x)^\top \widehat{\alpha}_{n,\lambda,m}$. As in the least-squares case, a much better way parametrizing is to write $\mathbf{K}_{mm} = \mathbf{T}^\top \mathbf{T}$ using a cholesky decomposition, where $\mathbf{T}$ is upper triangular matrix. One then notes that defining $\phi(x) = \mathbf{T}^{-\top} \mathbf{k}_m(x) \in \mathbb{R}^m$, ϕ is a feature map which defines the kernel k_m : $k_m(x, x') = \phi(x)^\top \phi(x')$. Hence, $\mathcal{H}_m = \{\alpha^\top \phi(\cdot) \ : \ \alpha \in \mathbb{R}^m\}$. If Φ is the matrix whose columns are the $\Phi_i = \phi(x_i)$, that is $\Phi^\top = \mathbf{K}_{n,m} \mathbf{T}^{-1}$, the problem Eq. (2.76) can be formulated as

$$\widehat{\alpha}_{n,\lambda,m} = \underset{\alpha \in \mathbb{R}^m}{\arg\min} \, \tfrac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(\Phi_i^\top \alpha) + \frac{\lambda}{2} \|\alpha\|^2, \tag{2.78}$$

and the minimizer of Eq. (2.76) can be expressed as $\widehat{f}_{n,\lambda,m} = \mathbf{k}(x)^\top (\mathbf{T}^{-1} \widehat{\alpha}_{n,\lambda,m})$. Of course, this comes at the price of computing and saving a cholesky factor $\mathbf{T}$, which is $O(m^3)$ in time and $O(m^2)$ in memory.

The following theorem presents a bias variance trade-off in the case where the dimension is reduced using Nyström subsampling. This is a rewriting of a theorem by Marteau-Ferey, Bach, and Rudi (2019) which can be found in Theorem 4.6.

Main theorem 2: Marteau-Ferey, Bach, and Rudi (2019)

There exists explicit constants n_0, n_1, m_0, m_1, C, and $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ as well as a constant λ_0 depending on b_2 such that for any $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \lambda_0$, whenever

$$n \geq n_0 \, \frac{1+\mathsf{b}_2}{\lambda} \log \frac{1+\mathsf{b}_2}{\lambda\delta}, \;\; n \geq n_1 \, \frac{\mathsf{df}_\lambda \vee \mathsf{q}^2}{\mathsf{r}_\lambda^2} \, \log \frac{2}{\delta}, \;\; \mathsf{C}\sqrt{b_\lambda} \leq \mathsf{r}_\lambda/2,$$

with $\mathsf{q}^2 = \mathsf{b}_1^2/\mathsf{b}_2$. Assume that dimension reduction has been applied using Nyström subsampling, where the samples are obtained with either one of the following values of m and subsampling techniques.

(i) $m \geq m_0 \frac{1+\mathsf{b}_2}{\lambda} \log \frac{8(1+\mathsf{b}_2)}{\lambda\delta}$ using uniform sampling.

(ii) $m \geq m_1 d_{\lambda/(1+\mathsf{b}_2)} \log \frac{8(1+\mathsf{b}_2)}{\lambda\delta}$ using approximate leverage scores with $t = \lambda/(1+\mathsf{b}_2)$.

With probability at least $1-\delta$, it holds

$$\mathcal{R}(\widehat{f}_{n,\lambda,m}) - \mathcal{R}(f_\rho) \leq \mathsf{C}_{\text{bias}} \, b_\lambda + \mathsf{C}_{\text{var}} \, \frac{\mathsf{df}_\lambda \vee \mathsf{q}^2}{n} \, \log \frac{2}{\delta}, \tag{2.79}$$

and $\mathfrak{R}\|\widehat{f}_{n,\lambda,m}\| \leq \mathfrak{R}\|f_\rho\| + 10$.

The proof of this proposition is done in chapter 4. To incorporate the fact that we reduce the dimension, the key quantity to bound is $\|\mathbf{H}^{1/2}(I - P_m)\|$ where P_m is the orthogonal projection on $\mathcal{H}_m$ (the subset of $\mathcal{H}$ defined by the Nyström points). We show that approximating $\mathcal{H}$ with $\mathcal{H}_m$ is statistically optimal as soon as $\|\mathbf{H}^{1/2}(I - P_m)\| \leq c\,\sqrt{\lambda}$ for a certain constant c (see Sec. 4.H .3 for more details). There is a subtlety here in the case where sampling with leverage scores is used. Indeed, the ideal goal is to approximate the operator $\mathbf{H}$ with as few Nystrom points as possible. However, we cannot compute any kind of approximation of the leverage scores of the associated empirical operator $\widehat{\mathbf{H}}_n$ since f_ρ is not known ($\widehat{\mathbf{H}}_n = \widehat{\mathbf{H}}_n(f_\rho)$). Instead, we know how to approximate leverage scores of the empirical covariance $\widehat{\Sigma}$ associated to the covariance operator Σ, and since $\mathbf{H} \preceq \mathsf{b}_2\Sigma$, we can use this to obtain a guarantee $\sqrt{\mathsf{b}_2}\|\Sigma^{1/2}(I - P_m)\| \leq c\,\sqrt{\lambda}$ using these leverage scores, which will imply the desired bound. That is why in point (ii) of Main theorem 2, the effective dimension of the covariance operator $d_{\lambda/(1+\mathsf{b}_2)}$ appears.

The previous results leads to the following **informal** upper rate of convergence, which is a reformulation of a sub-case of Theorem 4.7.

Main corollary 2: Marteau-Ferey, Bach, and Rudi (2019), Theorem 7

Fix $\mathsf{b}_1, \mathsf{b}_2 > 0$, $r > 1/2$, $Q > 0$,$R > 0$, and $b \geq 1$ and let $\mathcal{M} = \mathcal{M}(R, r, \mathsf{Q}, b, \mathsf{b}_1, \mathsf{b}_2)$. Let

$$\lambda = \left(\frac{\mathsf{Q}^2}{R^2 n}\right)^{\frac{1}{2br+1}}, \tag{2.80}$$

and let $\delta \in (0, 1/2]$. Assume that dimension reduction has been applied using Nyström subsampling, where the samples are obtained with either one of the following values of m and subsampling techniques.

(i) $m \geq m_0 \frac{1+\mathsf{b}_2}{\lambda} \log \frac{8(1+\mathsf{b}_2)}{\lambda\delta}$ using uniform sampling.

(ii) $m \geq m_1 d_{\lambda/(1+\mathsf{b}_2)} \log \frac{8(1+\mathsf{b}_2)}{\lambda\delta}$ using approximate leverage scores with $t = \lambda/(1+\mathsf{b}_2)$.

With probability at least $1 - 2\delta$, it holds

$$\forall \rho \in \mathcal{M},\ \mathcal{R}(\widehat{f}_{n,\lambda,m}) - \mathcal{R}(f_\rho) \leq \mathsf{C}\ R^2 \left(\frac{\mathsf{Q}^2}{R^2 n}\right)^{\frac{2br}{2br+1}} \log \frac{2}{\delta}, \tag{2.81}$$

where C is an explicit constant, provided $n \geq N$ and where N can be characterized using the proof of Theorem 4.7, depends on all the parameters of the model $\mathcal{M}$ as well as δ, (in particular, it depends polynomially on $\log \frac{1}{\delta}$).

Note that the above result can be made formal, but for the sake of readability, we have left it in this form. Note that contrary to the least squares case, we do not explicitely give the order of m as a function of n. This could be done by assuming another eigenvalue decrease for the effective dimension of the covariance matrix in the model parameters. However, for the sake of readability, we will keep this implicit.

2.3.3 A globally convergent optimization algorithm for GSC functions

Recall from Sec. 2.2.2 that Newton methods converge very fast in the vicinity of the optimum, that is when they are initialized at a point w_0 which lies in a certain Dikin ellipsoid (see Theorem 2.9). The problem is that this region of fast convergence is not directly accessible, and can be quite small. The purpose of this section is to present an algorithm introduced by Marteau-Ferey, Bach, and Rudi (2019), whose purpose is to enter this small region.

More formally, let F be a $\mathfrak{R}$ GSC function defined on a Hilbert space H. We define F_λ to be the λ-regularizer version of F, *i.e.*, $F_\lambda = F + \frac{\lambda}{2}\|\cdot\|_H^2$. Let $\mathsf{r}_\lambda := \sqrt{\lambda}/\mathfrak{R}$. We denote with $\overline{\mathsf{D}}(F_\lambda, c)$ the modified Dikin ellipsoid :

$$\overline{\mathsf{D}}(F_\lambda, c) := \{w \in H \ :\ \nu(F_\lambda, w) \leq c\ \mathsf{r}_\lambda\}. \tag{2.82}$$

Note that by definition of $\mathsf{r}(F_\lambda)$, we have $\mathsf{r}_\lambda \leq \mathsf{r}(F_\lambda)$ and hence $\overline{\mathsf{D}}(F_\lambda, c) \subset \mathsf{D}(F_\lambda, c)$. Moreover, we will denote with $\mathbf{H}(w)$, $\mathbf{H}_\lambda(w)$ the Hessians of F, F_λ at w, and with w_λ the minimizer of F_λ for $\lambda > 0$.

The aim of the section is to present an algorithm which minimizes F_λ up to ε error : we want to find x such that $\nu(F_\lambda, x)^2, F(x) - F_\lambda(x_\lambda) \leq \varepsilon$. To do so, we will proceed in two steps.

- We will start by showing a variant of Theorem 2.9 which shows that we can actually perform an approximate Newton method rather than an exact Newton method in the modified

Dikin ellipsoid, where the price to pay is a linear rather than quadratic convergence towards the optimum. This will be useful when solving the dimension reduced problem Eq. (2.78), where computing an exact Newton step would be too costly.

- We then explain the core of the algorithm by Marteau-Ferey, Bach, and Rudi (2019) to reach the Dikin ellipsoid of F_λ. Similarly to interior point methods, this method is based on approximately minimizing F_{λ_t} for a decreasing sequence $(\lambda_t)_{0\leq t\leq T}$ where $\lambda_T = \lambda$. It is based on the fact that the approximation $\widetilde{w}_{\lambda_t}$ of the minimizer w_{λ_t} will be in a modified Dikin ellipsoid of $F_{\lambda_{t+1}}$ with good properties, thus guaranteeing that we can perform a fast minimization of $F_{\lambda_{t+1}}$ starting from $\widetilde{w}_{\lambda_t}$ using an approximate Newton method. Crucially Marteau-Ferey, Bach, and Rudi (2019) show that under certain conditions, this method is independent of the conditioning of the problem (up to log factors).

At the end of this section, we will also show how approximate Newton steps can be computed in the setting where F_λ is given by Eq. (2.78), our dimension reduced problem.

The rest of this section is greatly inspired from certain sections in the work by Marteau-Ferey, Bach, and Rudi (2019), and in particular Secs. 4.2 , 4.3 , 4.B and 4.C .

Approximate Newton methods for GSC losses

In Theorem 2.9, we saw the behavior of a full Newton method inside the Dikin ellipsoid $\mathsf{D}(F_\lambda, \mathsf{c})$ for $\mathsf{c} \leq 1/2$, and *a fortiori* in the modified Dikin ellipsoid $\overline{\mathsf{D}}(F_\lambda, \mathsf{c})$ for $\mathsf{c} \leq 1/2$. However, computing the full Newton step $\Delta(F_\lambda, w)$ at point $w \in H$ requires the exact solving of a linear system, which can be very expensive when the dimension of H is large. A natural idea is to approximate the Newton iteration, leading to *approximate Newton methods* (ANM), which take the form

$$w_{t+1} = w_t - \widetilde{\Delta}(F_\lambda, w_t), \qquad \widetilde{\Delta}(F_\lambda, w_t) \approx \Delta(F_\lambda, w_t), \tag{2.83}$$

and the symbol $\approx$ will be formally defined. Marteau-Ferey, Bach, and Rudi (2019) generically consider any technique to compute approximate Newton steps $\widetilde{\Delta}(F_\lambda, w)$ which are *ρ-relative approximation* of $\Delta(F_\lambda, w)$, defined as follows (Deuflhard, 2011).

Definition 2.4 (relative approximation)**.** *Let $\rho < 1$, let $\mathbf{A}$ be an invertible positive definite Hermitian operator on H and b in H. A ρ-relative approximations of $z^* = \mathbf{A}^{-1}b$ is an element z satisfying $\|z - z^*\|_{\mathbf{A}} \leq \rho\|z^*\|_{\mathbf{A}}$. We denote with* $\mathrm{LinApprox}(\mathbf{A}, b, \rho)$ *the set of all ρ-relative approximations of $z^* = \mathbf{A}^{-1}b$.*

A ρ-relative approximation of the full Newton step of F_λ at w will therefore be any element $\widetilde{\Delta}(F_\lambda, w) \in \mathrm{LinApprox}(\mathbf{H}_\lambda(w), \nabla F_\lambda(w), \rho)$. Moreover, we will say that w is the result of k steps of an approximate Newton method of parameter $\rho < 1$ starting at w_0, and write $w \in \mathtt{ANM}_\rho(F, w_0, k)$, if there exists $w_1, ..., w_k$ such that $w = w_k$ and for any $t \in \{1, ..., k\}$, we have $w_{t-1} - w_t \in \mathrm{LinApprox}(\mathbf{H}_\lambda(w_{t-1}), \nabla F_\lambda(w_{t-1}), \rho)$.

The following result, which can be found in Lemma 4.2 and proved in Lemma 4.11 in Sec. 4.B .3, shows that when $w_0 \in \overline{\mathsf{D}}(F_\lambda, \mathsf{c})$ for a sufficiently small c, the convergence of the approximate Newton method is linear and does not depend on the condition number of the problem, as was the case for full Newton methods. Note that this type of result is not new, and had already been obtained, if not formalized in exactly the same way, in a work by Sun and Tran-Dinh (2017).

Proposition 2.4 (behavior of the approximate Newton method)**.** *Let* $\mathsf{c}_0 = 1/7$ *and* $\rho_0 = 1/7$. *For any* $w \in \overline{\mathsf{D}}(F_\lambda, \mathsf{c}_0)$, *it holds :*

$$\frac{1}{4}\nu^2(F_\lambda, w) \leq F_\lambda(w) - F_\lambda(w_\lambda) \leq \nu^2(F_\lambda, w). \tag{2.84}$$

As soon as $\widetilde{\Delta}(F_\lambda, w) \in \mathrm{LinApprox}(\mathbf{H}_\lambda(w), \nabla F_\lambda(w), \rho)$ *for* $\rho \leq \rho_0$, *it holds*

$$\nu(F_\lambda, w - \widetilde{\Delta}(F_\lambda, w)) \leq \tfrac{1}{2}\nu(F_\lambda, w). \tag{2.85}$$

Thus, if $w_0 \in \overline{\mathsf{D}}(F_\lambda, \mathsf{c}_0)$, *then any* $w \in \mathtt{ANM}_{\rho_0}(F_\lambda, w_0, k)$ *satisfies* $\nu(F_\lambda, w) \leq 2^{-k}\nu(F_\lambda, w_0)$.

Globally convergent scheme for ANM algorithms on GSC functions

We are now ready to introduce the globally convergent scheme proposed by Marteau-Ferey, Bach, and Rudi (2019). Note that in the literature, some other globalization schemes arrive to regions of interest by first-order methods or back-tracking schemes Agarwal, Bullins, and Hazan (2017); A. Erdogdu and Montanari (2015). However such approaches require a number of steps that is usually proportional to $\sqrt{L/\lambda}$, making them depend on the condition number of the problem.

Instead, our algorithm is based on the previous observation that as soon as $w \in \overline{\mathsf{D}}(F_\lambda, \mathsf{c}_0)$ (with say $\mathsf{c}_0 = 1/7$), t steps of ANM converge as fast as 2^{-t} : it is therefore interesting to apply ANM only in Dikin ellipsoids. Our idea is to start from a very large regularization parameter λ_0, such that we are sure that $w_0 = 0$ is in the convergence region $\overline{\mathsf{D}}(F_{\lambda_0}, w_0)$, and perform some steps of an approximate Newton methods to obtain an approximation w_1 of w_{λ_0} such that the solution enters in the Dikin ellipsoid of F_{λ_1}, for a certain $\lambda_1 < \lambda_0$, and to iterate this procedure until we enter the convergence region of F_{λ_K} where $\lambda_K = \lambda$. Formally, we propose the following globalization scheme, where the function $q(\cdot, \cdot)$ will help define the next λ_k from the previous one.

Proposed Globalization Scheme

Phase I: Getting in the Dikin ellispoid of F_λ

Inputs : $w_0 \in H$, $\lambda_0 > 0$, $t \in \mathbb{N}$, $T \in \mathbb{N}$, $\rho < 1$ and a function $q : H \times \mathbb{R}_+ \to [0, 1)$.

$k \leftarrow 0$

While $\lambda_k > \lambda$

$\quad w_{k+1} \leftarrow w \in \mathtt{ANM}_\rho(F_{\lambda_k}, w_k, t)$

$\quad \lambda_{k+1} \leftarrow \max(q(w_{k+1}, \lambda_k)\lambda_k, \lambda)$

$\quad k \leftarrow k + 1$

Phase II: reach a certain precision starting from inside the Dikin ellipsoid

Output : $w \in \mathtt{ANM}_\rho(F_\lambda, w_k, T)$

The main ingredient to guarantee that the scheme will work is the following lemma (see Lemma 4.13 in Sec. 4.C .1 for a proof).

Lemma 2.1. *Let* $\mathsf{c} < 1$ *and* $w \in H$. *Let* $q_\mathsf{c}(w) = 1 - \frac{2}{3(1+\mathfrak{R}\|w\|/\mathsf{c})}$.

$$\forall q \in [q_\mathsf{c}(w), 1),\ \forall \lambda > 0,\ w \in \overline{\mathsf{D}}(F_\lambda, \mathsf{c}/3) \implies w \in \overline{\mathsf{D}}(F_{q\lambda}, \mathsf{c}). \tag{2.86}$$

This will allow to show the loop invariant $w_k \in \overline{\mathsf{D}}(F_{\lambda_k}, \mathsf{c})$. Indeed assume that $w_{k-1} \in \overline{\mathsf{D}}(F_{\lambda_{k-1}}, \mathsf{c})$. Then $\nu(F_{\lambda_{k-1}}, w_{k-1}) \leq \mathsf{c}\sqrt{\lambda_{k-1}}/\mathfrak{R}$. By taking $t = 2$ and $\rho = 1/7$ as parameters of the

approximate Newton method, and settint $w_k = \texttt{ANM}_\rho(F_{\lambda_{k-1}}, w_{k-1}, t)$, by proposition 2.4, $w_k \in \overline{\mathsf{D}}(F_{\lambda_{k-1}}, \mathsf{c}/4)$. Setting $q = q(w_k)$, this implies that $w_k \in \overline{\mathsf{D}}(F_{q\lambda_{k-1}}, \mathsf{c}) = \overline{\mathsf{D}}(F_{\lambda_k}, \mathsf{c})$, by Lemma 2.1. Now we are ready to state our main theorem of this section.

Main theorem 3: Marteau-Ferey, Bach, and Rudi (2019), Theorem 1.

Let $\epsilon > 0$. If the proposed globalization scheme is performed with the following input parameters :

- $\lambda_0 = 7\mathfrak{R}\|\nabla F(0)\|$ and $w_0 = 0$ for the initialization points;
- $\rho \leq 1/7$ and $t = 2$ for the approximate Newton methods in phase I;
- $q(w, \lambda) = q_{1/7}(w)$ in order to update the value of λ;
- $T = \lceil \log_2 \sqrt{1 \vee (\lambda\epsilon^{-1}/\mathfrak{R}^2)} \rceil$ in the last approximate Newton method in phase II.

The proposed scheme with these parameters finishes, and if K denotes the number of passes through the Phase I loop, it holds

$$F_\lambda(w) - F_\lambda(w_\lambda) \leq \nu(F_\lambda, w)^2 \leq \epsilon, \qquad K \leq \lfloor (3 + 11\mathfrak{R}\|w_\lambda\|) \log(7\mathfrak{R}\|\nabla F(0)\|/\lambda) \rfloor .$$

Note that the theorem above (proven in Sec. 4.C .3) guarantees a solution with error ϵ with K steps of ANM each performing 2 iterations of approximate linear system solving, plus a final step of ANM which performs T iterations of approximate linear system solving.

Remarks. The proposed method does not depend on the condition number of the problem, but on the term $\mathfrak{R}\|w_\lambda\|$ which can be in the order of $\mathfrak{R}/\sqrt{\lambda}$ in the worst case, but is usually way smaller. For example, it is possible to prove that this term is bounded by an absolute constant not depending on λ, if at least one minimizer of F exists. Note that in the statistical setting, *i.e.*, when solving Eq. (2.78), then $\|w_\lambda\|$ is bounded by a term coming from the statistical problem, as can be seen in Main theorem 2.

In proposition 4.7, Marteau-Ferey, Bach, and Rudi (2019) show a variant of this adaptive method which can leverage the regularity of the solution with respect to the Hessian, i.e., depending on the smaller quantity $\mathfrak{R}\sqrt{\lambda}\|w_\lambda\|_{\mathbf{H}_\lambda^{-1}(w_\lambda)}$ instead of $\mathfrak{R}\|w_\lambda\|$.

Finally note that it is possible to use $q_k = q$ fixed for all the iterations and way smaller than the one in Theorem 4.1, depending on some regularity properties of $\mathbf{H}$ (see proposition 4.8 in Sec. 4.C .2).

Methods for computing approximate Newton steps in the setting of decomposable losses

In the previous section, our algorithm to compute an ε approximation of the minimizer of F_λ relies crucially on the capacity to perform approximate Newton methods for $F_{\widetilde{\lambda}}$ with $\widetilde{\lambda} \geq \lambda$. In this section, we will present different methods to compute such approximate Newton steps. They all approximate the true Newton step $\Delta(F_\lambda, w)$ with $\widetilde{\Delta}(F_\lambda, w)$ of the form $\widetilde{\mathbf{H}}_\lambda(w)^{-1}\nabla F_\lambda(w)$, where $\widetilde{\mathbf{H}}_\lambda(w)$ is an approximation of the Hessian $\mathbf{H}_\lambda(w)$ satisfying

$$\tfrac{1}{\sqrt{\kappa}}\mathbf{H}_\lambda(w) \preceq \widetilde{\mathbf{H}}_\lambda(w) \preceq \sqrt{\kappa}\mathbf{H}_\lambda(w), \tag{2.87}$$

for some $\kappa \geq 1$. In that case, it is easy to prove that $\widetilde{\Delta}(F_\lambda, w) \in \mathrm{LinApprox}(\mathbf{H}_\lambda(w), \nabla F_\lambda(w), \rho)$ with $\rho = 1 - 1/\sqrt{\kappa}$.

Assume now that F is defined on a finite dimensional Hilbert space $\mathbb{R}^m$, and is of the form

$$F(w) = \tfrac{1}{n}\sum_{i=1}^{n} \ell_i(\Phi_i^\top w), \tag{2.88}$$

where the Φ_i are bounded by one, and the ℓ_i are $\mathfrak{R}$-GSC. Let $\mathsf{b}_2(w) = \sup_i \ell_i^{(2)}(\Phi_i^\top w)$. The Hessian of F at w is in the form :

$$\nabla^2 F(w) = \tfrac{1}{n}\Psi_w \Psi_w^\top, \qquad \Psi_w = \Phi \operatorname{diag}\left(\sqrt{\ell_1^{(2)}(\Phi_i^\top w)}\right) \in \mathbb{R}^{m\times n}. \tag{2.89}$$

Note that in Sec. 2.1.3, we introduced sketching and subsampling methods which exactly compute $\widetilde{\Psi}_w \in \mathbb{R}^{m\times q}$ such that $\widetilde{\mathbf{H}}_\lambda = \frac{1}{n}\widetilde{\Psi}_w\widetilde{\Psi}_w^\top + \lambda\mathbf{I}$ satisfies Eq. (2.87) with probability at least $1-\delta$ as soon as $q \geq d_\lambda(\Psi_w)\log\frac{\mathsf{b}_2(w)}{\lambda\delta}$, where $d_\lambda(\Psi_w) = \operatorname{Tr}(\mathbf{H}_\lambda(w)^{-1}\mathbf{H}(w)) \leq d_{\lambda/\mathsf{b}_2(w)}(\Phi)$.

If, for example, we apply the sketching method by Pilanci and Wainwright (2017) to perform the approximate Newton steps (with κ such that $\rho \leq 1/7$), the total complexity of global method can be bounded by

$$\begin{aligned} &O\left((nm\log n + mq^2 + q^3)\left(\mathfrak{R}\|w_\lambda\|\log\frac{\mathfrak{R}\|\nabla F(0)\|}{\lambda} + \log_2(1 + (\lambda\epsilon^{-1}/\mathfrak{R}^2))\right)\right) \text{ in time,} \\ &q = O\left(d_{\lambda/\bar{\mathsf{b}}_2}\log\frac{\bar{\mathsf{b}}_2}{\lambda\delta}\right), \end{aligned} \tag{2.90}$$

where $\bar{\mathsf{b}}_2 = \sup_{w\in\Gamma_\lambda}\mathsf{b}_2(w)$, is the supremum of the bounds on the second derivatives of the ℓ_i over the path of regularized minimizers $\Gamma_\lambda = \{w_{\lambda'} \ : \ \lambda' \geq \lambda\}$, which included in the ball centered at 0 and of radius $\|w_\lambda\|$.

2.3.4 Statistical bounds for the whole algorithm

Note that the dimension reduced problem can be cast exactly as a problem of the form Eq. (2.88). Recall that in that case, $\widehat{f}_{n,\lambda,m} = f_{\widehat{\alpha}_{n,\lambda,m}}$ where $f_\alpha = \alpha^\top\phi(x)$, for $\phi(x)$ defined just before Eq. (2.78), and where $\widehat{\alpha}_{n,\lambda,m}$ is defined by

$$\widehat{\alpha}_{n,\lambda,m} = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^m} F_\lambda(\alpha), \qquad F(\alpha) = \tfrac{1}{n}\sum_{i=1}^{n}\ell_{z_i}(\Phi_i^\top\alpha), \tag{2.78}$$

and which can be optimized using the global scheme, with approximate Newton steps computed using one of the techniques explained in Sec. 2.1.3. We now state an *informal* result, which gives the spirit of the results by Marteau-Ferey, Bach, and Rudi (2019), Theorem 4.4, or in Sec. 4.D .6 on the complexity and statistical performance of the algorithm. For the sake of simplicity, assume here that the $\ell_z^{(2)}$ are all bounded by a constant b_2. Note that this result is the analog of the result obtained in the least-squares case for FALKON.

Main theorem 4: Marteau-Ferey, Bach, and Rudi (2019), Proposition 14

There exists explicit constants n_0, n_1, m_0, m_1, C, and $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ as well as a constant λ_0 depending on b_2 such that the following hold. Let $\varepsilon > 0$, $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \lambda_0$ and assume

$$n \geq n_0 \, \frac{1+\mathsf{b}_2}{\lambda} \log \frac{1+\mathsf{b}_2}{\lambda\delta}, \;\; n \geq n_1 \, \frac{\mathsf{df}_\lambda \vee \mathsf{q}^2}{\mathsf{r}_\lambda^2} \, \log \frac{2}{\delta}, \;\; \mathsf{C}(\sqrt{b_\lambda} + \sqrt{\varepsilon}) \leq \mathsf{r}_\lambda/2,$$

with $\mathsf{q}^2 = \mathsf{b}_1^2/\mathsf{b}_2$. Assume that dimension reduction has been applied using Nyström subsampling using approximate leverage scoresassociated with the empirical covariance with $t = \lambda/(1+\mathsf{b}_2)$ and

$$m \geq m_1 d_{\lambda/(1+\mathsf{b}_2)} \log \frac{8(1+\mathsf{b}_2)}{\lambda\delta}.$$

Assume we perform the globalization scheme using the parameters given in Main theorem Main theorem 3 with precision ε, and that we compute approximate Newton steps using a sketching method as the one by Pilanci and Wainwright (2017), in order to obtain $\widetilde{f}_{n,\lambda,m}$. With probability at least $1-\delta$, it holds

$$\mathcal{R}(\widetilde{f}_{n,\lambda,m}) - \mathcal{R}(f_\rho) \leq \mathsf{C}_{\text{bias}} \, b_\lambda + \mathsf{C}_{\text{var}} \, \frac{\mathsf{df}_\lambda \vee \mathsf{q}^2}{n} \, \log \frac{2}{\delta} + \varepsilon, \tag{2.91}$$

and the time complexity necessary to compute $\widetilde{f}_{n,\lambda,m}$ is of order

$$\begin{aligned} &(nm \log n + mq^2 + q^3) \left(\mathfrak{R}\|f_\rho\| \log \frac{\mathfrak{R}\|\nabla F(0)\|}{\lambda} + \log_2(1 + (\lambda\epsilon^{-1}/\mathfrak{R}^2)) \right) \\ &m, q = O\left(d_{\lambda/\mathsf{b}_2} \log \frac{\mathsf{b}_2}{\lambda\delta} \right), \end{aligned} \tag{2.92}$$

The "real" result is slightly more complex. In particular, it defines the effective dimension for all the Hessians that are crossed, and relates the empirical effective dimensions with the ideal ones statistically. This theorem can be extended to obtain upper rates of convergence as the previous corollaries (see Theorem 7 by Marteau-Ferey, Bach, and Rudi (2019)).

The main takeaway of this result is that we can compute a statistically optimal estimator in time roughly of order $O(nd_\lambda + d_\lambda^3)$ (we hide logarithmic constants). In particular, in the well-specified case, Eq. (2.92) shows that the complexity is independent from the condition number of the problem. This is a very desirable property in classification problems where we are brought to consider very small values of λ, such as in the two examples in Fig. 2.1 taken from the work by Marteau-Ferey, Bach, and Rudi (2019), where we compare our method with a first order competitor, which has a worse dependance on the condition number.

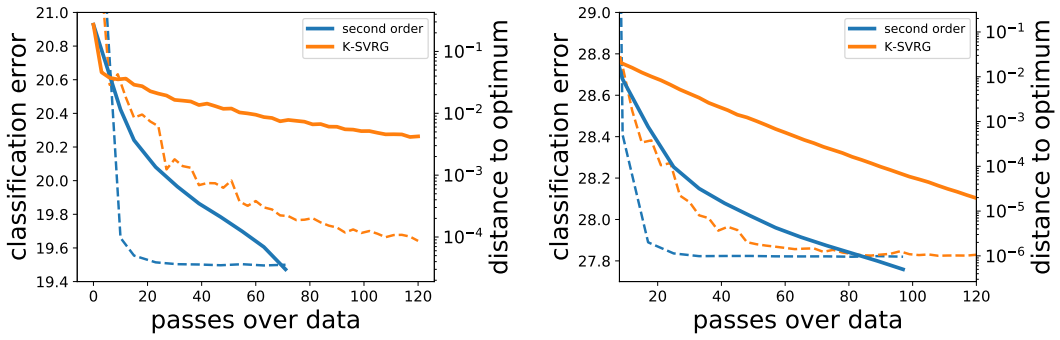

Figure 2.1: Training loss and test error as as function of the number of passes on the data for our algorithm vs. K-SVRG. on the **(left)** Susy and **(right)** Higgs data sets.

Chapter 3

Fast rates for regularized empirical risk minimization

This chapter is a verbatim of the work :

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2294–2340. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/marteau-ferey19a.html.

Contents

3.1 Introduction . . . 94
3.2 Main Assumptions and Results . . . 96
3.3 Slow convergence rates . . . 99
3.4 Faster Rates with Source Conditions . . . 100
3.5 Rates for source and capacity . . . 101
3.6 Sketch of the proof . . . 103
3.7 Conclusion . . . 104
3.A Setting, definitions, assumptions . . . 105
3.B Generalized self-concordant losses . . . 107
3.C Main result, simplified . . . 110
3.D Main result, refined analysis . . . 115
3.E Explicit bounds for the simplified case . . . 123
3.F Explicit bounds for the refined case . . . 125
3.G Additional lemmas . . . 129

3.1 Introduction

Regularized empirical risk minimization remains a cornerstone of statistics and supervised learning, from the early days of linear regression (Hoerl and Kennard, 1976) and neural networks (Geman, Bienenstock, and Doursat, 1992), then to spline smoothing (Wahba, 1990) and more generally kernel-based methods (Shawe-Taylor and Cristianini, 2004). While the regularization by the squared Euclidean norm is applied very widely, the statistical analysis of the resulting learning methods is still not complete.

The main goal of this paper is to provide a sharp non-asymptotic analysis of regularized empirical risk minimization (ERM), or more generally regularized M-estimation, that is estimators obtained as the unique solution of

$$\min_{\theta \in \mathcal{H}} \quad \frac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(\theta) + \frac{\lambda}{2} \|\theta\|^2, \tag{3.1}$$

where $\mathcal{H}$ is a Hilbert space (possibily infinite-dimensional) and $\ell_z(\theta)$ is the convex loss associated with an observation z and the estimator $\theta \in \mathcal{H}$. We assume that the observations z_i, $i = 1, \dots, n$ are independent and identically distributed, and that the minimum of the associated unregularized expected risk $L(\theta)$ is attained at a certain $\theta^\star \in \mathcal{H}$.

In this paper, we focus on dimension-independent results (thus ultimately extending the analysis in the finite-dimensional setting from Ostrovskii and Bach, 2018). For this class of problems, two main classes of problems have been studied, depending on the regularity assumptions on the loss.

Convex Lipschitz-continuous losses (with respect to the parameter θ), such as for logistic regression or the support vector machine, lead to general *non-asymptotic* bounds for the excess risk of the form (Sridharan, Shalev-Shwartz, and Srebro, 2009):

$$\frac{B^2}{\lambda n} + \lambda \|\theta^\star\|^2, \tag{3.2}$$

where B is a uniform upper bound on the Lipschitz constant for all losses $\theta \mapsto \ell_z(\theta)$. The bound above already has a form that takes into account two separate terms: a *variance term* $B^2/(\lambda n)$ which depends on the sample size n but not on the optimal predictor $\theta^\star$, and a *bias term* $\lambda \|\theta^\star\|^2$ which depends on the optimal predictor but not on the sample size n. All our bounds will have this form but with smaller quantities (but asking fore more assumptions). Without further assumptions, in Eq. (3.2), λ is taken proportional to $1/\sqrt{n}$, and we get the usual optimal slow rate in excess risk of $O(1/\sqrt{n})$ associated with such a general set-up (see, e.g., Cesa-Bianchi, Mansour, and Shamir, 2015).

For the specific case of quadratic losses of the form $\ell_z(\theta) = \frac{1}{2}(y - \theta \cdot \Phi(x))^2$, where $z = (x, y)$, and $y \in \mathbb{R}$ and $\Phi(x) \in \mathcal{H}$, the situation is much richer. Without further assumptions, the same rate $O(1/\sqrt{n})$ is achieved, but stronger assumptions lead to faster rates (Caponnetto and De Vito, 2007). In particular, the decay of the eigenvalues of the Hessian $\mathbb{E}\big[\Phi(x) \otimes \Phi(x)\big]$ (often called the *capacity condition*) leads to an improved variance term, while the finiteness of some bounds on $\theta^\star$ for norms other than the plain Hilbertian norms $\|\theta^\star\|$ (often called the *source condition*) leads to an improved bias term. Both of these assumptions lead to faster rates than $O(1/\sqrt{n})$ for the excess risk, with the proper choice of the regularization parameter λ. For least-squares, these rates are then optimal and provide a better understanding of properties of the problem that influence the generalization capabilities of regularized ERM (see, e.g. Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Steinwart, Hush, and Scovel, 2009; Fischer and Steinwart, 2017; Blanchard and Mücke, 2018).

Our main goal in this paper is to bridge the gap between Lipschitz-continuous and quadratic losses by improving on slow rates for general classes of losses beyond least-squares. We first note that: (a) there has to be an extra regularity assumption because of lower bounds (Cesa-Bianchi, Mansour, and Shamir, 2015), and (b) asymptotically, we should obtain bounds that approach the local quadratic approximation of $\ell_z(\theta)$ around $\theta^\star$ with the same optimal behavior as for plain least-squares.

Several frameworks are available for such an extension with extra assumptions on the losses, such as "exp-concavity" (Koren and Levy, 2015; Mehta, 2016), strong convexity (Van de Geer, 2008) or a generalized notion of self-concordance (Bach, 2010; Ostrovskii and Bach, 2018). In this paper, we focus on self-concordance, which links the second and third order derivatives of the loss. This notion is quite general and corresponds to widely used losses in machine learning, and does not suffer from constants which can be exponential in problem parameters (e.g., $\|\theta^\star\|$) when applied to generalized linear models like logistic regression. See Sec. 3.1.1 for a comparison to related work.

With this self-concordance assumption, we will show that our problem behaves like a quadratic problem corresponding to the local approximation around $\theta^\star$, in a totally non-asymptotic way, which is the core technical contribution of this paper. As we have already mentioned, this phenomenon is naturally expected in the asymptotic regime, but is hard to capture in the non-asymptotic setting without constants which explode exponentially with the problem parameters.

The paper is organized as follows: in Sec. 3.2, we present our main assumptions and informal results, as well as our bias-variance decomposition. In order to introduce precise results gradually, we start in Sec. 3.3 with a result similar to Eq. (3.2) for our set-up to show that we recover with a simple argument the result from Sridharan et al. (2009), which itself applies more generally. Then, in Sec. 3.4 we introduce the source condition allowing for a better control of the bias. Finally, in Sec. 3.5, we detail the capacity condition leading to an improved variance term, which, together with the improved bias leads to fast rates (which are optimal for least-squares).

3.1.1 Related work

Fast rates for empirical risk minimization. Rates faster than $O(1/\sqrt{n})$ can be obtained with a variety of added assumptions, such as some form of strong convexity (Sridharan, Shalev-Shwartz, and Srebro, 2009; Boucheron and Massart, 2011), noise conditions for classification (Steinwart and Scovel, 2007), or extra conditions on the loss, such as self-concordance (Bach, 2010) or exp-concavity (Koren and Levy, 2015; Mehta, 2016), whose partial goal is to avoid exponential constants. Note that Bach (2010) already considers logistic regression with Hilbert spaces, but only for well-specified models and a fixed design, and without the sharp and simpler results that we obtain in this paper.

Avoiding exponential constants for logistic regression. The problem of exponential constants (i.e., leading factors in the rates scaling as e^{RD} where D is the radius of the optimal predictor, and R the radius of the design) is long known. In fact, Hazan et al. (2014) showed a lower bound, explicitly constructing an adversarial distribution (i.e., an ill-specified model) for which the problem manifests in the finite-sample regime with $n = O(e^{RD})$. Various attempts to address this problem are found in the literature. For example, Ostrovskii and Bach (2018, App. C) prove the optimal d/n rate in the non-regularized d-dimensional setting but, multiplied with the curvature parameter ρ which is at worst exponential but is shown to grow at most as $(RD)^{3/2}$ in the case of Gaussian design. Another approach is due to Foster et al. (2018): they

establish "1-mixability" of the logistic loss, then apply Vovk's aggregating algorithm in the online setting, and then proceed via online-to-batch conversion. While this result allows to obtain the fast $O(d/n)$ rate (and its counterparts in the nonparametric setting) without exponential constants, the resulting algorithm is *improper* (i.e., the canonical parameter $\eta = \Phi(x) \cdot \theta^\star$, see below, is estimated by a *non-linear* functional of $\Phi(x)$).

A closely related approach is to use the notion of exp-concavity instead of mixability (Rakhlin and Sridharan, 2015; Koren and Levy, 2015; Mehta, 2016). The two close notions are summarized in the so-called central condition (due to Van Erven et al. (2015)) which fully characterizes when the fast $O(d/n)$ rates (up to log factors and in high probability) are available for improper algorithms. However, when proper learning algorithms are concerned, this analysis requires η-mixability (or η-exp-concavity) of the *overall* loss $\ell_z(\theta)$ for which the η parameter scales with the radius of the set of predictors. This scaling is exponential for the logistic loss, leading to exponential constants.

3.2 Main Assumptions and Results

Let $\mathcal{Z}$ be a Polish space and Z be a random variable on $\mathcal{Z}$ with distribution ρ. Let $\mathcal{H}$ be a separable (non-necessarily finite-dimensional) Hilbert space, with norm $\|\cdot\|$, and let $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ be a loss function, we denote by $\ell_z(\cdot)$ the function $\ell(z, \cdot)$. Our goal is to minimize the expected risk with respect to $\theta \in \mathcal{H}$:

$$\inf_{\theta \in \mathcal{H}} \; L(\theta) = \mathbb{E}\left[\ell_Z(\theta)\right].$$

Given $(z_i)_{i=1}^n \in \mathcal{Z}^n$, we will consider the following estimator based on regularized empirical risk minimization given $\lambda > 0$ (note that the minimizer is unique in this case):

$$\widehat{\theta}_\lambda^\star = \operatorname*{arg\,min}_{\theta \in \mathcal{H}} \quad \frac{1}{n} \sum_{i=1}^n \ell_{z_i}(\theta) + \frac{\lambda}{2} \|\theta\|^2,$$

where we assume the following.

Assumption 3.1 (i.i.d. data)**.** *The samples $(z_i)_{1 \le i \le n}$ are independently and identically distributed according to ρ.*

The goal of this work is to provide upper bounds in high probability for the so-called *excess risk*

$$L(\widehat{\theta}_\lambda^\star) - \inf_{\theta \in \mathcal{H}} \; L(\theta),$$

and thus to provide a general framework to measure the quality of the estimator $\widehat{\theta}_\lambda^\star$. Algorithms for obtaining such estimators have been extensively studied, in both finite-dimensional regimes, where a direct optimization over θ is performed, typically by gradient descent or stochastic versions thereof (see, e.g., Bottou and Bousquet, 2008; Shalev-Shwartz, Singer, Srebro, and Cotter, 2011) and infinite-dimensional regimes, where kernel-based methods are traditionally used (see, e.g., Keerthi, Duan, Shevade, and Poo, 2005; Gerfo, Rosasco, Odone, Vito, and Verri, 2008; Dieuleveut and Bach, 2016; Tu, Roelofs, Venkataraman, and Recht, 2016; Rudi, Carratino, and Rosasco, 2017, and references therein).

Example 3.1 (Supervised learning)**.** *Although formulated as a general M-estimation problem (see, e.g., Lehmann and Casella, 2006), our main motivation comes from supervised learning, with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the data space and $\mathcal{Y}$ the target space. We will consider, as examples, losses with both real-valued outputs but also the multivariate case. For learning real-valued outputs, consider we have a bounded representation of the input space $\Phi : \mathcal{X} \to \mathcal{H}$ (potentially implicit when using kernel-based methods, Aronszajn, 1950). We will provide bounds for the following losses.*

- *The square loss* $\ell_z(\theta) = \frac{1}{2}\left(y - \theta \cdot \Phi(x)\right)^2$, *which is not Lipschitz-continuous.*
- *The Huber losses* $\ell_z(\theta) = \psi(y - \theta \cdot \Phi(x))$ *where* $\psi(t) = \sqrt{1+t^2} - 1$ *or* $\psi(t) = \log \frac{e^t + e^{-t}}{2}$ *(Hampel, Ronchetti, Rousseeuw, and Stahel, 2011), which are Lipschitz-continuous.*
- *The logistic loss* $\ell_z(\theta) = \log(1 + e^{-y\theta \cdot \Phi(x)})$ *commonly used in binary classification where* $y \in \{-1, 1\}$, *which is Lipschitz-continuous.*

Our framework goes beyond real-valued outputs, and can be applied to all generalized linear models (GLM) (McCullagh and Nelder, 1989), including softmax regression: we consider a representation function $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ *and an a priori measure* μ *on* $\mathcal{Y}$. *The loss we consider in this case is*

$$\ell_z(\theta) = -\theta \cdot \Phi(x, y) + \log \int_{\mathcal{Y}} \exp\left(\theta \cdot \Phi(x, y')\right) d\mu(y'),$$

which corresponds to the negative conditional log-likelihood when modelling y *given* x *by the distribution* $p(y|x, \theta) \sim \frac{\exp(\theta \cdot \Phi(x,y))}{\int_{\mathcal{Y}} \exp(\theta \cdot \Phi(x,y'))d\mu(y')} d\mu(y)$. *Our framework applies to all of these generalized linear models with almost surely bounded features* $\Phi(x, y)$, *such as conditional random fields (Lafferty, McCallum, and Pereira, 2001).*

We can now introduce the main technical assumption on the loss ℓ.

Assumption 3.2 (Generalized self-concordance)**.** *For any* $z \in \mathcal{Z}$, *the function* $\ell_z(\cdot)$ *is convex and three times differentiable. Moreover, there exists a set* $\varphi(z) \subset \mathcal{H}$ *such that it holds :*

$$\forall \theta \in \mathcal{H},\ \forall h, k \in \mathcal{H},\ \left|\nabla^3 \ell_z(\theta)[k, h, h]\right| \leq \sup_{g \in \varphi(z)} |k \cdot g|\ \nabla^2 \ell_z(\theta)[h, h].$$

This is a generalization of the assumptions introduced by Bach (2010), by allowing a varying term $\sup_{g\in\varphi(z)} |k \cdot g|$ instead of a uniform bound proportional to $\|k\|$. This is crucial for the fast rates we want to show.

Example 3.2 (Checking assumptions)**.** *For the losses in Example 3.1, this condition is satisfied with the following corresponding set-function* φ.

- *For the square loss* $\ell_z(\theta) = \frac{1}{2}\left(y - \theta \cdot \Phi(x)\right)^2$, $\varphi(z) = \{0\}$.
- *For the Huber losses* $\ell_z(\theta) = \psi(y - \theta \cdot \Phi(x))$, *if* $\psi(t) = \sqrt{1+t^2} - 1$, *then* $\varphi(z) = \{3\Phi(x)\}$ *and if* $\psi(t) = \log \frac{e^t + e^{-t}}{2}$, *then* $\varphi(z) = \{2\Phi(x)\}$ *(Ostrovskii and Bach, 2018). For the logistic loss* $\ell_z(\theta) = \log(1 + e^{-y\theta \cdot \Phi(x)})$, *we have* $\varphi(z) = \{y\Phi(x)\}$ *(here,* $\varphi(z)$ *is reduced to a point).*
- *For generalized linear models,* $\nabla^3 \ell_z(\theta)$ *is a third-order cumulant, and thus* $\left|\nabla^3 \ell_z(\theta)[k, h, h]\right| \leq \mathbb{E}_{p(y|x,\theta)} |k \cdot \Phi(x, y) - k \cdot \mathbb{E}_{p(y'|x,\theta)} \Phi(x, y')| \cdot |h \cdot \Phi(x, y) - h \cdot \mathbb{E}_{p(y'|x,\theta)} \Phi(x, y')|^2 \leq 2 \sup_{y\in\mathcal{Y}} |k \cdot \Phi(x, y)|\ \nabla^2 \ell_z(\theta)[h, h]$. *Therefore* $\varphi(z) = \{2\Phi(x, y'),\ y' \in \mathcal{Y}\}$ *(which is not a singleton).*

Moreover we require the following two technical assumptions to guarantee that $L(\theta)$ and its first and second derivatives are well defined for any $\theta \in \mathcal{H}$.

Assumption 3.3 (Boundedness)**.** *There exists* $R \geq 0$ *such that* $\sup_{g\in\varphi(Z)} \|g\| \leq R$ *almost surely.*

Assumption 3.4 (Definition in 0)**.** $|\ell_Z(0)|$, $\|\nabla \ell_Z(0)\|$ *and* $\mathrm{Tr}(\nabla^2 \ell_Z(0))$ *are almost surely bounded.*

The assumptions above are usually easy to check in practice. In particular, if the support of ρ is bounded, the mappings $z \mapsto \ell_z(0), \nabla \ell_z(0), \mathrm{Tr}(\nabla^2 \ell_z(0))$ are continuous, and φ is uniformly bounded on bounded sets, then they hold. The main regularity assumption we make on our statistical problems follows.

Assumption 3.5 (Existence of a minimizer)**.** *There exists* $\theta^\star \in \mathcal{H}$ *such that* $L(\theta^\star) = \inf_{\theta\in\mathcal{H}} L(\theta)$.

While Assumption 3.3 is standard in the analysis of such models (Caponnetto and De Vito, 2007; Sridharan, Shalev-Shwartz, and Srebro, 2009; Steinwart, Hush, and Scovel, 2009; Bach, 2014), Assumption 3.5 imposes that the model is "well-specified", that is, for supervised learning situations from Example 3.1, we have chosen a rich enough representation Φ. It is possible to study the non-realizable case in our setting by requiring additional technical assumptions (see Steinwart, Hush, and Scovel (2009) or discussion after (3.6)), but this is out of scope of this paper. Note that our well-specified assumption (for logistic regression for simplicity of arguments) is weaker than requiring $f^\star(x) = \mathbb{E}[Y|X]$ being equal to $\theta^\star \cdot \Phi(x)$. We can now introduce the main definitions allowing our bias-variance decomposition.

Definition 3.1 (Hessian, Bias, Degrees of freedom)**.** *Let $L_\lambda(\theta) = L(\theta) + \frac{\lambda}{2}\|\theta\|^2$; define the* expected Hessian $\mathbf{H}(\theta)$*, the regularized Hessian* $\mathbf{H}_\lambda(\theta)$*, the* bias Bias_λ *and the* degrees of freedom df_λ *as:*

$$\mathbf{H}(\theta) = \mathbb{E}\left[\nabla^2 \ell_Z(\theta)\right], \quad \textit{and} \quad \mathbf{H}_\lambda(\theta) = \mathbf{H}(\theta) + \lambda I, \tag{3.3}$$

$$\mathsf{Bias}_\lambda = \|\mathbf{H}_\lambda(\theta^\star)^{-1/2} \nabla L_\lambda(\theta^\star)\|, \tag{3.4}$$

$$\mathsf{df}_\lambda = \mathbb{E}\left[\|\mathbf{H}_\lambda(\theta^\star)^{-1/2} \nabla \ell_Z(\theta^\star)\|^2\right]. \tag{3.5}$$

Note that the bias and degrees of freedom only depend on the optimum $\theta^\star \in \mathcal{H}$ and not on the minimizer $\theta^\star_\lambda$ of the regularized expected risk. Moreover, the degrees of freedom df_λ correspond to the usual Fisher information term commonly seen in the asymptotic analysis of M-estimation (Van der Vaart, 2000; Lehmann and Casella, 2006), and correspond to the usual quantities introduced in the analysis of least-squares (Caponnetto and De Vito, 2007). Indeed, in the least-squares case, we recover exactly $\mathsf{Bias}_\lambda = \lambda \|\mathbf{C}_\lambda^{-1/2} \theta^\star\|$ and $\mathsf{df}_\lambda = \mathrm{Tr}(\mathbf{C}\mathbf{C}_\lambda^{-1})$, where $\mathbf{C}$ is the covariance operator $\mathbf{C} = \mathbb{E}[\Phi(x) \otimes \Phi(x)]$ and $\mathbf{C}_\lambda = \mathbf{C} + \lambda I$.

Our results will rely on the quadratic approximation of the losses around $\theta^\star$. Borrowing tools from the analysis of Newton's method (Nesterov and Nemirovskii, 1994), this will only be possible in the vicinity of $\theta^\star$. The proper notion of vicinity is the so-called *radius of the Dikin ellipsoid*, which we define as follows:

$$\mathsf{r}_\lambda(\theta) \quad \text{such that} \quad 1/\mathsf{r}_\lambda(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \sup_{g \in \varphi(z)} \|\mathbf{H}_\lambda{}^{-1/2}(\theta) g\|. \tag{3.6}$$

Our most refined bounds will depend whether the bias term is small enough compared to $\mathsf{r}_\lambda(\theta^\star)$. We believe that in the non realizable setting, the results we obtain would still hold when the bias term is smaller than the Dikin radius, although one would have to modify the definitions to incorporate the fact that $\theta^\star$ is not in $\mathcal{H}$. The following informal result summarizes all of our results.

Theorem 3.1 (General bound, informal)**.** *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $\lambda > 0$. Under Assumptions 3.1 to 3.5, whenever*

$$n \geq C_0 \frac{R^2 \mathsf{df}_\lambda \ \log \frac{2}{\delta}}{\lambda},$$

then with probability at least $1 - 2\delta$, it holds

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \mathsf{C}_{\mathrm{bias}} \ \mathsf{Bias}_\lambda^2 + \mathsf{C}_{\mathrm{var}} \ \frac{\mathsf{df}_\lambda \ \log \frac{2}{\delta}}{n},$$

where $C_0, \mathsf{C}_{\mathrm{bias}}$ and $\mathsf{C}_{\mathrm{var}}$ are either universal or depend only on $R\|\theta^\star\|$.

Assumptions	Bias	Variance	Optimal λ	Optimal Rate	
None	λ	$\frac{1}{\lambda n}$	$n^{-1/2}$	$n^{-1/2}$	Theorem 3.2 and Cor. 3.1
Source	λ^{2r+1}	$\frac{1}{\lambda n}$	$n^{-\frac{1}{2r+2}}$	$n^{-\frac{2r+1}{2r+2}}$	Theorem 3.3 and Cor. 3.2
Source + Capacity	λ^{2r+1}	$\frac{1}{\lambda^{1/\alpha} n}$	$n^{-\frac{\alpha}{2r\alpha+\alpha+1}}$	$n^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}}$	Theorem 3.4 and Cor. 3.3

Table 3.1: Summary of convergence rates, without constants except λ, for source condition (Asm. 3.6): $\theta^\star \in \mathrm{Im}(\mathbf{H}(\theta^\star)^r)$, $r \in (0, 1/2]$, capacity condition (Asm. 3.7): $\mathsf{df}_\lambda = O(\lambda^{-1/\alpha})$, $\alpha \geq 1$

.

This mimics a usual bias-variance decomposition, with a bias term Bias_λ^2 and a variance term proportional to df_λ/n. In particular in the rest of the paper we quantify the constants and the rates under various regularity assumptions, and specify the good choices of the regularization parameter λ. In Table 3.1, we summarize the different assumptions and corresponding rates.

3.3 Slow convergence rates

Here we bound the quantity of interest without any regularity assumption (e.g., source of capacity condition) beyond some boundedness assumptions on the learning problem. We consider the various bounds on the derivatives of the loss ℓ:

$$\mathsf{B}_1(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta)\|, \ \mathsf{B}_2(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \mathrm{Tr}(\nabla^2 \ell_z(\theta)), \ \overline{\mathsf{B}}_1 = \sup_{\|\theta\| \leq \|\theta^\star\|} \mathsf{B}_1(\theta), \ \overline{\mathsf{B}}_2 = \sup_{\|\theta\| \leq \|\theta^\star\|} \mathsf{B}_2(\theta).$$

Example 3.3 (Bounded derivatives)**.** *In all the losses considered above, assume the feature representation ($\Phi(x)$ for the Huber losses and the square loss, $y\Phi(x)$ for the logistic loss, and $\Phi(x, y)$ for GLMs) is bounded by $\bar{R}$. Then the losses considered above apart from the square loss are Lipschitz-continuous and B_1 is uniformly bounded by $\bar{R}$. For these losses, B_2 is also uniformly bounded by $\bar{R}^2$. Using Example 3.2, one can take $\bar{R}$ to be equal to a constant times R (1/2 and 1/3 for the respective Huber losses, 1 for logistic regression and 1/2 for canonical GLMs). For the square loss (where $R = 0$ because the third-order derivative is zero), $\overline{\mathsf{B}}_2 \leq \bar{R}^2$ and $\overline{\mathsf{B}}_1 \leq \bar{R}\|y\|_\infty + \bar{R}^2\|\theta^\star\|$, where $\|y\|_\infty$ is an almost sure bound on the output y.*

Theorem 3.2 (Basic result)**.** *Let $n \in \mathbb{N}$ and $0 < \lambda \leq \overline{\mathsf{B}}_2$. Let $\delta \in (0, 1/2]$. If*

$$n \geq 512 \left(\|\theta^\star\|^2 R^2 \vee 1\right) \log \frac{2}{\delta}, \ n \geq 24 \frac{\overline{\mathsf{B}}_2}{\lambda} \log \frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \ n \geq 256 \frac{R^2 \overline{\mathsf{B}}_1^2}{\lambda^2} \log \frac{2}{\delta},$$

then with probability at least $1 - 2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq 84 \frac{\overline{\mathsf{B}}_1^2}{\lambda n} \log \frac{2}{\delta} + 2\lambda\|\theta^\star\|^2. \tag{3.7}$$

This result shown in Sec. 3.C .3 as a consequence of Theorem 3.6 (also see the proof sketch in Sec. 3.6) matches the one obtained with Lipschitz-continuous losses (Sridharan, Shalev-Shwartz, and Srebro, 2009) and the one for least-squares when assuming the existence of $\theta^\star$ (Caponnetto and De Vito, 2007). The following corollary (proved as Theorem 3.8 in Sec. 3.E) gives the bound optimized in λ, with explicit rates.

Corollary 3.1 (Basic Rates)**.** *Let $\delta \in (0, 1/2]$. Under Assumptions 3.1 to 3.5, when $n \geq N, \lambda = C_0\sqrt{\log(2/\delta)/n}$, then with probability at least $1 - 2\delta$,*

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq C_1 \ n^{-1/2} \ \log^{1/2} \frac{2}{\delta}.$$

with $C_0 = 16\overline{\mathsf{B}}_1 \max(1, R), C_1 = 48\overline{\mathsf{B}}_1 \max(1, R) \max(1, \|\theta^\star\|^2)$ and with N defined in Eq. (3.41) and satisfying N=$O(\mathrm{poly}(\overline{\mathsf{B}}_1, \overline{\mathsf{B}}_2, R\|\theta^\star\|))$ where poly *denotes a polynomial function of the inputs.*

Both bias and variance terms are of order $O(1/\sqrt{n})$ and we recover up to constants terms the result of Sridharan et al. (2009). In the next section, we will improve both bias and variance terms to obtain faster rates.

3.4 Faster Rates with Source Conditions

Here we provide a more refined bound, where we introduce a *source condition* on $\theta^\star$ allowing to improve the bias term and to achieve learning rates as fast as $O(n^{-2/3})$. We first define the localized versions of $\overline{\mathsf{B}}_1, \overline{\mathsf{B}}_2$:

$$\mathsf{B}_1^\star = \mathsf{B}_1(\theta^\star), \qquad \mathsf{B}_2^\star = \mathsf{B}_2(\theta^\star),$$

and recall the definition of the bias

$$\mathsf{Bias}_\lambda = \|\mathbf{H}_\lambda(\theta^\star)^{-1/2}\nabla L_\lambda(\theta^\star)\|. \tag{3.8}$$

Note that since $\theta^\star$ is the minimizer of L, we have $\nabla L(\theta^\star) = 0$, so that $\nabla L_\lambda(\theta^\star) = \nabla L(\theta^\star) + \lambda\theta^\star = \lambda\theta^\star$, and $\mathsf{Bias}_\lambda = \lambda\|\mathbf{H}_\lambda(\theta^\star)^{-1/2}\theta^\star\|$. This characterization is always bounded by $\lambda\|\theta^\star\|^2$, but allows a finer control of the regularity of $\theta^\star$, leading to improved rates compared to Sec. 3.3.

Note that in the least-squares case, we recover exactly the bias of ridge regression $\mathsf{Bias}_\lambda = \lambda\|\mathbf{C}_\lambda^{-1/2}\theta^\star\|$, where $\mathbf{C}$ is the covariance operator $\mathbf{C} = \mathbb{E}\left[\Phi(x)\otimes\Phi(x)\right]$.

Using self-concordance, we will relate quantities at $\theta^\star$ to quantities at $\theta_\lambda^\star$ using:

$$\mathsf{t}_\lambda = \sup_{z\in\mathrm{supp}(\rho)} \sup_{g\in\varphi(z)} |(\theta_\lambda^\star - \theta^\star)\cdot g|.$$

The following theorem, proved in Sec. 3.D .4, relates Bias_λ to the excess risk.

Theorem 3.3 (Decomposition with refined bias)**.** *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Whenever*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda}\log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}, \quad n \geq \triangle_2\,\frac{(\mathsf{B}_1^\star R)^2}{\lambda^2}\,\log\frac{2}{\delta},$$

then with probability at least $1-2\delta$, it holds

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq \mathsf{C}_{\mathrm{bias}}\,\mathsf{Bias}_\lambda^2 + \mathsf{C}_{\mathrm{var}}\,\frac{(\mathsf{B}_1^\star)^2}{\lambda n}\,\log\frac{2}{\delta}, \tag{3.9}$$

where $\square_1 \leq e^{\mathsf{t}_\lambda/2}, \triangle_1 \leq 2304e^{4\mathsf{t}_\lambda}(1/2 \vee R\|\theta^\star\|), \triangle_2 \leq 256e^{2\mathsf{t}_\lambda}, \mathsf{C}_{\mathrm{bias}} \leq 6e^{2\mathsf{t}_\lambda}, \mathsf{C}_{\mathrm{var}} \leq 256e^{3\mathsf{t}_\lambda}$.

It turns out that the *radius of the Dikin ellipsoid* $\mathsf{r}_\lambda(\theta^\star)$ defined in Eq. (3.6) provides the sufficient control over the constants above: when the bias is of the same order of the radius of the Dikin ellipsoid, the quantities $\mathsf{C}_{\mathrm{bias}}, \mathsf{C}_{\mathrm{var}}, \triangle_1, \triangle_2$ become universal constants instead of depending exponentially on $R\|\theta^\star\|$, as shown by the lemma below, proved in Lemma 3.4 in Sec. 3.D .

Lemma 3.1. *When $\mathsf{Bias}_\lambda \leq \frac{\mathsf{r}_\lambda(\theta^\star)}{2}$ then $\mathsf{t}_\lambda \leq \log 2$ else $\mathsf{t}_\lambda \leq 2R\|\theta^\star\|$.*

Interestingly, regularity of $\theta^\star$, like the source condition below, can induce this effect, allowing a better dependence on λ for the bias term.

Assumption 3.6 (Source condition)**.** *There exists $r \in (0, 1/2]$ and $v \in \mathcal{H}$ such that $\theta^\star = \mathbf{H}(\theta^\star)^r v$.*

In particular we denote by $\mathsf{L} := \|v\|$. Assumption 3.6 is commonly made in least-squares regression (Caponnetto and De Vito, 2007; Steinwart, Hush, and Scovel, 2009; Blanchard and Mücke, 2018) and is equivalent to requiring that, when expressing $\theta^\star$ with respect to the eigenbasis

of $\mathbf{H}(\theta^\star)$, i.e., $\theta^\star = \sum_{j\in\mathbb{N}} \alpha_j u_j$, where λ_j, u_j is the eigendecomposition of $\mathbf{H}(\theta^\star)$, and $\alpha_j = \theta \cdot u_j$, then α_j decays as λ_j^r. In particular, with this assumption, defining $\beta_j = v \cdot u_j$,

$$\mathsf{Bias}_\lambda^2 = \lambda^2 \sum_j \frac{\alpha_j^2}{\lambda_j + \lambda} = \lambda^2 \sum_j \frac{\lambda_j^{2r}\beta_j^2}{\lambda_j + \lambda} \leq \lambda^2 \Big(\sup_j \frac{\lambda_j^{2r}}{\lambda_j + \lambda} \Big) \sum_j \beta_j^2 \leq \lambda^{1+2r} \|v\|^2.$$

Note moreover that $\mathbf{H}(\theta^\star) \preccurlyeq \mathsf{B}_2^\star \mathbf{C}$, meaning that the usual sufficient conditions leading to the source conditions for least-squares also apply here. For example, for logistic regression, if the log-odds ratio is smooth enough, then it is in $\mathcal{H}$. So, when $\mathcal{H}$ corresponds to a Sobolev space of smoothness m and the marginal of ρ on the input space is a density bounded away from 0 and infinity with bounded support, then the source condition corresponds essentially to requiring $\theta^\star$ to be $(1+2r)m$-times differentiable (see discussion after Thm. 9 of Steinwart, Hush, and Scovel, 2009, for more details). A precise example can be found in Sec. 4.1 of Pillaud-Vivien, Rudi, and Bach (2018).

In conclusion, the effect of additional regularity for $\theta^\star$ as Assumption 3.6, has two beneficial effects: (a) on one side it allows to obtain faster rates as shown in the next corollary, (b) as mentioned before, somewhat surprisingly, it reduces the constants to universal, since it allows the bias to go to zero faster than the Dikin radius (indeed, the squared radius $\mathsf{r}_\lambda^2(\theta^\star)$ is always larger than λ/R^2, which is strictly larger than $\lambda^{1+2r}\|v\|^2$ if $r > 0$ and λ small enough). This is why we do not the get exponential constants imposed by Hazan et al. (2014).

Corollary 3.2 (Rates with source condition)**.** *Let $\delta \in (0, 1/2]$. Under Assumptions 3.1 to 3.5 and Assumption 3.6, whenever $n \geq N$ and $\lambda = (C_0/n)^{1/(2+2r)}$, then with probability at least $1 - 2\delta$,*

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \;\leq\; C_1 \; n^{-\frac{1+2r}{2+2r}} \log \frac{2}{\delta},$$

with $C_0 = 256\ (\mathsf{B}_1^\star/\mathsf{L})^2$, $C_1 = 8\ (256)^\gamma((\mathsf{B}_1^\star)^\gamma \mathsf{L}^{1-\gamma})^2, \gamma = \frac{1+2r}{2+2r}$ and with N defined in Eq. (3.48) and satisfying $N = O(\mathrm{poly}(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, R, \log(1/\delta)))$.

The corollary above, derived in Sec. 3.F , is obtained by minimizing in λ the r.h.s. side of Eq. (3.9) in Theorem 3.3, and considering that when $\theta^\star$ satisfies the source condition, then $\mathsf{Bias}_\lambda \leq \lambda^{1+2r}\mathsf{L}$, while the variance is still of the form $1/(\lambda n)$. When r is close to 0, the rate $1/\sqrt{n}$ is recovered. When instead the target function is more regular, implying $r = 1/2$, a rate of $n^{-2/3}$ is achieved. Two considerations are in order: (a) the obtained rate is the same as least-squares and minimax optimal (Caponnetto and De Vito, 2007; Steinwart, Hush, and Scovel, 2009; Blanchard and Mücke, 2018), (b) the fact that regularized ERM is adaptive to the regularity of the function up to $r = 1/2$ is a byproduct of Tikhonov regularization as already shown for the least-squares case by Gerfo et al. (2008). Using different regularization techniques may remove the limit $r = 1/2$.

3.5 Fast Rates with both Source and Capacity Conditions

In this section, we consider improved results with a finer control of the effective dimension df_λ (often called degrees of freedom), which, together with the source condition allows to achieve rates as fast as $1/n$:

$$\mathsf{df}_\lambda = \mathbb{E}\left[\|\mathbf{H}_\lambda(\theta^\star)^{-1/2} \nabla \ell_Z(\theta^\star)\|^2 \right],$$

As mentioned earlier this definition of df_λ corresponds to the usual asymptotic term in M-estimation. Moreover, in the case of least-squares, it corresponds to the standard notion of effective dimension $\mathsf{df}_\lambda = \mathrm{Tr}(\mathbf{C}\mathbf{C}_\lambda^{-1})$ (Caponnetto and De Vito, 2007; Blanchard and Mücke,

2018). Note that by definition, we always have $\mathsf{df}_\lambda \leqslant \mathsf{B}_1^{\star 2}/\lambda$, but we can have in general a much finer control. For example, for least-squares, $\mathsf{df}_\lambda = O(\lambda^{-1/\alpha})$ if the eigenvalues of the covariance operator $\mathbf{C}$ decay as $\lambda_j(\mathbf{C}) = O(j^{-\alpha})$, for $\alpha \geq 1$. Moreover note that since $\mathbf{C}$ is trace-class, by Asm. 3.3, the eigenvalues form a summable sequence and so $\mathbf{C}$ satisfies $\lambda_j(\mathbf{C}) = O(j^{-\alpha})$ with α always larger than 1.

Example 3.4 (Generalized linear models)**.** *For generalized linear models, an extra assumption makes the degrees of freedom particularly simple: if the probabilistic model is well-specified, that is, there exists $\theta^\star$ such that almost surely, $p(y|x) = p(y|x,\theta^\star) = \frac{\exp(\theta^\star \cdot \Phi(x,y))}{\int_{\mathcal{Y}} \exp(\theta^\star \cdot \Phi(x,y'))d\mu(y')}$, then from the usual Bartlett identities (Bartlett, 1953) relating the expected squared derivatives and Hessians, we have $\mathbb{E}\left[\nabla \ell_z(\theta^\star) \otimes \nabla \ell_z(\theta^\star)\right] = \mathbf{H}(\theta^\star)$, leading to $\mathsf{df}_\lambda = \operatorname{Tr}(\mathbf{H}_\lambda(\theta^\star)^{-1}\mathbf{H}(\theta^\star))$.*

As we have seen in the previous example there are interesting problems for which $\mathsf{df}_\lambda = \operatorname{Tr}(\mathbf{H}(\theta^\star) + \lambda I)^{-1}\mathbf{H}(\theta^\star))$. Since we have $\mathbf{H}(\theta^\star) \preceq \mathsf{B}_2^\star \mathbf{C}$, df_λ still enjoys a polynomial decay depending on the eigenvalue decay of $\mathbf{C}$ as observed for least-squares. In the finite-dimensional setting where $\mathcal{H}$ is of dimension d, note that in this case, df_λ is always bounded by d. Now we are ready to state our result in the most general form, proved in Sec. 3.D .4.

Theorem 3.4 (General bound)**.** *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Whenever*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \quad n \geq \triangle_2 \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{\mathsf{r}_\lambda(\theta^\star)^2} \log \frac{2}{\delta},$$

with $(\mathsf{Q}^\star)^2 = \mathsf{B}_1^{\star 2}/\mathsf{B}_2^\star$, then with probability at least $1 - 2\delta$, it holds

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq \mathsf{C}_{\text{bias}}\, \mathsf{Bias}_\lambda^2 + \mathsf{C}_{\text{var}} \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n} \log \frac{2}{\delta}, \tag{3.10}$$

where, $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}, \square_1 \leq 414$, $\triangle_1$, $\triangle_2 \leq 5184$ when $\mathsf{Bias}_\lambda \leq \mathsf{r}_\lambda(\theta^\star)/2$;
otherwise C_{bias}, C_{var}, $\square_1 \leq 256 e^{6R\|\theta^\star\|}$, $\triangle_1$, $\triangle_2 \leq 2304(1 + R\|\theta^\star\|)^2 e^{8R\|\theta^\star\|}$.

As shown in the theorem above, the variance term depends on df_λ/n, implying that, when df_λ has a better dependence in λ than $1/\lambda$, it is possible to achieve faster rates. We quantify this with the following assumption.

Assumption 3.7 (Capacity condition)**.** *There exists $\alpha > 0$ and $\mathsf{Q} \geq 0$ such that $\mathsf{df}_\lambda \leq \mathsf{Q}\lambda^{-1/\alpha}$.*

Assumption 3.7 is standard in the context of least-squares, (Caponnetto and De Vito, 2007) and in many interesting settings is implied by the eigenvalue decay order of $\mathbf{H}(\theta^\star)$, or $\mathbf{C}$ as discussed above. In the following corollary we quantify the effect of df_λ in the learning rates.

Corollary 3.3. *Let $\delta \in (0, 1/2]$. Under Assumptions 3.1 to 3.5, Assumption 3.6 and Assumption 3.7, when $n \geq N$ and $\lambda = (C_0/n)^{\alpha/(1+\alpha(1+2r))}$, then with probability at least $1 - 2\delta$,*

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq C_1 n^{-\frac{\alpha(1+2r)}{1+\alpha(1+2r)}} \log \frac{2}{\delta},$$

with $C_0 = 256(\mathsf{Q}/\mathsf{L})^2$, $C_1 = 8(256)^\gamma\, (\mathsf{Q}^\gamma\, \mathsf{L}^{1-\gamma})^2$, $\gamma = \frac{\alpha(1+2r)}{1+\alpha(1+2r)}$ and N defined in Eq. (3.48) and satisfying $N = O(\operatorname{poly}(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, \mathsf{Q}, R, \log(1/\delta)))$.

The result above is derived in Cor. 3.4 in Sec. 3.F and is obtained by bounding Bias_λ with $\lambda^{1+2r}\mathsf{L}$ due to the source condition, and df_λ with $\lambda^{-1/\alpha}$ due to the capacity condition and then optimizing the r.h.s. of Eq. (3.10) in λ. Note that (a) the learning rate under the considered assumptions is the same as least-squares and minimax optimal (Caponnetto and De Vito, 2007), and (b) when $\alpha = 1$ the same rate of Cor. 3.2 is achieved, which can be as fast as $n^{-2/3}$, otherwise, when $\alpha \gg 1$, we achieve a learning rate in the order of $1/n$, for $\lambda = n^{-1/(1+2r)}$.

3.6 Sketch of the proof

In this section we will use the notation $\|v\|_{\mathbf{A}} := \|\mathbf{A}^{1/2}v\|$, with $v \in \mathcal{H}$ and $\mathbf{A}$ a bounded positive semi-definite operator on $\mathcal{H}$. Here we prove that the excess risk decomposes using the bias term Bias_λ defined in Eq. (3.8) and a variance term V_λ, where V_λ is defined as

$$V_\lambda := \|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star_\lambda)}, \text{ with } \widehat{L}_\lambda(\cdot) = \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(\cdot) + \frac{\lambda}{2}\|\cdot\|^2,$$

which in turn is a random variable that concentrate in high probability to $\sqrt{\mathsf{df}_\lambda/n}$.

Required tools. To proceed with the proof we need two main tools. The first is a result on the equivalence of norms of the empirical Hessian $\widehat{\mathbf{H}}_\lambda(\theta) = \nabla^2\widehat{L}_\lambda(\theta)$ w.r.t. the true Hessian $\mathbf{H}_\lambda(\theta) = \nabla^2 L_\lambda(\theta)$ for $\lambda > 0$ and $\theta \in \mathcal{H}$. The result is proven in Lemma 3.6 of Sec. 3.D .3, using Bernstein inequalities for Hermitian operators (Tropp, 2012), and essentially states that for $\delta \in (0,1]$, whenever $n \geq \frac{24\mathsf{B}_2(\theta)}{\lambda}\log\frac{8\mathsf{B}_2(\theta)}{\lambda\delta}$, then with probability $1-\delta$, it holds

$$\|\cdot\|_{\mathbf{H}_\lambda(\theta)} \leq 2\|\cdot\|_{\widehat{\mathbf{H}}_\lambda(\theta)}, \qquad \|\cdot\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)} \leq 2\|\cdot\|_{\mathbf{H}_\lambda^{-1}(\theta)}. \tag{3.11}$$

The second result is about localization properties induced by generalized self-concordance on the risk. We express the result with respect to a generic probability μ (we will use it with $\mu = \rho$ and $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$). Let μ be a probability distribution with support contained in the support of ρ. Denote by $L_\mu(\theta)$ the risk $L_\mu(\theta) = \mathbb{E}_{z\sim\mu}[\ell_z(\theta)]$ and by $L_{\mu,\lambda}(\theta) = L_\mu(\theta) + \frac{\lambda}{2}\|\theta\|^2$ (then $L_{\mu,\lambda} = L_\lambda$ when $\mu = \rho$, or $\widehat{L}_\lambda$ when $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$).

Proposition 3.1. *Under Assumptions 3.2 to 3.4, the following holds: (a) $L_{\mu,\lambda}(\theta), \nabla L_{\mu,\lambda}(\theta), \mathbf{H}_{\mu,\lambda}(\theta)$ are defined for all $\theta \in \mathcal{H}, \lambda \geq 0$, (b) for all $\lambda > 0$, there exists a unique $\theta^\star_{\mu,\lambda} \in \mathcal{H}$ minimizing $L_{\mu,\lambda}$ over $\mathcal{H}$, and (c) for all $\lambda > 0$ and $\theta \in \mathcal{H}$,*

$$\mathbf{H}_{\mu,\lambda}(\theta) \preceq e^{\mathsf{t}(\theta-\theta^\star_{\mu,\lambda})}\mathbf{H}_{\mu,\lambda}(\theta^\star_{\mu,\lambda}), \tag{3.12}$$

$$L_{\mu,\lambda}(\theta) - L_{\mu,\lambda}(\theta^\star_{\mu,\lambda}) \leq \psi(\mathsf{t}(\theta-\theta^\star_{\mu,\lambda}))\|\theta-\theta^\star_{\mu,\lambda}\|^2_{\mathbf{H}_{\mu,\lambda}(\theta^\star_{\mu,\lambda})}, \tag{3.13}$$

$$\underline{\phi}(\mathsf{t}(\theta-\theta^\star_{\mu,\lambda}))\|\theta-\theta^\star_{\mu,\lambda}\|_{\mathbf{H}_{\mu,\lambda}(\theta)} \leq \|\nabla L_{\mu,\lambda}(\theta)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta)}, \tag{3.14}$$

(d) Eqs. (3.12) and (3.13) hold also for $\lambda = 0$, provided that $\theta^\star_{\mu,0}$ exists. Here. $\underline{\phi}(t) = (1-e^{-t})/t$ and $\psi(t) = (e^t - t - 1)/t^2$.

The result above is proved in Sec. 3.B .1 and is essentially an extension of results by Bach (2010) applied to $L_{\mu,\lambda}$ under Assumptions 3.2 to 3.4.

Sketch of the proof. Now we are ready to decompose the excess risk using our bias and variance terms. In particular we will sketch the decomposition without studying the terms that lead to constants terms. For the complete proof of the decomposition see Theorem 3.7 in Sec. 3.D .1. Since $\theta^\star$ exists by Assumption 3.5, using Eq. (3.13), applied with $\mu = \rho$ and $\lambda = 0$, we have $L(\theta) - L(\theta^\star) \leq \psi(\mathsf{t}(\theta-\theta^\star))\|\theta-\theta^\star\|^2_{\mathbf{H}(\theta^\star)}$ for any $\theta \in \mathcal{H}$. By setting $\theta = \widehat{\theta}^\star_\lambda$, we obtain

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \psi(\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star))\|\widehat{\theta}^\star_\lambda - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}.$$

The term $\psi(\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star))$ will become a constant. For the sake of simplicity, in this sketch of proof we will not deal with it nor with other terms of the form $\mathsf{t}(\cdot)$ leading to constants. On the other hand, the term $\|\widehat{\theta}^\star_\lambda - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}$ will yield our bias and variance terms. Using the fact that $\mathbf{H}(\theta^\star) \preceq \mathbf{H}(\theta^\star) + \lambda I =: \mathbf{H}_\lambda(\theta^\star)$, by adding and subtracting $\theta^\star_\lambda$, we have

$$\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}(\theta^\star)} \le \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \le \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)},$$

so

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \;\le\; \text{const.} \;\times\; (\|\theta^\star_\lambda - \theta^\star\|^2_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)})^2.$$

By applying Eq. (3.12) with $\mu = \rho$ and $\theta = \theta^\star$, we have $\mathbf{H}_\lambda(\theta^\star) \preceq e^{\mathsf{t}_\lambda}\mathbf{H}_\lambda(\theta^\star_\lambda)$ and so we further bound $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)}$ with $e^{\mathsf{t}_\lambda/2}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$ obtaining

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \;\le\; \text{const.} \;\times\; (\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + e^{\mathsf{t}_\lambda/2}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)})^2.$$

The term $\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)}$ will lead to the *bias terms*, while the term $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$ will lead to the *variance term*.

Bounding the bias terms. Recall the definition of bias $\mathsf{Bias}_\lambda = \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}$ and of the constant $\mathsf{t}_\lambda := \mathsf{t}(\theta^\star - \theta^\star_\lambda)$. We bound $\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)}$ by applying Eq. (3.14) with $\mu = \rho$ and $\theta = \theta^\star$

$$\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)} \;\le\; 1/\underline{\phi}(\mathsf{t}_\lambda)\; \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \;=\; 1/\underline{\phi}(\mathsf{t}_\lambda)\; \mathsf{Bias}_\lambda.$$

Bounding the variance terms. To bound the term $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$, we assume n large enough to apply Eq. (3.11) in high probability. Thus, we obtain

$$\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} \le 2\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)}.$$

Applying Eq. (3.14) with $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\theta = \widehat{\theta}^\star_\lambda$, since $L_{\mu,\lambda} = \widehat{L}_\lambda$ for the given choice of μ,

$$\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)} \;\le\; \|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)} \;/\; \underline{\phi}(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda)),$$

and applying Eq. (3.11) in high probability again, we obtain

$$\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)} \le 2\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}.$$

Bias-variance decomposition. A technical part of the proof relates $\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}$ with $\|\nabla \widehat{L}_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} =: V_\lambda$, by many applications of Prop. 3.1. Here we assume it is done, obtaining

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \;\le\; \text{const.} \;\times\; (\mathsf{Bias}_\lambda^2 + V_\lambda^2).$$

From V_λ to $\sqrt{\mathsf{df}_\lambda/n}$. By construction, $\nabla \widehat{L}_\lambda(\theta^\star_\lambda) = \frac{1}{n}\sum_{i=1}^n \zeta_i$, with $\zeta_i := \nabla \ell_{z_i}(\theta^\star_\lambda) + \lambda\theta^\star_\lambda$. Moreover since the z_i's are i.i.d. samples from ρ, $\mathbb{E}[\zeta_i] = \nabla L_\lambda(\theta^\star_\lambda)$. Finally since $\theta^\star_\lambda$ is the minimizer of L_λ, $\nabla L_\lambda(\theta^\star_\lambda) = 0$. Thus $\nabla \widehat{L}_\lambda(\theta^\star_\lambda)$ is the average of n i.i.d. zero-mean random vectors, and so the variance of V_λ is exactly

$$\mathbb{E}\left[V_\lambda^2\right] = \frac{1}{n}\mathbb{E}\left[\|\mathbf{H}_\lambda^{-1/2}(\theta^\star)\nabla \ell_Z(\theta^\star)\|^2\right] = \frac{\mathsf{df}_\lambda}{n}.$$

Finally, by using Bernstein inequality for random vectors (e.g., Yurinsky, 1995, Thm. 3.3.4), we bound V_λ roughly with $\sqrt{\mathsf{df}_\lambda \log(2/\delta)/n}$ in high probability.

3.7 Conclusion

In this paper we have presented non-asymptotic bounds with faster rates than $O(1/\sqrt{n})$, for regularized empirical risk minimization with self-concordant losses such as the logistic loss. It would be interesting to extend our work to algorithms used to minimize the empirical risk, in particular stochastic gradient descent or Newton's method.

Organization of the Appendix

3.A Setting, definitions, assumptions

3.B Preliminary results on self concordant losses

3.B .1 Basic results on self-concordance **(proof of proposition 3.1)**

3.B .2 Localization properties for $\mathfrak{t}_\lambda$ **(proof of Lemma 3.1)**

3.C Main result, simplified

3.C .1 Analytic decomposition of the risk

3.C .2 Concentration lemmas

3.C .3 Final result **(proof of Theorem 3.2)**

3.D Main result, refined analysis

3.D .1 Analytic decomposition of the risk

3.D .2 Analytic decomposition of terms related to the variance

3.D .3 Concentration lemmas

3.D .4 Final result **(proof of Theorems 3.3 and 3.4)**

3.E Explicit bounds for the simplified case **(proof of Cor. 3.1)**

3.F Explicit bounds for the refined case **(proof of Cors. 3.2 and 3.3)**

3.G Additional lemmas

3.G .1 Self-concordance and sufficient conditions to define L

3.G .2 Bernstein inequalities for operators

3.A Setting, definitions, assumptions

Let $\mathcal{Z}$ be a Polish space and Z a random variable on $\mathcal{Z}$ whith law ρ. Let $\mathcal{H}$ be a separable (non-necessarily finite) Hilbert space and let $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ be a loss function; we denote by $\ell_z(\cdot)$ the function $\ell(z, \cdot)$. Our goal is to solve

$$\inf_{\theta \in \mathcal{H}} L(\theta), \quad \text{with} \quad L(\theta) = \mathbb{E}\left[\ell_Z(\theta)\right].$$

Given $(z_i)_{i=1}^n$ we will consider the following estimator

$$\widehat{\theta}^\star_\lambda = \arg\min_{\theta \in \mathcal{H}} \widehat{L}_\lambda(\theta), \quad \text{with} \quad \widehat{L}_\lambda(\theta) := \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(\theta) + \frac{\lambda}{2}\|\theta\|^2.$$

The goal of this work is to give upper bounds in high probability to the so called *excess risk*

$$L(\widehat{\theta}^\star_\lambda) - \inf_{\theta \in \mathcal{H}} L(\theta).$$

In the rest of this introduction we will introduce the basic assumptions required to make $\widehat{\theta}^\star_\lambda$ and the *excess risk* well defined, and we will introduce basic objects that are needed for the proofs.

First we introduce some notation we will use in the rest of the appendix: let $\lambda \geq 0$, $\theta \in \mathcal{H}$ and $\mathbf{A}$ be a bounded positive semidefinite Hermitian operator on $\mathcal{H}$, we denote by $\mathbf{I}$, the identity operator and

$$\|f\|_{\mathbf{A}} := \|\mathbf{A}^{1/2} f\|, \tag{3.15}$$

$$\mathbf{A}_\lambda := \mathbf{A} + \lambda \mathbf{I}, \tag{3.16}$$

$$\ell_z^\lambda(\theta) := \ell_z(\theta) + \frac{\lambda}{2}\|\theta\|^2, \tag{3.17}$$

$$L_\lambda(\theta) := L(\theta) + \frac{\lambda}{2}\|\theta\|^2. \tag{3.18}$$

Now we recall the assumptions we require on the loss function $\ell, \rho, (z_i)_{1 \leq i \leq n}$.

Assumption 3.1 (i.i.d. data)**.** *The samples $(z_i)_{1 \leq i \leq n}$ are independently and identically distributed according to ρ.*

Assumption 3.8 (Generalized self-concordance)**.** *The mapping $z \mapsto \ell_z(\theta)$ is measurable for all $\theta \in \mathcal{H}$ and for any $z \in \mathcal{Z}$, the function ℓ_z is convex and three times differentiable. Moreover, there exists a set $\varphi(z) \subset \mathcal{H}$ such that it holds:*

$$\forall \theta \in \mathcal{H},\ \forall h, k \in \mathcal{H},\ \left|\nabla^3 \ell_z(\theta)[k, h, h]\right| \leq \sup_{g \in \varphi(z)} |k \cdot g|\ \nabla^2 \ell_z(\theta)[h, h].$$

Assumption 3.3 (Boundedness)**.** *There exists $R \geq 0$ such that $\sup_{g \in \varphi(Z)} \|g\| \leq R$ almost surely.*

Assumption 3.4 (Definition in 0)**.** *$|\ell_Z(0)|$, $\|\nabla \ell_Z(0)\|$ and $\mathrm{Tr}(\nabla^2 \ell_Z(0))$ are almost surely bounded.*

Introduce the following definitions.

Definition 3.2. *Let $\lambda > 0$, $\theta \in \mathcal{H}$. We introduce*

$$\mathsf{B}_1(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta)\|, \qquad \mathsf{B}_2(\theta) = \sup_{z \in \mathrm{supp}(\rho)} \mathrm{Tr}\left(\nabla^2 \ell_z(\theta)\right). \tag{3.19}$$

$$\mathbf{H}(\theta) = \mathbb{E}\left[\nabla^2 \ell_Z(\theta)\right], \qquad \mathbf{H}_\lambda(\theta) = \mathbf{H}(\theta) + \lambda \mathbf{I}. \tag{3.20}$$

$$\theta_\lambda^\star = \arg\min_{\theta \in \mathcal{H}} L_\lambda(\theta). \tag{3.21}$$

Proposition 3.2. *Under Assumptions 3.3, 3.4 and 3.8, $\mathsf{B}_1(\theta), \mathsf{B}_2(\theta), L(\theta), \nabla L(\theta), \mathbf{H}(\theta), \theta_\lambda^\star$ exist for any $\theta \in \mathcal{H}, \lambda > 0$. Moreover $\nabla L = \mathbb{E}\left[\nabla \ell_Z(\theta)\right]$, $\mathbf{H}(\theta) = \nabla^2 L(\theta)$ and $\mathbf{H}(\theta)$ is trace class.*

Proof. We start by proving, using the assumptions, that $\mathsf{B}_2, \mathsf{B}_1$ and $\theta \mapsto \sup_{z \in \mathrm{supp}(\rho)} |\ell_z(\theta)|$ are all locally bounded (see Lemmas 3.11 to 3.13). This allows us to show that $\ell_z(\theta)$, $\nabla \ell_z(\theta)$ and $\mathrm{Tr}(\nabla^2 \ell_z(\theta))$ are uniformly integrable on any ball of finite radius. The fact that $\theta_\lambda^\star$ exists is due to the strong convexity of the function L_λ. □

Proposition 3.3. *Under Assumptions 3.1, 3.4 and 3.8, when $\lambda > 0$, $\widehat{\theta}_\lambda^\star$ exists and is unique.*

Proof. By Assumption 3.1 we know that $z_1, \ldots, z_n$ are in the support of ρ. Thus, by Assumption 3.4, $\frac{1}{n}\sum_{i=1}^n \ell_{z_i}$ is finite valued in 0. Since $\frac{1}{n}\sum_{i=1}^n \ell_{z_i}$ is convex three times differentiable as a sum of such functions, it is real-valued on $\mathcal{H}$ and hence $\widehat{L}_\lambda$ is real-valued on $\mathcal{H}$; by strong convexity, $\widehat{\theta}_\lambda^\star$ exists and is unique. □

Recall that we also make the following regularity assumption.

Assumption 3.5 (Existence of a minimizer)**.** *There exists $\theta^\star \in \mathcal{H}$ such that $L(\theta^\star) = \inf_{\theta\in\mathcal{H}} L(\theta)$.*

Finally we conclude with the following definitions that will be used later.

Definition 3.3. *For $\theta \in \mathcal{H}$, denote by $\mathsf{t}(\theta)$ the function*

$$\mathsf{t}(\theta) = \sup_{z\in\mathrm{supp}(\rho)} \left(\sup_{g\in\varphi(z)} |\theta \cdot g| \right),$$

and define

$$\mathsf{Bias}_\lambda = \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}, \tag{3.22}$$

$$\widehat{\mathsf{Var}}_\lambda = \|\mathbf{H}_\lambda^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\|^2 \ \|\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta_\lambda^\star)}, \tag{3.23}$$

$$\mathsf{df}_\lambda = \mathbb{E}\left[\|\nabla \ell_Z(\theta^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star)}\right], \tag{3.24}$$

$$\mathsf{t}_\lambda = \mathsf{t}(\theta^\star - \theta_\lambda^\star), \tag{3.25}$$

$$\mathsf{r}_\lambda(\theta) \quad \textit{such that} \quad 1/\mathsf{r}_\lambda(\theta) = \sup_{z\in\mathrm{supp}(\rho)} \left(\sup_{g\in\varphi(z)} \|g\|_{\mathbf{H}_\lambda^{-1}(\theta)} \right). \tag{3.26}$$

3.B Preliminary results on self concordant losses

In this section, we show how our definition/assumption of self concordance (see Assumption 3.8) enables a fine control on the excess risk. In particular, we clearly relate the difference in function values to the quadratic approximation at the optimum as well as the renormalized gradient. We start by presenting a general bounds in Sec. 3.B .1 before applying them to the problem of localizing the optimum Sec. 3.B .2.

3.B .1 Basic results on self-concordance

In this section, as in the rest of the appendix, we are under the conditions of Assumption 3.8. **In this section only**, we give ourselves a probability measure μ on $\mathcal{Z}$. We will apply the results of this section to $\mu = \rho, \widehat{\rho}, \delta_z$, where $\widehat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and z is sampled from ρ.

First of all, let us introduce the following notation. For any probability measure μ on $\mathcal{Z}$ and any $\theta \in \mathcal{H}$, define

- $R^\mu = \sup_{z\in\mathrm{supp}(\mu)} \left(\sup_{g\in\varphi(z)} \|g\|\right)$,
- $\mathsf{t}^\mu(\theta) = \sup_{z\in\mathrm{supp}(\mu)} \left(\sup_{g\in\varphi(z)} |\theta \cdot g|\right)$.

In order to be able to define $L_\mu(\theta) = \mathbb{E}_\mu[\ell_z(\theta)]$ and to derive under the expectation, we assume that Assumptions 3.3 and 3.4 are satisfied for μ (replace ρ by μ in the assumption).

Since μ and ℓ satisfy Assumptions 3.3, 3.4 and 3.8, proposition 3.8 ensures that we can define $L_\mu(\theta) = \mathbb{E}_\mu[\ell_z(\theta)]$ and $L_{\mu,\lambda}(\theta) = L_\mu(\theta) + \frac{\lambda}{2}\|\theta\|^2$, as well as their respective Hessians $\mathbf{H}_\mu(\theta)$ and $\mathbf{H}_{\mu,\lambda}(\theta)$.

The following result is greatly inspired from results in (Bach, 2010) on generalized self concordant losses, and their refinement in (Ostrovskii and Bach, 2018). However, while Eqs. (3.27), (3.29)

and (3.30) appear more or less explicitly, Eq. (3.28) provides an easier way to deal with certain bounds afterwards and was not used in this form before.

Proposition 3.4 (using the self-concordance of ℓ)**.** *Let $\theta_0, \theta_1 \in \mathcal{H}$ and $\lambda \geq 0$. Assume that $(\ell_z)_z$ and μ satisfy Assumptions 3.3, 3.4 and 3.8. We have the following inequalities:*

- *Bounds on Hessians*

$$\mathbf{H}_{\mu,\lambda}(\theta_1) \preceq \exp\left(\mathfrak{t}^\mu(\theta_1 - \theta_0)\right) \mathbf{H}_{\mu,\lambda}(\theta_0). \tag{3.27}$$

- *Bounds on gradients (if $\lambda > 0$)*

$$\underline{\phi}\left(\mathfrak{t}^\mu(\theta_1 - \theta_0)\right) \ \|\theta_1 - \theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)} \leq \|\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)}, \tag{3.28}$$

$$\|\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)} \leq \overline{\phi}\left(\mathfrak{t}^\mu(\theta_1 - \theta_0)\right) \ \|\theta_1 - \theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)}, \tag{3.29}$$

  where $\overline{\phi}(t) = (e^t - 1)/t$ and $\underline{\phi}(t) = (1 - e^{-t})/t$.

- *Bounds on function values*

$$L_{\mu,\lambda}(\theta_1) - L_{\mu,\lambda}(\theta_0) - \nabla L_{\mu,\lambda}(\theta_0)(\theta_1 - \theta_0) \leq \psi\left(\mathfrak{t}^\mu(\theta_1 - \theta_0)\right) \|\theta_1 - \theta_0\|^2_{\mathbf{H}_{\mu,\lambda}(\theta_0)}, \tag{3.30}$$

  where $\psi(t) = (e^t - t - 1)/t^2$.

Proof. First of all, note that for any μ and λ, given $\theta \in \mathcal{H}$ and $k, h \in \mathcal{H}$,

$$\begin{aligned}
\left|\nabla^3 L_{\mu,\lambda}(\theta)[h,k,k]\right| &= \left|\mathbb{E}_{z\sim\mu}\left[\nabla^3 \ell_z^\lambda(\theta)[h,k,k]\right]\right| \\
&\leq \mathbb{E}_{z\sim\mu}\left[\left|\nabla^3 \ell_z(\theta)[h,k,k]\right|\right] \\
&\leq \mathbb{E}_{z\sim\mu}\left[\sup_{g\in\varphi(z)} |h\cdot g| \ \nabla^2 \ell_z(\theta)[k,k]\right] \\
&\leq \mathfrak{t}^\mu(h) \ \mathbb{E}_{z\sim\mu}\left[\nabla^2 \ell_z(\theta)[k,k]\right] = \mathfrak{t}^\mu(h) \ \nabla^2 L_\mu(\theta)[k,k].
\end{aligned}$$

This yields the following fundamental inequality :

$$\left|\nabla^3 L_{\mu,\lambda}(\theta)[h,k,k]\right| \leq \ \mathfrak{t}^\mu(h) \ \nabla^2 L_{\mu,\lambda}(\theta)[k,k]. \tag{3.31}$$

We now define, for any $t \in \mathbb{R}$, $\theta_t := \theta_0 + t(\theta_1 - \theta_0)$.

Point 1. For the first inequality, let $h \in \mathcal{H}$ be a fixed vector, and consider the function $\varphi : t \in \mathbb{R} \mapsto \nabla^2 L_{\mu,\lambda}(\theta_t)[h,h]$. Since $\varphi'(t) = \nabla^3 L_{\mu,\lambda}(\theta_t)[\theta_1 - \theta_0, h, h]$, using Eq. (3.31), we get that $\varphi'(t) \leq \mathfrak{t}^\mu(\theta_1 - \theta_0) \ \varphi(t)$. Using Lemma 3.10, we directly find that $\varphi(1) \leq \exp(\mathfrak{t}^\mu(\theta_1 - \theta_0))\varphi(0)$, which, rewriting the definition of φ, yields

$$\nabla^2 L_{\mu,\lambda}(\theta_1)[h,h] \leq \exp(\mathfrak{t}^\mu(\theta_1 - \theta_0))\nabla^2 L_{\mu,\lambda}(\theta_0)[h,h].$$

This being true for any direction h, we have (3.27).

Point 2. To prove Eq. (3.28), let us look at the quantity $(\theta_1 - \theta_0) \cdot (\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0))$. Since $\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0) = \int_0^1 \nabla^2 L_{\mu,\lambda}(\theta_t)(\theta_1 - \theta_0)dt$, we have

$$(\theta_1 - \theta_0) \cdot (\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0)) = \int_0^1 \nabla^2 L_{\mu,\lambda}(\theta_t)[\theta_1 - \theta_0, \theta_1 - \theta_0]dt.$$

Applying Eq. (3.27) to θ_0 and θ_t and the reverse, we find that

$$\forall t \in [0,1],\ e^{-t\mathfrak{t}^\mu(\theta_1-\theta_0)}\nabla^2 L_{\mu,\lambda}(\theta_0) \preceq \nabla^2 L_{\mu,\lambda}(\theta_t).$$

Hence, integrating the previous equation, we have

$$(\theta_1 - \theta_0) \cdot (\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0)) \geq \underline{\phi}\left(\mathfrak{t}^\mu(\theta_1 - \theta_0)\right) \|\theta_1 - \theta_0\|^2_{\mathbf{H}_{\mu,\lambda}(\theta_0)}.$$

Finally, bounding $(\theta_1-\theta_0)\cdot(\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0))$ by $\|\theta_1-\theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)} \|\nabla L_{\mu,\lambda}(\theta_1)-\nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)}$, and simplifying by $\|\theta_1 - \theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)}$, we obtain Eq. (3.28).

Point 3. To prove Eq. (3.29), first write

$$\begin{aligned}
\|\nabla L_{\mu,\lambda}(\theta_1) - \nabla L_{\mu,\lambda}(\theta_0)\|_{\mathbf{H}^{-1}_{\mu,\lambda}(\theta_0)} &= \| \int_0^1 \mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)(\theta_1 - \theta_0)dt\| \\
&= \| \int_0^1 \mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\ \mathbf{H}^{1/2}_{\mu,\lambda}(\theta_0)(\theta_1 - \theta_0)dt\| \\
&\leq \left(\int_0^1 \|\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\|\ dt \right)\ \|\theta_1 - \theta_0\|_{\mathbf{H}_{\mu,\lambda}(\theta_0)}.
\end{aligned}$$

Then apply Eq. (3.27) to have

$$\forall t \in [0,1],\ \mathbf{H}_{\mu,\lambda}(\theta_t) \preceq e^{t\mathfrak{t}^\mu(\theta_1-\theta_0)}\mathbf{H}_{\mu,\lambda}(\theta_0).$$

This implies

$$\forall t \in [0,1],\ \mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0) \preceq e^{t\mathfrak{t}^\mu(\theta_1-\theta_0)}\ I.$$

And hence in particular

$$\forall t \in [0,1],\ \|\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\| \leq e^{t\mathfrak{t}^\mu(\theta_1-\theta_0)}.$$

Finally, integrating this, we get

$$\int_0^1 \|\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\mathbf{H}_{\mu,\lambda}(\theta_t)\mathbf{H}^{-1/2}_{\mu,\lambda}(\theta_0)\|\ dt \leq \overline{\phi}\left(\mathfrak{t}^\mu(\theta_1 - \theta_0)\right).$$

Thus Eq. (3.29) is proved.

Point 4. To prove Eq. (3.30), define $\forall t \in \mathbb{R},\ \varphi(t) = L_{\mu,\lambda}(\theta_t) - L_{\mu,\theta}(\theta_0) - t\ \nabla L_{\mu,\lambda}(\theta_0)(\theta_1 - \theta_0)$ We have $\varphi''(t) = \|\theta_1 - \theta_0\|^2_{\mathbf{H}_{\mu,\lambda}(\theta_t)} \leq e^{t\ \mathfrak{t}^\mu(\theta_1-\theta_0)}\varphi''(0)$. Then using the fact that $\varphi(0), \varphi'(0) = 0$ and integrating this inequality two times, we get the result. □

Proof. **of proposition 3.1.** First note that since the support of μ is included in the support of ρ, Assumption 3.3 and Assumption 3.4 also hold for μ. Hence, since Assumptions 3.2 to 3.4 are satisfied, by proposition 3.8, $L_{\mu,\lambda}$, $\nabla L_{\mu,\lambda}$ and $\nabla^2 L_{\mu,\lambda}$ are well-defined.

Assuming the existence of a minimizer $\theta^\star_{\mu,\lambda}$ of $L_{\mu,\lambda}$, the reported equations are the same than those of proposition 3.4 when taking $\theta_1 = \theta$ and $\theta_0 = \theta^\star_{\mu,\lambda}$, with the fact that $\mathfrak{t}^\mu(v) \leq \mathfrak{t}(v)$ for any $v \in \mathcal{H}$ since the support of μ is a subset of the support of ρ, and $\nabla L_{\mu,\lambda}(\theta^\star_{\mu,\lambda}) = 0$. Note that since $L_{\mu,\lambda}$ is defined on $\mathcal{H}$, if $\lambda > 0$, then $\theta^\star_{\mu,\lambda}$ always exists and is unique by strong convexity. □

3.B .2 Localization properties for t_λ

The aim of this section is to localize the optima $\theta^\star_\lambda$ and $\widehat{\theta}^\star_\lambda$ using the re-normalized gradient. This type of result is inspired by Proposition 2 of (Bach, 2010) or Proposition 3.5 of (Ostrovskii and Bach, 2018). However, their proof is based on a slightly different result, namely Eq. (3.28), and its formulation is slightly different. Indeed, while the two propositions mentioned above concentrate on performing a quadratic approximation directly, we bound the term that could have been too large in that quadratic approximation.

Proposition 3.5 (localisation)**.** *Let $\theta \in \mathcal{H}$, then the following holds*

$$\|\nabla L_\lambda(\theta)\|_{\mathbf{H}_\lambda^{-1}(\theta)} \leq \frac{\mathsf{r}_\lambda(\theta)}{2} \quad \Longrightarrow \quad \mathsf{t}(\theta - \theta^\star_\lambda) = \mathsf{t}_\lambda \leq \log 2, \tag{3.32}$$

$$\|\nabla \widehat{L}_\lambda(\theta)\|_{\mathbf{H}_\lambda{}^{-1}(\theta)} \, \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\|^2 \leq \frac{\mathsf{r}_\lambda(\theta)}{2} \quad \Longrightarrow \; \mathsf{t}(\theta - \widehat{\theta}^\star_\lambda) \leq \log 2. \tag{3.33}$$

Proof. To prove Eq. (3.32), we first write

$$\mathsf{t}(\theta - \theta^\star_\lambda) = \sup_{z \in \mathrm{supp}(\rho)} \sup_{g \in \varphi(z)} |(\theta - \theta^\star_\lambda) \cdot g| \leq \|\theta - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta)} \sup_{z \in \mathrm{supp}(\rho)} \sup_{g \in \varphi(z)} \|g\|_{\mathbf{H}_\lambda^{-1}(\theta)}.$$

Now we use Eq. (3.14) to bound $\|\theta - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta)}$, and putting things together, we get

$$\mathsf{t}(\theta - \theta^\star_\lambda)\underline{\phi}\left(\mathsf{t}(\theta - \theta^\star_\lambda)\right) \leq \frac{\|\nabla L_\lambda(\theta)\|_{\mathbf{H}_\lambda^{-1}(\theta)}}{\mathsf{r}_\lambda(\theta)}.$$

Using the fact that $t\underline{\phi}(t) = 1 - e^{-t}$ is an increasing function, we see that if $t\underline{\phi}(t) \leq 1/2$, then $t \leq \log 2$ hence the result.

To prove Eq. (3.33), we use the same reasoning. First, we bound

$$\mathsf{t}(\theta - \widehat{\theta}^\star_\lambda) = \sup_{z \in \mathrm{supp}(\rho)} \sup_{g \in \varphi(z)} \left|(\theta - \widehat{\theta}^\star_\lambda) \cdot g\right| \leq \|\theta - \widehat{\theta}^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta)} \, \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\| \sup_{z \in \mathrm{supp}(\rho)} \sup_{g \in \varphi(z)} \|g\|_{\mathbf{H}_\lambda^{-1}(\theta)}.$$

Now using Eq. (3.14) to the function $\widehat{L}_\lambda$, we get

$$\mathsf{t}(\theta - \widehat{\theta}^\star_\lambda)\underline{\phi}\left(\mathsf{t}^{\hat{\rho}}(\theta - \widehat{\theta}^\star_\lambda)\right) \leq \|\nabla \widehat{L}_\lambda(\theta)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)} \, \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\| \, \frac{1}{\mathsf{r}_\lambda(\theta)}.$$

Now using the fact that $\mathsf{t}^{\hat{\rho}}(\theta - \widehat{\theta}^\star_\lambda) \leq \mathsf{t}(\theta - \widehat{\theta}^\star_\lambda)$ and that $\underline{\phi}$ is a decreasing function, and that $\|\nabla \widehat{L}_\lambda(\theta)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta)} \leq \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\| \, \|\nabla \widehat{L}_\lambda(\theta)\|_{\mathbf{H}_\lambda{}^{-1}(\theta)}$, this yields

$$\mathsf{t}(\theta - \widehat{\theta}^\star_\lambda)\underline{\phi}\left(\mathsf{t}(\theta - \widehat{\theta}^\star_\lambda)\right) \leq \|\nabla \widehat{L}_\lambda(\theta)\|_{\mathbf{H}_\lambda{}^{-1}(\theta)} \, \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\mathbf{H}_\lambda{}^{1/2}(\theta)\|^2 \, \frac{1}{\mathsf{r}_\lambda(\theta)}.$$

We conclude using the same argument as before. □

3.C Main result, simplified

In this section, we perform a simplified analysis in the case where we assume nothing on Bias_λ more than just the fact that $\theta^\star$ exists. In this section we assume that ℓ_z and ρ satisfy Assumptions 3.3 to 3.5 and 3.8.

Definition 3.4 (Definition of $\overline{\mathsf{B}}_1$, $\overline{\mathsf{B}}_2$ and $\overline{\mathsf{df}}_\lambda$)**.** *Under assumptions Assumptions 3.3 to 3.5 and 3.8, the following quantities are well-defined and real-valued.*

$$\overline{\mathsf{B}}_1 = \sup_{\|\theta\| \le \|\theta^\star\|} \mathsf{B}_1(\theta) \quad \overline{\mathsf{B}}_2 = \sup_{\|\theta\| \le \|\theta^\star\|} \mathsf{B}_2(\theta), \quad \overline{\mathsf{df}}_\lambda = \mathbb{E}\left[\|\nabla \ell_z(\theta^\star_\lambda)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}\right].$$

Proposition 3.6. *The quantities in definition 3.4 are finite and moreover*

$$\overline{\mathsf{df}}_\lambda \le \frac{\overline{\mathsf{B}}_1^2}{\lambda}.$$

Proof. These are well defined thanks to Lemmas 3.11 and 3.12. □

Definition 3.5 (Constants)**.** *In this section, we will use the following constants.*

$$\mathsf{K}_{\mathrm{var}} = \frac{1+\psi(\log 2)}{\underline{\phi}(\log 2)^2} \le 4, \quad \triangle = 2\sqrt{2}\left(1+\frac{1}{2\sqrt{3}}\right) \le 4,$$

$$\mathsf{C}_{\mathrm{bias}} = 1 + \frac{\mathsf{K}_{\mathrm{var}}}{8} \le 2, \qquad \mathsf{C}_{\mathrm{var}} = 2\mathsf{K}_{\mathrm{var}}\triangle^2 \le 84.$$

3.C .1 Analytic results

Theorem 3.5 (Analytic decomposition)**.** *For any $\lambda > 0$ and $n \in \mathbb{N}$, if $\frac{R}{\sqrt{\lambda}}\widehat{\mathsf{Var}}_\lambda \le \frac{1}{2}$,*

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \le \mathsf{K}_{\mathrm{var}}\, \widehat{\mathsf{Var}}^2_\lambda + \lambda\|\theta^\star\|^2, \tag{3.34}$$

where $\mathsf{K}_{\mathrm{var}}$ is defined in definition 3.5.

Proof. First decompose the excess risk of $\widehat{\theta}^\star_\lambda$ in the following way:

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) = \underbrace{L_\lambda(\widehat{\theta}^\star_\lambda) - L_\lambda(\theta^\star_\lambda)}_{\text{variance}} + \underbrace{L(\theta^\star_\lambda) - L(\theta^\star)}_{\text{bias}} + \underbrace{\frac{\lambda}{2}\left(\|\theta^\star_\lambda\|^2 - \|\widehat{\theta}^\star_\lambda\|^2\right)}_{\text{mixed}}.$$

1) Variance term: For the variance term, use Eq. (3.13)

$$L_\lambda(\widehat{\theta}^\star_\lambda) - L_\lambda(\theta^\star_\lambda) \le \psi\left(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda)\right)\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|^2_{\mathbf{H}_\lambda(\theta^\star_\lambda)}.$$

2) Bias term: For the bias term, note that since $\|\theta^\star_\lambda\| \le \|\theta^\star\|$,

$$L(\theta^\star_\lambda) - L(\theta^\star) = L_\lambda(\theta^\star_\lambda) - L_\lambda(\theta^\star) + \frac{\lambda}{2}\|\theta^\star\|^2 - \frac{\lambda}{2}\|\theta^\star_\lambda\|^2 \le \frac{\lambda}{2}\|\theta^\star\|^2.$$

3) Mixed term: For the mixed term, since $\|\theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1}} \le \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\|\|\theta^\star_\lambda\| \le \lambda^{-1/2}\|\theta^\star_\lambda\| \le \lambda^{-1/2}\|\theta^\star\|$, we have

$$\begin{aligned}
\frac{\lambda}{2}\left(\|\theta^\star_\lambda\|^2 - \|\widehat{\theta}^\star_\lambda\|^2\right) &= \frac{\lambda}{2}\left(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\right)\cdot\left(\theta^\star_\lambda + \widehat{\theta}^\star_\lambda\right) \\
&\le \frac{\lambda}{2}\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}\ \left(\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1}} + 2\|\theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1}}\right) \\
&\le \frac{1}{2}\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|^2_{\mathbf{H}_\lambda(\theta^\star_\lambda)} + \sqrt{\lambda}\|\theta^\star\|\ \|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} \\
&\le \|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|^2_{\mathbf{H}_\lambda(\theta^\star_\lambda)} + \frac{\lambda}{2}\|\theta^\star\|^2.
\end{aligned}$$

where we get the last inequality by using $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$.

4) Putting things together

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq \left(1 + \psi\left(\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star)\right)\right) \|\theta_\lambda^\star - \widehat{\theta}_\lambda^\star\|^2_{\mathbf{H}_\lambda(\theta_\lambda^\star)} + \lambda\|\theta^\star\|^2.$$

By using Eq. (3.14) we have

$$\begin{aligned}\|\theta_\lambda^\star - \widehat{\theta}_\lambda^\star\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} &\leq \|\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\| \; \|\theta_\lambda^\star - \widehat{\theta}_\lambda^\star\|_{\widehat{\mathbf{H}}_\lambda(\theta_\lambda^\star)} \\ &\leq \frac{1}{\underline{\phi}\left(\mathsf{t}^{\widehat{\rho}}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star)\right)} \; \|\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\| \; \|\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta_\lambda^\star)}.\end{aligned}$$

Note that by multiplying and dividing for $\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)$,

$$\begin{aligned}\|\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta_\lambda^\star)} &= \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| = \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\mathbf{H}_\lambda{}^{-1/2}(\theta_\lambda^\star)\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \\ &\leq \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\mathbf{H}_\lambda{}^{-1/2}(\theta_\lambda^\star)\|\|\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \\ &= \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\mathbf{H}_\lambda{}^{-1/2}(\theta_\lambda^\star)\|\|\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}.\end{aligned}$$

Then,

$$\begin{aligned}\|\theta_\lambda^\star - \widehat{\theta}_\lambda^\star\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} &\leq \frac{1}{\underline{\phi}\left(\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star)\right)} \; \|\mathbf{H}_\lambda{}^{1/2}(\theta_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta_\lambda^\star)\|^2 \; \|\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \\ &= \frac{1}{\underline{\phi}\left(\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star)\right)} \; \widehat{\mathsf{Var}}_\lambda.\end{aligned}$$

Now we know that using Eq. (3.33), if $\widehat{\mathsf{Var}}_\lambda \leq \frac{\mathsf{r}_\lambda(\theta_\lambda^\star)}{2}$, then $\mathsf{t}(\theta_\lambda^\star - \widehat{\theta}_\lambda^\star) \leq \log 2$, which yields the following bound:

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq \frac{(1 + \psi(\log 2))}{\underline{\phi}(\log 2)^2}\widehat{\mathsf{Var}}_\lambda + \lambda\|\theta^\star\|^2.$$

Finally, we can bound $\frac{1}{\mathsf{r}_\lambda(\theta_\lambda^\star)} \leq \frac{R}{\lambda^{1/2}}$ to have the final form of the proposition.

□

3.C .2 Probabilistic results

Lemma 3.2 (bounding $\|\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)}$)**.** *Let $n \in \mathbb{N}$, $\lambda > 0$ and $\delta \in (0, 1]$. For $k \geq 1$, if*

$$n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{2}{\delta}, n \geq k^2 2\log\frac{2}{\delta},$$

then with probability at least $1 - \delta$,

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leq \triangle/2\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2) \;\log\frac{2}{\delta}}{n}} + \frac{2}{k}\sqrt{\lambda}\|\theta^\star\|$$

where $\triangle$ is defined in definition 3.5.

Proof. **1)** First use Bernstein inequality for random vectors (e.g. Thm. 3.3.4 of Yurinsky, 1995): for any $n \in \mathbb{N}$ and $\delta \in (0,1]$, with probability at least $1-\delta$, we have

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leq \frac{2M\log\frac{2}{\delta}}{n} + \sigma\sqrt{\frac{2\log\frac{2}{\delta}}{n}},$$

where $M = \sup_{z\in\mathrm{supp}(\rho)}\|\nabla\ell_z^\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}$ and $\sigma = \mathbb{E}\left[\|\nabla\ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}\right]^{1/2}$.

2) Using the fact that $\nabla\ell_z^\lambda(\theta_\lambda^\star) = \nabla\ell_z(\theta_\lambda^\star) + \lambda\theta_\lambda^\star$, we bound M as follows:

$$M = \sup_{z\in\mathrm{supp}(\rho)}\|\nabla\ell_z^\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} \leq \sup_{z\in\mathrm{supp}(\rho)}\|\nabla\ell_z(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda(\theta_\lambda^\star)} + \lambda\|\theta_\lambda^\star\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leq \frac{\overline{\mathsf{B}}_1}{\sqrt{\lambda}} + \sqrt{\lambda}\|\theta^\star\|,$$

where in the last inequality, we use the fact that $\|\theta_\lambda^\star\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leq \frac{1}{\sqrt{\lambda}}\|\theta_\lambda^\star\| \leq \frac{1}{\sqrt{\lambda}}\|\theta^\star\|$. Similarly, we bound σ

$$\sigma \leq \mathbb{E}\left[\|\nabla\ell_z(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)}\right]^{1/2} + \lambda\|\theta_\lambda^\star\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leq \overline{\mathsf{df}}_\lambda^{1/2} + \sqrt{\lambda}\|\theta^\star\|.$$

3) Injecting these bounds in the concentration inequality,

$$\begin{aligned}\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leq{}& \sqrt{\frac{2\overline{\mathsf{B}}_2\ \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2(\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}} + \sqrt{\frac{2\overline{\mathsf{df}}_\lambda\ \log\frac{2}{\delta}}{n}} \\ &+ \sqrt{\lambda}\|\theta^\star\|\left(\frac{2\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}\right),\end{aligned}$$

where we have decomposed $\frac{2\overline{\mathsf{B}}_1^2\log\frac{2}{\delta}}{\sqrt{\lambda}n} = \sqrt{\frac{2\overline{\mathsf{B}}_2\ \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2(\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}}$ for the first term. Reordering the terms, this yields

$$\begin{aligned}\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leq{}& \left(1 + \sqrt{\frac{2\overline{\mathsf{B}}_2\ \log\frac{2}{\delta}}{\lambda n}}\right)\sqrt{\frac{2\overline{\mathsf{df}}_\lambda \vee \left(\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2\right)\ \log\frac{2}{\delta}}{n}} \\ &+ \sqrt{\lambda}\|\theta^\star\|\left(\frac{2\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}\right).\end{aligned}$$

4) Now assuming that

$$n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{2}{\delta}, n \geq k^2 2\log\frac{2}{\delta},$$

this yields

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2}\nabla\widehat{L}_\lambda(\theta_\lambda^\star)\| \leq \left(1 + \frac{1}{2\sqrt{3}}\right)\sqrt{\frac{2\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}} + \frac{2}{k}\sqrt{\lambda}\|\theta^\star\|.$$

□

Combining the two previous lemmas, we get:

Lemma 3.3 (Bounding $\widehat{\mathsf{Var}}_\lambda$)**.** *Let $n \in \mathbb{N}$ and $0 < \lambda \leq \overline{\mathsf{B}}_2$. Let $\delta \in (0,1]$. If for $k \geq 1$*

$$n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta},\ n \geq 2k^2\log\frac{2}{\delta},$$

then with probability at least $1-2\delta$,

$$\widehat{\mathsf{Var}}_\lambda \leq \triangle\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}} + \frac{4}{k}\sqrt{\lambda}\|\theta^\star\|,$$

where $\triangle$ is a constant defined in definition 3.5.

Proof. Recall that $\widehat{\mathsf{Var}}_\lambda = \|\mathbf{H}_\lambda{}^{1/2}(\theta^\star_\lambda)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\|^2\ \|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)}$. Using Lemma 3.6, under the conditions of this lemma, we have $\|\mathbf{H}_\lambda{}^{1/2}(\theta^\star_\lambda)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\|^2 \leq 2$. Combining this with the bound for $\|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)}$ obtained in Lemma 3.2, we get the result (the probability $1-2\delta$ comes from the fact that we perform a union bound).

□

3.C .3 Final result

Theorem 3.6 (General bound, simplified setting)**.** *Let $n \in \mathbb{N}$ and $0 < \lambda \leq \overline{\mathsf{B}}_2$. Let $\delta \in (0,1]$. If*

$$n \geq 512\left(\|\theta^\star\|^2R^2 \vee 1\right)\log\frac{2}{\delta},\ n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta},\ n \geq 16\triangle^2R^2\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)}{\lambda}\ \log\frac{2}{\delta},$$

then with probability at least $1-2\delta$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \mathsf{C}_{\text{var}}\ \frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)}{n}\ \log\frac{2}{\delta} + \mathsf{C}_{\text{bias}}\lambda\|\theta^\star\|^2,$$

where $\triangle, \mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ are defined in definition 3.5.

Proof. **1)** Recall the analytical decomposition in Theorem 3.5. For any $\lambda > 0$ and $n \in \mathbb{N}$, if $\frac{R}{\sqrt{\lambda}}\widehat{\mathsf{Var}}_\lambda \leq \frac{1}{2}$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \mathsf{K}_{\text{var}}\ \widehat{\mathsf{Var}}_\lambda^2 + \lambda\|\theta^\star\|^2,$$

where K_{var} is defined in definition 3.5.

2) Now apply Lemma 3.3 for a given $k \geq 1$. If

$$n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta},\ n \geq 2k^2\log\frac{2}{\delta},$$

then with probability at least $1-2\delta$,

$$\widehat{\mathsf{Var}}_\lambda \leq \triangle\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n}} + \frac{4}{k}\sqrt{\lambda}\|\theta^\star\|,$$

where $\triangle$ is a constant defined in definition 3.5.

In order to satisfy the condition to have the analytical decomposition, namely $\frac{R}{\sqrt{\lambda}}\widehat{\mathsf{Var}}_\lambda \le \frac{1}{2}$, it is therefore sufficient to have

$$\triangle\, R\sqrt{\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{\lambda n}} \le \frac{1}{4}, \qquad \frac{4}{k}R\|\theta^\star\| \le \frac{1}{4}.$$

3) Thus, if we choose $k = 16(R\|\theta^\star\| \vee 1)$, we have both $k \ge 1$ and the second condition in the previous equation. Moreover, the condition $n \ge 2k^2 \log\frac{2}{\delta}$ becomes $n \ge 512(R^2\|\theta^\star\|^2 \vee 1)\log\frac{2}{\delta}$. Hence, under the conditions of this theorem, we can apply the analytical decomposition :

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \le \mathsf{K}_{\mathrm{var}}\,\widehat{\mathsf{Var}}_\lambda^2 + \lambda\|\theta^\star\|^2 \le 2\mathsf{K}_{\mathrm{var}}\triangle^2\,\frac{\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2)\ \log\frac{2}{\delta}}{n} + \left(1 + \mathsf{K}_{\mathrm{var}}\frac{32}{k^2}\right)\lambda\|\theta^\star\|^2.$$

In the last inequality, we have used $(a+b)^2 \le 2a^2 + 2b^2$ to separate the terms coming from $\widehat{\mathsf{Var}}_\lambda^2$.

Finally, using the fact that $k \ge 16$ and hence that $\frac{32}{k^2} \le \frac{1}{8}$, we get the constants in the theorem. □

Proof. **of Theorem 3.2** Since $\forall\lambda > 0$, $\mathsf{df}_\lambda \le \frac{\overline{\mathsf{B}}_1^2}{\lambda}$, and since $\lambda \le \overline{\mathsf{B}}_2$, $\mathsf{df}_\lambda \vee \overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2 \le \frac{\overline{\mathsf{B}}_1^2}{\lambda}$. From definition 3.5, we get that $\triangle \le 4$, $\mathsf{C}_{\mathrm{bias}} \le 2$, $\mathsf{C}_{\mathrm{var}} \le 84$. Thus, we can use these bounds in Theorem 3.6 to obtain the result.

□

3.D Main result, refined analysis

In subsection Sec. 3.D .1 we split the excess risk in terms of bias and variance, that will be controlled in Sec. 3.D .3, the final result is Theorem 3.4 in Sec. 3.D .4, while in Sec. 3.F a version with explicit dependence in λ, n is reported.

Constants First, we introduce three constants that will be crucial for the final bound.

Definition 3.6.

$$\mathsf{B}_1^\star = \mathsf{B}_1(\theta^\star), \qquad \mathsf{B}_2^\star = \mathsf{B}_2(\theta^\star), \qquad \mathsf{Q}^* = \mathsf{B}_1^\star/\sqrt{\mathsf{B}_2^\star}.$$

In the following sections, we also will use the following functions of t_λ and $\widetilde{\mathsf{t}}_\lambda$ which we will treat as constants (see proposition 3.7).

Definition 3.7.

$$K_{\mathrm{bias}}(\mathsf{t}_\lambda) = 2\frac{\psi(\mathsf{t}_\lambda + \log 2)}{\underline{\phi}(\mathsf{t}_\lambda)^2} \le 2e^{3\mathsf{t}_\lambda}, \qquad K_{\mathrm{var}}(\mathsf{t}_\lambda) = 2\frac{\psi(\mathsf{t}_\lambda + \log 2)e^{\mathsf{t}_\lambda}}{\underline{\phi}(\log 2)^2} \le 8e^{2\mathsf{t}_\lambda},$$

$$\square_1(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}, \qquad \square_2(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}\left(1 + e^{\mathsf{t}_\lambda}\right) \le 2e^{3\mathsf{t}_\lambda/2}$$

$$\mathsf{C}_{\mathrm{bias}} = \psi(\mathsf{t}_\lambda + \log 2)\left(\frac{2}{\underline{\phi}(\mathsf{t}_\lambda)} + \frac{e^{\mathsf{t}_\lambda}}{\underline{\phi}(\log 2)^2}\right) \le 6e^{2\mathsf{t}_\lambda}, \qquad \mathsf{C}_{\mathrm{var}} = \frac{64\psi(\mathsf{t}_\lambda + \log 2)e^{2\mathsf{t}_\lambda}}{\underline{\phi}(\log 2)^2} \le 256e^{3\mathsf{t}_\lambda}$$

$$\triangle_1 = 576\square_1^2\square_2^2(1/2 \vee \widetilde{\mathsf{t}}_\lambda)^2 \le 2304e^{4\mathsf{t}_\lambda}(\widetilde{\mathsf{t}}_\lambda \vee 1/2)^2, \qquad \triangle_2 = 256\square_1^4 \le 256e^{2\mathsf{t}_\lambda}.$$

Note that theses functions are all increasing in t_λ and $\widetilde{\mathsf{t}}_\lambda$, and are lower bounded by strictly positive constants.

For the second bounds, we use the fact that $\psi(t) \le \frac{e^t}{2}$ and $1/\underline{\phi}(t) \le e^t$ to bound all the quantities using only exponentials of t_λ.

A priori, these constants will depend on λ. However, we can always bound t_λ and $\widetilde{\mathsf{t}}_\lambda$ in the following way.

Lemma 3.4. *Recall the definitions of* $\mathsf{t}_\lambda := \mathsf{t}(\theta^\star_\lambda - \theta^\star)$ *and* $\widetilde{\mathsf{t}}_\lambda := \frac{\mathsf{Bias}_\lambda}{\mathsf{r}_\lambda(\theta^\star)}$. *We have the following cases.*

- *If* $\widetilde{\mathsf{t}}_\lambda \leq \frac{1}{2}$, *then* $\mathsf{t}_\lambda \leq \log 2$,
- *else,* $\widetilde{\mathsf{t}}_\lambda \leq R\|\theta^\star\|$ *and* $\mathsf{t}_\lambda \leq 2R\|\theta^\star\|$.

Proof. The first point is a direct application of Eq. (3.32). One can obtain the second by noting that $\mathsf{t}(\theta^\star_\lambda - \theta^\star) \leq R\|\theta^\star_\lambda - \theta^\star\|$. Since $\|\theta^\star_\lambda\| \leq \|\theta^\star\|$, we have the bound on t_λ. For the bound on $\widetilde{\mathsf{t}}_\lambda$, since $\mathsf{Bias}_\lambda \leq \sqrt{\lambda}\|\theta^\star\|$ and $\frac{1}{\mathsf{r}_\lambda(\theta^\star)} \leq \frac{R}{\sqrt{\lambda}}$, we have the wanted bound.

□

Hence, we can always bound the constants in definition 3.7 by constants independant of λ.

Proposition 3.7. *If* $\widetilde{\mathsf{t}}_\lambda \leq 1/2$, *then* $\mathsf{t}_\lambda \leq \log 2$ *and*

$$\begin{aligned} K_{\text{bias}}(\mathsf{t}_\lambda) &\leq 4, & K_{\text{var}}(\mathsf{t}_\lambda) &\leq 7, & \square_1(\mathsf{t}_\lambda) &\leq 2, & \square_2(\mathsf{t}_\lambda) &\leq 5 \\ \triangle_1(\mathsf{t}_\lambda, \widetilde{\mathsf{t}}_\lambda) &\leq 5184, & \triangle_2(\mathsf{t}_\lambda) &\leq 1024, & \mathsf{C}_{\text{bias}} &\leq 6, & \mathsf{C}_{\text{var}} &\leq 414. \end{aligned}$$

Else,

$$\begin{aligned} K_{\text{bias}}(\mathsf{t}_\lambda) &\leq 2e^{6R\|\theta^\star\|}, & K_{\text{var}}(\mathsf{t}_\lambda) &\leq 8e^{4R\|\theta^\star\|}, & \square_1(\mathsf{t}_\lambda) &\leq e^{R\|\theta^\star\|}, \\ \square_2(\mathsf{t}_\lambda) &\leq 2e^{3R\|\theta^\star\|}, & \triangle_1(\mathsf{t}_\lambda, \widetilde{\mathsf{t}}_\lambda) &\leq 2304(R\|\theta^\star\|)^2 e^{8R\|\theta^\star\|}, & \triangle_2(\mathsf{t}_\lambda) &\leq 256e^{4R\|\theta^\star\|}, \\ \mathsf{C}_{\text{bias}} &\leq 6e^{4R\|\theta^\star\|}, & \mathsf{C}_{\text{var}} &\leq 256e^{6R\|\theta^\star\|}. \end{aligned}$$

Proof. For the first bound, we use the fact that $\mathsf{t}_\lambda \leq \log 2$ and plug that in the expressions above as these functions are increasing in t_λ. We compute them numerically from the definition.

For the second set of bounds, we simply inject the bounds for t_λ and $\widetilde{\mathsf{t}}_\lambda$ in the second bounds of definition 3.7. □

3.D .1 Analytic decomposition of the risk

In this section, we make use of self-concordance to control certain quantities required to control the variance, with respect to our main quantities Bias_λ, r_λ and df_λ. The excess risk has been already decomposed in Sec. 3.6.

Theorem 3.7 (Analytic decomposition)**.** *Let* $\lambda > 0$ *and* K_{bias} *and* K_{var} *be the increasing functions of* t_λ *described in Eq.* (3.37). *When* $\widehat{\mathsf{Var}}_\lambda \leq \mathsf{r}_\lambda(\theta^\star_\lambda)/2$, *then*

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \;\leq\; K_{\text{bias}}(\mathsf{t}_\lambda)\ \mathsf{Bias}_\lambda^2 \;+\; K_{\text{var}}(\mathsf{t}_\lambda)\ \widehat{\mathsf{Var}}_\lambda^2. \tag{3.35}$$

Moreover $K_{\text{bias}}(\mathsf{t}_\lambda), K_{\text{var}}(\mathsf{t}_\lambda) \leq 7$ *if* $\mathsf{Bias}_\lambda \leq \frac{1}{2}\mathsf{r}_\lambda(\theta^\star)$, *otherwise* $K_{\text{bias}}(\mathsf{t}_\lambda), K_{\text{var}}(\mathsf{t}_\lambda) \leq 8e^{6\|\theta^\star\|\ R}$ *(see proposition 3.7 in Sec. 3.D for more precise bounds).*

Proof. Since $\theta^\star$ exists by Assumption 3.5, using Eq. (3.13), applied with $\mu = \rho$ and $\lambda = 0$, we have $L(\theta) - L(\theta^\star) \leq \psi(\mathsf{t}(\theta - \theta^\star))\|\theta - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}$, for any $\theta \in \mathcal{H}$. By setting $\theta = \widehat{\theta}^\star_\lambda$, we obtain

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \psi(\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star))\|\widehat{\theta}^\star_\lambda - \theta^\star\|^2_{\mathbf{H}(\theta^\star)}.$$

Using the fact that $\mathbf{H}(\theta^\star) \preceq \mathbf{H}(\theta^\star) + \lambda I =: \mathbf{H}_\lambda(\theta^\star)$, by adding and subtracting $\theta^\star_\lambda$, we have

$$\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}(\theta^\star)} \leq \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \leq \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)},$$

and analogously since $\mathsf{t}(\cdot)$ is a (semi)norm, $\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star) \leq \mathsf{t}_\lambda + \mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star)$, so

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \;\leq\; \psi(\mathsf{t}_\lambda + \mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda))\; (\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + \|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)})^2.$$

By applying Eq. (3.12) with $\mu = \rho$ and $\theta = \theta^\star$, we have $\mathbf{H}_\lambda(\theta^\star) \preceq e^{\mathsf{t}_\lambda}\mathbf{H}_\lambda(\theta^\star_\lambda)$ and so

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \;\leq\; \psi(\mathsf{t}_\lambda + \mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda))\; (\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} + e^{\mathsf{t}_\lambda/2}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)})^2. \tag{3.36}$$

The terms t_λ and $\|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)}$ are related to the *bias terms*, while the terms $\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda)$ and $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)}$ are related to the *variance term*.

Bounding the bias terms. Recall the definition of the bias $\mathsf{Bias}_\lambda = \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}$. We bound $\mathsf{t}_\lambda = \mathsf{t}(\theta^\star_\lambda - \theta^\star)$, by Lemma 3.1 and the term $\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)}$ by applying Eq. (3.14) with $\mu = \rho$ and $\theta = \theta^\star$

$$\|\theta^\star - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star)} \;\leq\; 1/\underline{\phi}(\mathsf{t}_\lambda)\; \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \;=\; 1/\underline{\phi}(\mathsf{t}_\lambda)\; \mathsf{Bias}_\lambda.$$

Bounding the variance terms. First we bound the term $\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} := \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda)\|$, by multiplying and dividing for $\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{-1/2}$, we have

$$\begin{aligned}\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} &= \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{-1/2}\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{1/2}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda)\| \\ &\leq \|\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)^{-1/2}\|\|\widehat{\theta}^\star_\lambda - \theta^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)}.\end{aligned}$$

Applying Eq. (3.14) with $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\theta = \widehat{\theta}^\star_\lambda$, since $L_{\mu,\lambda} = \widehat{L}_\lambda$ for the given choice of μ, we have

$$\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\widehat{\mathbf{H}}_\lambda(\theta^\star_\lambda)} \;\leq\; \|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)} \;/\; \underline{\phi}(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda))$$

and since $\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)} := \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|$, by multiplying and dividing by $\mathbf{H}_\lambda(\theta^\star_\lambda)$, we have:

$$\begin{aligned}\|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\| &= \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\| \\ &\leq \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta^\star_\lambda)\mathbf{H}_\lambda(\theta^\star_\lambda)^{1/2}\|\|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}.\end{aligned}$$

Then

$$\|\theta^\star_\lambda - \widehat{\theta}^\star_\lambda\|_{\mathbf{H}_\lambda(\theta^\star_\lambda)} \;\leq\; \frac{1}{\underline{\phi}(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda))}\; \|\mathbf{H}_\lambda^{1/2}(\theta^\star_\lambda)\mathbf{H}_\lambda{}^{-1/2}(\theta^\star_\lambda)\|^2\; \|\nabla \widehat{L}_\lambda(\theta^\star_\lambda)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\theta^\star_\lambda)} = \frac{\widehat{\mathsf{Var}}_\lambda}{\underline{\phi}(\mathsf{t}(\theta^\star_\lambda - \widehat{\theta}^\star_\lambda))}.$$

To conclude this part of the proof we need to bound $\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda)$. Since we require $\widehat{\mathsf{Var}}_\lambda/\mathsf{r}_\lambda(\theta^\star_\lambda) \leq 1/2$, by proposition 3.5 we have $\mathsf{t}(\widehat{\theta}^\star_\lambda - \theta^\star_\lambda) \leq \log 2$.

Gathering the terms. By gathering the results of the previous paragraphs

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \;\leq\; \psi(\mathsf{t}_\lambda + \log 2)\,\big(\; 1/\underline{\phi}(\mathsf{t}_\lambda)\;\mathsf{Bias}_\lambda \;+\; e^{\mathsf{t}_\lambda/2}/\underline{\phi}(\log 2)\;\widehat{\mathsf{Var}}_\lambda\;\big)^2$$

Using the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, we have the desired result, with

$$K_{\text{bias}}(\mathsf{t}_\lambda) = 2\psi(\mathsf{t}_\lambda + \log 2)/\underline{\phi}(\mathsf{t}_\lambda)^2,\;\; K_{\text{var}}(\mathsf{t}_\lambda) = 2\psi(\mathsf{t}_\lambda + \log 2)e^{\mathsf{t}_\lambda}/\underline{\phi}(\log 2)^2. \tag{3.37}$$

which are bounded in definition 3.7 and proposition 3.7 of Sec. 3.D . □

3.D .2 Analytic bounds for terms related to the variance

In this lemma, we aim to control the essential supremum and the variance of the random vector $\mathbf{H}_\lambda{}^{-1/2}(\theta^\star_\lambda)\nabla\ell^\lambda_z(\theta^\star_\lambda)$ relating it to quantities at $\theta^\star$. The results will be used to control the variance via Bernstein concentration inequalities, so we are going to control its essential supremum and its variance.

Lemma 3.5 (Control of $\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\nabla\ell^\lambda_z(\theta^\star_\lambda)$)**.** *For any $0 < \lambda \leq \mathsf{B}^\star_2$, we have*

1. *A bound on the essential supremum:*

$$\sup_{z\in\text{supp}(\rho)} \|\nabla\ell_z(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)} \leq \square_1\;\frac{\mathsf{B}^\star_1}{\sqrt{\lambda}} + 2\square_2\;\frac{\mathsf{B}^\star_2}{\lambda}\mathsf{Bias}_\lambda.$$

2. *A bound on the variance*

$$\mathbb{E}\left[\|\nabla\ell^\lambda_z(\theta^\star_\lambda)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)}\right]^{1/2} \leq \square_1\;\sqrt{\mathsf{df}_\lambda} + \sqrt{2}\square_2\;\sqrt{\frac{\mathsf{B}^\star_2}{\lambda}}\mathsf{Bias}_\lambda,$$

where $\square_1, \square_2$ are increasing functions of t_λ : $\square_1(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}$ $\square_2(\mathsf{t}_\lambda) = e^{\mathsf{t}_\lambda/2}\,(1 + e^{\mathsf{t}_\lambda})$.

Proof. Start by noting that if $\lambda \leq \mathsf{B}^\star_2$, then $\sup_{z\in\text{supp}(\rho)} \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\nabla^2\ell^\lambda_z(\theta^\star)^{1/2}\|^2 \leq 1 + \frac{\mathsf{B}^\star_2}{\lambda} \leq 2\frac{\mathsf{B}^\star_2}{\lambda}$. Moreover, note that for any vector $h \in \mathcal{H}$, multiplying and dividing by $\nabla^2\ell_z(\theta^\star)^{1/2}$,

$$\begin{aligned}
\|h\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} &:= \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\;h\| = \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\;\nabla^2\ell_z(\theta^\star)^{1/2}\;\nabla^2\ell_z(\theta^\star)^{-1/2}\;h\| \\
&\leq \|\mathbf{H}_\lambda{}^{-1/2}(\theta^\star)\;\nabla^2\ell_z(\theta^\star)^{1/2}\|\;\|\nabla^2\ell_z(\theta^\star)^{-1/2}\;h\| \\
&\leq \sqrt{\frac{2\mathsf{B}^\star_2}{\lambda}}\|\nabla^2\ell_z(\theta^\star)^{-1/2}\;h\| \\
&= \sqrt{\frac{2\mathsf{B}^\star_2}{\lambda}}\|h\|_{\nabla^2\ell_z(\theta^\star)^{-1}} \quad ,
\end{aligned}$$

where the last bound is mentioned at the beginning of the proof. Similarly, we can show

$$\|h\|_{\nabla^2\ell_z(\theta^\star)} \leq \sqrt{\frac{2\mathsf{B}^\star_2}{\lambda}}\;\|h\|_{\mathbf{H}_\lambda(\theta^\star)}, \qquad \|h\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leq \sqrt{\frac{2\mathsf{B}^\star_2}{\lambda}}\|h\|_{\nabla^2\ell_z(\theta^\star)^{-1}} \quad . \tag{3.38}$$

Essential supremum. Let $z \in \text{supp}(\rho)$. First note that using Eq. (3.27), we have

$$\|\nabla\ell^\lambda_z(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star_\lambda)} \leq e^{\mathsf{t}_\lambda/2}\|\nabla\ell^\lambda_z(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}.$$

Now bound

$$\|\nabla\ell^\lambda_z(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} \leq \|\nabla\ell^\lambda_z(\theta^\star_\lambda) - \nabla\ell^\lambda_z(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)} + \|\nabla\ell^\lambda_z(\theta^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}.$$

Since $\nabla \ell_z^\lambda(\theta^\star) = \nabla \ell_z(\theta^\star) + \lambda\theta^\star$, the last term is bounded by

$$\mathsf{Bias}_\lambda + \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \leq \mathsf{Bias}_\lambda + \frac{\mathsf{B}_1^\star}{\sqrt{\lambda}}.$$

For the first term, start by using Eq. (3.38).

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \leq \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\, \|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\nabla^2 \ell_z^\lambda(\theta^\star)^{-1}}.$$

Using Eq. (3.29) on ℓ_z^λ, we find

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\nabla^2 \ell_z^\lambda(\theta^\star)^{-1}} \leq \overline{\phi}(\mathsf{t}_\lambda)\, \|\theta_\lambda^\star - \theta^\star\|_{\nabla^2 \ell_z^\lambda(\theta^\star)}.$$

Applying once again Eq. (3.38), we bound

$$\|\theta_\lambda^\star - \theta^\star\|_{\nabla^2 \ell_z^\lambda(\theta^\star)} \leq \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}}\, \|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)}.$$

Finally, using Eq. (3.28) on L_λ, we get

$$\|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \leq \frac{1}{\underline{\phi}(\mathsf{t}_\lambda)} \mathsf{Bias}_\lambda.$$

Hence, putting things together, we get

$$\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \leq \frac{2\mathsf{B}_2^\star}{\lambda} \frac{\overline{\phi}(\mathsf{t}_\lambda)}{\underline{\phi}(\mathsf{t}_\lambda)} \mathsf{Bias}_\lambda = \frac{2\mathsf{B}_2^\star}{\lambda} e^{\mathsf{t}_\lambda} \mathsf{Bias}_\lambda.$$

We the combine all our different computation to get the bound.

Variance. We start by using Eq. (3.27) to show that

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta_\lambda^\star)}\right]^{1/2} \leq e^{\mathsf{t}_\lambda/2} \mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star)}\right]^{1/2}.$$

Then we use the triangle inequality

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star)}\right]^{1/2} \leq \mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star)}\right]^{1/2} + \mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star)}\right]^{1/2}.$$

We can easily bound the last term on the right hand side by $\mathsf{Bias}_\lambda + \mathsf{df}_\lambda$. For the first term, we proceed as in the previous case to obtain

$$\forall z \in \mathrm{supp}(\rho),\ \|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)} \leq \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \overline{\phi}(\mathsf{t}_\lambda) \|\theta_\lambda^\star - \theta^\star\|_{\nabla^2 \ell_z^\lambda(\theta^\star)}.$$

Now taking the expectancy of this inequality squared,

$$\begin{aligned}\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star)}\right]^{1/2} &\leq \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \overline{\phi}(\mathsf{t}_\lambda) \mathbb{E}\left[\|\theta_\lambda^\star - \theta^\star\|^2_{\nabla^2 \ell_z^\lambda(\theta^\star)}\right]^{1/2} \\ &= \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \overline{\phi}(\mathsf{t}_\lambda) \|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)},\end{aligned}$$

where the last equality comes from $\mathbb{E}\left[\nabla^2 \ell_Z^\lambda(\theta^\star)\right] = \mathbf{H}_\lambda(\theta^\star)$. Now applying Eq. (3.28) to L_λ, we obtain

$$\|\theta_\lambda^\star - \theta^\star\|_{\mathbf{H}_\lambda(\theta^\star)} \leq \frac{1}{\underline{\phi}(\mathsf{t}_\lambda)} \mathsf{Bias}_\lambda.$$

Regrouping all these bounds, we obtain

$$\mathbb{E}\left[\|\nabla \ell_z^\lambda(\theta_\lambda^\star) - \nabla \ell_z^\lambda(\theta^\star)\|^2_{\mathbf{H}_\lambda{}^{-1}(\theta^\star)}\right]^{1/2} \leq e^{\mathsf{t}_\lambda} \sqrt{\frac{2\mathsf{B}_2^\star}{\lambda}} \mathsf{Bias}_\lambda.$$

Hence the final bound is proved, regrouping all our computations.

□

3.D .3 Concentration lemmas

Here we concentrate in high probability the quantities obtained in the analytical decomposition. Details on the proof technique are given in Sec. 3.6 of the paper.

Lemma 3.6 (Equivalence of empirical and expected Hessian)**.** *Let $\theta \in \mathcal{H}$ and $n \in \mathbb{N}$. For any $\delta \in (0,1]$, $\lambda > 0$, if*

$$n \geq 24 \frac{\mathsf{B}_2(\theta)}{\lambda} \log \frac{8\mathsf{B}_2(\theta)}{\lambda\delta}, \tag{3.39}$$

then with probability at least $1-\delta$: $\mathbf{H}_\lambda(\theta) \preceq 2\widehat{\mathbf{H}}_\lambda(\theta)$, or equivalently

$$\|\mathbf{H}_\lambda{}^{1/2}(\theta)\widehat{\mathbf{H}}_\lambda^{-1/2}(\theta)\|^2 \leq 2.$$

Proof. By Remark 8 and the definition of $\mathsf{B}_2(\theta)$, the condition we require on n is sufficient to apply proposition 3.10, in particular Eq. (3.51), to $\mathbf{H}_\lambda(\theta), \widehat{\mathbf{H}}_\lambda(\theta)$, for $t = 1/2$, which provides the desired result. □

Lemma 3.7 (Concentration of the empirical gradient)**.** *Let $n \in \mathbb{N}, \delta \in (0,1]$, $0 < \lambda \leq \mathsf{B}_2^\star$. For any $k \geq 4$, if $n \geq k^2 \square_2^2 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{2}{\delta}$, then with probability at least $1-\delta$, we have*

$$\|\nabla \widehat{L}_\lambda(\theta_\lambda^\star)\|_{\mathbf{H}_\lambda{}^{-1}(\theta_\lambda^\star)} \leq \frac{2\sqrt{3}}{k} \mathsf{Bias}_\lambda + 2\square_1 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log \frac{2}{\delta}}{n}}. \tag{3.40}$$

Here, $\square_1, \square_2$ are defined in Lemma 3.5 in Sec. 3.D and $(\mathsf{Q}^\star)^2 = (\mathsf{B}_1^\star)^2/\mathsf{B}_2^\star$.

Proof. **1)** First let us concentrate $\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2} \nabla \widehat{L}_\lambda(\theta_\lambda^\star)$ using a Bernstein-type inequality.

We can see $\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2} \nabla \widehat{L}_\lambda(\theta_\lambda^\star)$ as the mean of n i.i.d. random variables distributed from the law of the vector $\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2} \nabla \ell_z(\theta_\lambda^\star)$.

As we have shown in Lemma 3.5, the essential supremum and variance of this vector is bounded, then we can use Bernstein inequality for random vectors (e.g. Thm. 3.3.4 of Yurinsky, 1995): for any $\lambda > 0$, any $n \in \mathbb{N}$ and $\delta \in (0,1]$, with probability at least $1-\delta$, we have

$$\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{-1/2} \nabla \widehat{L}_\lambda(\theta_\lambda^\star)\| \leq \frac{2M \log \frac{2}{\delta}}{n} + \sigma \sqrt{\frac{2 \log \frac{2}{\delta}}{n}},$$

where $M = \sup_{z\in\mathrm{supp}(\rho)} \|\nabla\ell_z(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star_\lambda)}$ and $\sigma = \mathbb{E}\left[\|\nabla\ell_z(\theta^\star_\lambda)\|^2_{\mathbf{H}_\lambda^{-1}(\theta^\star_\lambda)}\right]^{1/2}$.

2) Using the bounds obtained in Lemma 3.5,

$$M \le \square_1 \ \frac{\mathsf{B}^\star_1}{\sqrt{\lambda}} + 2\square_2 \ \frac{\mathsf{B}^\star_2}{\lambda}\mathsf{Bias}_\lambda, \ \sigma \le \square_1 \ \sqrt{\mathsf{df}_\lambda} + \sqrt{2}\square_2 \ \frac{\sqrt{\mathsf{B}^\star_2}}{\sqrt{\lambda}}\mathsf{Bias}_\lambda.$$

3) Injecting these in the Bernstein inequality,

$$\begin{aligned}
\|\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\| &\le \frac{2\left(\square_1 \ \mathsf{B}^\star_1/\sqrt{\lambda} + 2\square_2 \ (\mathsf{B}^\star_2/\lambda)\mathsf{Bias}_\lambda\right) \ \log\frac{2}{\delta}}{n} \\
&\quad + \left(\square_1 \ \mathsf{df}_\lambda^{1/2} + \sqrt{2}\square_2 \ \sqrt{\mathsf{B}^\star_2/\lambda}\mathsf{Bias}_\lambda\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}} \\
&= \left[\frac{4\square_2 \ \mathsf{B}^\star_2 \ \log\frac{2}{\delta}}{\lambda n} + \sqrt{\frac{4\square_2^2 \ \mathsf{B}^\star_2 \ \log\frac{2}{\delta}}{\lambda n}}\right]\mathsf{Bias}_\lambda \\
&\quad + \sqrt{\frac{2\square_1^2 \ \mathsf{df}_\lambda \ \log\frac{2}{\delta}}{n}} + \sqrt{\frac{2 \ \mathsf{B}^\star_2 \ \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2\square_1^2 \ (\mathsf{B}^\star_1)^2/\mathsf{B}^\star_2 \ \log\frac{2}{\delta}}{n}}.
\end{aligned}$$

In the last inequality, we have regrouped the terms with a factor Bias_λ and we have separated the first term of the decomposition in the following way :

$$\frac{2\square_1\mathsf{B}^\star_1\log\frac{2}{\delta}}{\sqrt{\lambda}n} = \sqrt{\frac{2 \ \mathsf{B}^\star_2 \ \log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2\square_1^2 \ (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

Hence, we can bound the second line of the last inequality:

$$\sqrt{\frac{2\square_1^2\mathsf{df}_\lambda\log\frac{2}{\delta}}{n}} + \sqrt{\frac{2\mathsf{B}^\star_2\log\frac{2}{\delta}}{\lambda n}}\sqrt{\frac{2\square_1^2(\mathsf{Q}^\star)^2\log\frac{2}{\delta}}{n}} \le \left(1 + \sqrt{\frac{2\mathsf{B}^\star_2\log\frac{2}{\delta}}{\lambda n}}\right)\sqrt{\frac{2\square_1^2\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\log\frac{2}{\delta}}{n}}.$$

Thus, if we assume that $n \ge k^2\square_2^2\frac{\mathsf{B}^\star_2}{\lambda}\log\frac{2}{\delta}$,

$$\|\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\| \le \left(\frac{4}{k^2} + \frac{2}{k}\right)\mathsf{Bias}_\lambda + \left(1 + \frac{\sqrt{2}}{k}\right)\sqrt{\frac{2\square_1^2 \ \mathsf{df}_\lambda \vee (\mathsf{B}^\star_1)^2/\mathsf{B}^\star_2 \ \log\frac{2}{\delta}}{n}}.$$

In particular, for $k \ge 4$,

$$\|\mathbf{H}_\lambda(\theta^\star_\lambda)^{-1/2}\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\| \le \frac{3}{k}\mathsf{Bias}_\lambda + 2\square_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

□

Lemma 3.8 (control of $\widehat{\mathsf{Var}}_\lambda$)**.** *Let $n \in \mathbb{N}$, $\delta \in (0,1]$ and $0 < \lambda \leq \mathsf{B}_2^\star$. Assume that for a certain $k \geq 5$,*

$$n \geq k^2 \square_2^2 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}.$$

Then with probability at least $1 - 2\delta$, we have

$$\widehat{\mathsf{Var}}_\lambda \leq \frac{6}{k} \mathsf{Bias}_\lambda + 4\square_1 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log \frac{2}{\delta}}{n}}.$$

Here, $\square_1, \square_2$ are defined in Lemma 3.5

Proof.

- First we apply Lemma 3.6 to $\theta = \theta_\lambda^\star$. Since $\mathsf{B}_2(\theta_\lambda^\star) \leq e^{\mathsf{t}_\lambda} \ \mathsf{B}_2^\star = \square_1^2 \ \mathsf{B}_2^\star$, we see that the condition

$$n \geq 24 \frac{\mathsf{B}_2(\theta_\lambda^\star)}{\lambda} \log \frac{8\mathsf{B}_2(\theta_\lambda^\star)}{\lambda\delta}$$

is satisfied if

$$n \geq 24\square_1^2 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}.$$

Because $k \geq 5$ and $\square_2 \geq \square_1$, and we see that the assumption of this lemma imply the conditions above and hence Lemma 3.6 is satisfied. In particular, $\|\mathbf{H}_\lambda(\theta_\lambda^\star)^{1/2} \widehat{\mathbf{H}}_\lambda(\theta_\lambda^\star)^{-1/2}\|^2 \leq 2$.

- Note that the condition of this proposition also imply the conditions of Lemma 3.7, because $\lambda \leq \mathsf{B}_2^\star$ and $\square_1 \geq 1$ imply $\frac{\square_1^2 \mathsf{B}_2^\star}{\lambda\delta} \geq \frac{1}{\delta}$.

□

3.D .4 Final results

First, we find conditions on n such that the hypothesis $\widehat{\mathsf{Var}}_\lambda \leq \frac{\mathsf{r}_\lambda(\theta_\lambda^\star)}{2}$ is satisfied.

Lemma 3.9. *Let $n \in \mathbb{N}$, $\delta \in (0,1]$, $0 < \lambda \leq \mathsf{B}_2^\star$ and*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \ n \geq \triangle_2 \ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{\mathsf{r}_\lambda(\theta^\star)^2} \ \log \frac{2}{\delta},$$

then with probability at least $1 - 2\delta$

$$\frac{\widehat{\mathsf{Var}}_\lambda}{\mathsf{r}_\lambda(\theta_\lambda^\star)} \leq \square_1 \frac{\widehat{\mathsf{Var}}_\lambda}{\mathsf{r}_\lambda(\theta^\star)} \leq \frac{1}{2},$$

where $\square_1$, $\triangle_1, \triangle_2$ are constants defined in definition 3.7.

Proof. Recall that $\widetilde{\mathsf{t}}_\lambda = \frac{\mathsf{Bias}_\lambda}{\mathsf{r}_\lambda(\theta^\star)}$.

Using Lemma 3.8, we see that under the conditions of this lemma, we have

$$\square_1 \frac{\widehat{\mathsf{Var}}_\lambda}{\mathsf{r}_\lambda(\theta^\star)} \leq \frac{6\square_1 \ \mathsf{Bias}_\lambda}{k \ \mathsf{r}_\lambda(\theta^\star)} + 4\square_1^2 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log \frac{2}{\delta}}{n \ \mathsf{r}_\lambda(\theta^\star)^2}}.$$

Thus, taking $k = 24\square_1(1/2 \vee \widetilde{\mathsf{t}}_\lambda)$ and $n \geq 256\square_1^4 \frac{\mathsf{df}_\lambda \vee \mathsf{Q}^\star}{\mathsf{r}_\lambda(\theta^\star)^2} \ \log \frac{2}{\delta}$, both terms in the sum are bounded by $1/4$ hence the result.

Note that here, we have defined

$$\triangle_1 = 576\square_1^2\square_2^2(1/2 \vee \widetilde{\mathsf{t}}_\lambda)^2, \ \triangle_2 = 256\square_1^4,$$

hence the constants in the definition above. □

Proof. **of Theorem 3.4** First we recall that $\triangle_1$, $\triangle_2$, $\square_1$, C_{bias} and C_{var} are defined in definition 3.7, and bounded in proposition 3.7.

First note that, given the requirements on n, by Lemma 3.9, we have $\widehat{\mathsf{Var}}_\lambda \leq \frac{\mathsf{r}_\lambda(\theta^\star_\lambda)}{2}$ with probability at least $1 - 2\delta$. Thus, we are in a position to apply Theorem 3.7 :

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq K_{\text{bias}} \ \mathsf{Bias}_\lambda^2 + K_{\text{var}} \ \widehat{\mathsf{Var}}_\lambda^2,$$

with $K_{\text{bias}}, K_{\text{var}}$ defined in the proof of the theorem. Note that in the proof of Lemma 3.9, we have taken $k = 24\square_1(1/2 \vee \widetilde{\mathsf{t}}_\lambda) \geq 12$. Hence, using Lemma 3.8, we find

$$\widehat{\mathsf{Var}}_\lambda \leq \frac{1}{2}\mathsf{Bias}_\lambda + 4\square_1 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

Hence,

$$\widehat{\mathsf{Var}}_\lambda^2 \leq \frac{1}{2}\mathsf{Bias}_\lambda^2 + 32\square_1^2\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n},$$

which yields the wanted result with $\mathsf{C}_{\text{bias}} = K_{\text{bias}} + \frac{1}{2}K_{\text{var}}$ and $\mathsf{C}_{\text{var}} = 32\square_1^2 K_{\text{var}}$.

□

Proof. **of Theorem 3.3**

We get this theorem as a corollary of Theorem 3.4. Indeed, $\forall\lambda \leq \mathsf{B}_2^\star$, $\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^* \leq \frac{(\mathsf{B}_1^\star)^2}{\lambda}$, hence the result.

□

3.E Explicit bounds for the simplified case

In this section, assume that Assumptions 3.1, 3.3 to 3.5 and 3.8 hold.

Define the following constant N :

$$N = 36A^2\log^2\left(6A^2\frac{1}{\delta}\right) \ \vee \ 256\frac{1}{A^2} \ \log\frac{2}{\delta} \ \vee \ 512\left(\|\theta^\star\|^2R^2 \vee 1\right)\log\frac{2}{\delta}, \tag{3.41}$$

where $A = \frac{\overline{\mathsf{B}}_2}{\overline{\mathsf{B}}_1}$.

We have the following slow rates theorem.

Theorem 3.8 (Quantitative slow rates result)**.** *Let $n \in \mathbb{N}$. Let $\delta \in (0, 1]$. Setting*

$$\lambda = 16((R \vee 1)\overline{\mathsf{B}}_1) \ \frac{1}{\sqrt{n}} \ \log^{1/2}\frac{2}{\delta},$$

if $n \geq N$, with probability at least $1-2\delta$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq 48 \ \max(R,1)\max(\|\theta^\star\|^2,1)\overline{\mathsf{B}}_1 \frac{1}{\sqrt{n}} \ \log^{1/2}\frac{2}{\delta}, \tag{3.42}$$

and $N = O\left(\mathrm{poly}(\overline{\mathsf{B}}_1, \overline{\mathsf{B}}_2, R\|\theta^\star\|)\right)$ is given explicitly in Eq. (3.41). Here, poly *denotes a certain rational function of the inputs.*

Proof. Note that $\overline{\mathsf{df}}_\lambda \leq \frac{\overline{\mathsf{B}}_1^2}{\lambda}$. Hence, if $\lambda \leq \overline{\mathsf{B}}_2$, then $\overline{\mathsf{df}}_\lambda \vee (\overline{\mathsf{B}}_1^2/\overline{\mathsf{B}}_2) \leq \frac{\overline{\mathsf{B}}_1^2}{\lambda}$.

1) Let us reformulate Theorem 3.6. Let $n \in \mathbb{N}$ and $0 < \lambda \leq \overline{\mathsf{B}}_2$. Let $\delta \in (0,1]$. If

$$n \geq 512\left(\|\theta^\star\|^2 R^2 \vee 1\right)\log\frac{2}{\delta}, \ n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \ n \geq 16\triangle^2 \ \frac{R^2\overline{\mathsf{B}}_1^2}{\lambda^2} \ \log\frac{2}{\delta},$$

then with probability at least $1-2\delta$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \mathsf{C}_{\mathrm{var}} \ \frac{\overline{\mathsf{B}}_1^2}{\lambda n} \ \log\frac{2}{\delta} + \mathsf{C}_{\mathrm{bias}}\lambda\|\theta^\star\|^2,$$

where $\triangle, \mathsf{C}_{\mathrm{bias}}, \mathsf{C}_{\mathrm{var}}$ are defined in definition 3.5.

2) Now setting $\lambda = 16R\overline{\mathsf{B}}_1 \log^{1/2}\frac{2}{\delta}\frac{1}{n^{1/2}}$, we see that the inequality

$$n \geq 16\triangle^2 \ \frac{R^2\overline{\mathsf{B}}_1^2}{\lambda^2} \ \log\frac{2}{\delta}$$

is automatically satisfied since $\triangle \leq 4$. Hence, if

$$n \geq 512\left(\|\theta^\star\|^2 R^2 \vee 1\right)\log\frac{2}{\delta}, \ n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \ 0 < \lambda \leq \overline{\mathsf{B}}_2,$$

then

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \frac{\mathsf{C}_{\mathrm{var}}}{256} \ \frac{1}{R^2}\lambda + \mathsf{C}_{\mathrm{bias}}\lambda\|\theta^\star\|^2 \leq \left(\frac{\mathsf{C}_{\mathrm{var}}}{256} + \mathsf{C}_{\mathrm{bias}}\right)\max(\frac{1}{R^2}, \|\theta^\star\|^2)\lambda.$$

Since by definition 3.5, $\mathsf{C}_{\mathrm{var}} \leq 84$ and $\mathsf{C}_{\mathrm{bias}} \leq 2$, we get

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq 3\max(\frac{1}{R^2}, \|\theta^\star\|^2)\lambda.$$

3) Having our fixed $\lambda = 16\frac{\overline{\mathsf{B}}_1 R \ \log^{1/2}\frac{2}{\delta}}{n^{1/2}}$, let us look for conditions for which

$$n \geq 512\left(\|\theta^\star\|^2 R^2 \vee 1\right)\log\frac{2}{\delta}, \ n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}, \ 0 < \lambda \leq \overline{\mathsf{B}}_2,$$

are satisfied.

To deal with $n \geq 24\frac{\overline{\mathsf{B}}_2}{\lambda}\log\frac{8\overline{\mathsf{B}}_2}{\lambda\delta}$, bound

$$\frac{\overline{\mathsf{B}}_2}{\lambda} \leq \frac{1}{16}\frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1 \ \log^{1/2}\frac{2}{\delta}} \ n^{1/2} \leq \frac{1}{8}\frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1} \ n^{1/2},$$

where we have used the fact that $\log^{1/2} \frac{2}{\delta} \geq \frac{1}{2}$. apply Lemma 3.14 with $a_1 = 3, a_2 = 1, A = \frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1}$ to get the following condition:

$$n \geq 4a_1^2 A^2 \log^2\left(\frac{2a_1 a_2 A^2}{\delta}\right),$$

which we express as

$$n \geq 36A^2 \log^2\left(6A^2 \frac{1}{\delta}\right).$$

To deal with the bound $\lambda < \overline{\mathsf{B}}_2$, we need only apply the definition to obtain

$$n \geq 256 \frac{R^2 \overline{\mathsf{B}}_1^2}{\overline{\mathsf{B}}_2^2} \log \frac{2}{\delta}.$$

Thus, we can concentrate all these bounds as $n \geq N$ where

$$N = 36A^2 \log^2\left(6A^2 \frac{1}{\delta}\right) \vee 256 \frac{1}{A^2} \log \frac{2}{\delta} \vee 512 \left(\|\theta^\star\|^2 R^2 \vee 1\right) \log \frac{2}{\delta},$$

where $A = \frac{\overline{\mathsf{B}}_2}{R\overline{\mathsf{B}}_1}$.

4) Since R is only an upper bound, we can replace R by $R \vee 1$. In this case, we see that $A \leq \frac{\overline{\mathsf{B}}_2}{\overline{\mathsf{B}}_1}$ and $\max(\frac{1}{R\vee 1}, (R \vee 1)\|\theta^\star\|^2) \leq (R \vee 1)(\|\theta^\star\| \vee 1)^2$ hence the final bounds.

□

3.F Explicit bounds for the refined case

In this part, we continue to assume Assumptions 3.1, 3.3 to 3.5 and 3.8. We present a classification of distributions ρ and show that we can achieve better rates than the classical slow rates.

Definition 3.8 (class of distributions)**.** *Let $\alpha \in [1, +\infty]$ and $r \in [0, 1/2]$.*
We denote with $\mathcal{P}_{\alpha,r}$ the set of probability distributions ρ such that there exists $\mathsf{L}, \mathsf{Q} \geq 0$,

- $\mathsf{Bias}_\lambda \leq \mathsf{L}\ \lambda^{\frac{1+2r}{2}}$
- $\mathsf{df}_\lambda \leq \mathsf{Q}^2\ \lambda^{-1/\alpha}$,

where this holds for any $0 < \lambda \leq 1$. For simplicity, if $\alpha = +\infty$, we assume that $\mathsf{Q} \geq \mathsf{Q}^\star$.

Note that given our assumptions, we always have

$$\rho \in \mathcal{P}_{1,0},\ \mathsf{L} = \|\theta^\star\|,\ \mathsf{Q} = \mathsf{B}_1^\star. \tag{3.43}$$

We also define

$$\lambda_1 = \left(\frac{\mathsf{Q}}{\mathsf{Q}^\star}\right)^{2\alpha} \wedge 1, \tag{3.44}$$

such that

$$\forall \lambda \leq \lambda_1,\ \mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \leq \frac{\mathsf{Q}^2}{\lambda^{1/\alpha}}.$$

Interpretation of the classes

- The bias term Bias_λ characterizes the regularity of the objective $\theta^\star$. In a sense, if r is big, then this means $\theta^\star$ is very regular and will be easier to estimate. The following results reformulates this intuition.

  Remark 5 (source condition)**.** *Assume there exists* $0 \leq r \leq 1/2$ *and* $v \in \mathcal{H}$ *such that*

  $$P_{\mathbf{H}(\theta^\star)}\theta^\star = \mathbf{H}(\theta^\star)^r v.$$

  Then we have

  $$\forall \lambda > 0,\ \mathsf{Bias}_\lambda \leq \mathsf{L}\ \lambda^{\frac{1+2r}{2}},\ \mathsf{L} = \|\mathbf{H}(\theta^\star)^{-r}\theta^\star\|.$$

- The effective dimension df_λ characterizes the size of the space $\mathcal{H}$ with respect to the problem. The higher α, the smaller the space. If $\mathcal{H}$ is finite dimensional for instance, $\alpha = +\infty$.

We will give explicit bounds for the performance of $\widehat{\theta}^\star_\lambda$ depending on which class ρ belongs to, i.e., as a function of α, r.

Well -behaved problems $\mathsf{r}_\lambda(\theta^\star)$ has a limiting role. However, as soon as we have some sort of regularity, this role is no longer limiting, i.e. this quantity does not appear in the final rates and the constants in these rates have no dependence on the problem. This motivates the following definition.

We say that a problem is well behaved if the following equation holds.

$$\forall \delta \in (0, \frac{1}{2}],\ \exists \lambda_0(\delta) \in (0,1],\ \forall 0 < \lambda \leq \lambda_0(\delta),\ \frac{\mathsf{L}\lambda^{1/2+r}}{\mathsf{r}_\lambda(\theta^\star)}\ \log\frac{2}{\delta} \leq \frac{1}{2}. \tag{3.45}$$

Remark 6 (well-behaved problems)**.** *Note that Eq.* (3.45) *is satisfied if one of the following holds.*

- *If* $R = 0$*, then the condition holds for* $\lambda_0 = 1$*.*
- *If* $r > 0$*, then the condition holds for* $\lambda_0 = (2\mathsf{L}R\log\frac{2}{\delta})^{-1/r} \wedge 1$*.*
- *If there exists* $\mu \in [0,1)$ *and* $\mathsf{F} \geq 0$ *such that* $\mathsf{r}_\lambda(\theta^\star) \geq \frac{1}{\mathsf{F}}\lambda^{\mu/2}$*, then this holds for* $\lambda_0 = (2R\mathsf{F}\log\frac{2}{\delta})^{-2/(1-\mu+2r)} \wedge 1$*.*

Moreover, if Eq. (3.45) *is satisfied, than for any* $\lambda \leq \lambda_0$*,* $\mathsf{t}_\lambda \leq \log 2$*.*

Note that the first possible condition corresponds to the case where the loss functions are quadratic in θ (if the loss is the square loss for instance). The second condition corresponds to having a strict source condition, i.e. something strictly better than just $\theta^\star \in \mathcal{H}$. Finally, the third condition corresponds to the fact that the radius r_λ decreases slower than the original bound of $\mathsf{r}_\lambda \geq \frac{\lambda^{1/2}}{R}$, and hence it is not limiting.

Note that a priori, using only the assumptions, our problems do not satisfy Eq. (3.45) (see Eq. (3.43), and the fact that $\mathsf{r}_\lambda \geq \frac{\sqrt{\lambda}}{R}$).

3.F .1 Quantitative bounds

In this section, for any given pair (α, r) characterizing the regularity and size of the problem, we associate

$$\beta = \frac{1}{1+2r+1/\alpha}, \ \gamma = \frac{\alpha(1+2r)}{\alpha(1+2r)+1}.$$

In what follows, we define

$$N = \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta} \ \vee \ \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)}, \tag{3.46}$$

where $A = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$, λ_0 is given by Eq. (3.45) and λ_1 is given by Eq. (3.44) : $\lambda_1 = \frac{\mathsf{Q}^{2\alpha}}{(\mathsf{Q}^\star)^{2\alpha}}$.

Theorem 3.9 (Quantitative results when Eq. (3.45) is satisfied and $\alpha < \infty$ or $r > 0$)**.** *Let $\rho \in \mathcal{P}_{\alpha,r}$ and that we have either $\alpha < \infty$ or $r > 0$. Let $\delta \in (0, \frac{1}{2}]$.*
If Eq. (3.45) is satisfied, and

$$n \geq N, \ \lambda = \left(256 \ \left(\frac{\mathsf{Q}}{\mathsf{L}}\right)^2 \ \frac{1}{n}\right)^\beta,$$

then with probability at least $1-2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq 8 \ (256)^\gamma \ \left(\mathsf{Q}^\gamma \ \mathsf{L}^{1-\gamma}\right)^2 \ \frac{1}{n^\gamma}\log\frac{2}{\delta},$$

where N is defined in Eq. (3.46).

Proof. Using the definition of λ_1, as soon as $\lambda \leq \lambda_1$ we have $\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \leq \mathsf{Q}^2\lambda^{-1/\alpha}$.

Let us formulate Theorem 3.4 using the fact that $\rho \in \mathcal{P}_{\alpha,r}$.

Let $\delta \in (0,1]$, $0 < \lambda \leq \mathsf{B}_2^\star \wedge \lambda_1$ and $n \in \mathbb{N}$ such that

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda}\log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}, \ n \geq \triangle_2 \ \frac{\mathsf{Q}^2}{\lambda^{1/\alpha}\mathsf{r}_\lambda(\theta^\star)^2} \ \log\frac{2}{\delta},$$

then with probability at least $1-2\delta$

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq \mathsf{C}_{\text{bias}} \ \mathsf{L}^2\lambda^{1+2r} + \mathsf{C}_{\text{var}} \ \frac{\mathsf{Q}^2}{\lambda^{1/\alpha}n} \ \log\frac{2}{\delta},$$

where $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}$ are defined in definition 3.7. Now let us distinguish the two cases of our theorem.

Assume that ρ satisfies Eq. (3.45) . In this case the proof proceeds as follows. Note that as soon as $\lambda \leq \lambda_0$, we have $\frac{\mathsf{Bias}_\lambda}{\mathsf{r}_\lambda(\theta^\star)} \leq \frac{1}{2}$ and hence the bounds in proposition 3.7 apply.

1) First, we find a simple condition to guarantee

$$\mathsf{r}_\lambda(\theta^\star)^2\lambda^{1/\alpha} \geq \triangle_2 \ \mathsf{Q}^2\frac{1}{n} \ \log\frac{2}{\delta}.$$

Using the fact that Eq. (3.45) is satisfied, we see that if $\lambda \leq \lambda_0$, then $\mathsf{r}_\lambda \geq 2\mathsf{L}\lambda^{1/2+r}\log\frac{2}{\delta}$. Hence, this condition is satisfied if

$$\lambda \leq \lambda_0,\ 4\mathsf{L}^2\lambda^{1+2r+1/\alpha} \geq \triangle_2\ \mathsf{Q}^2\frac{1}{n}.$$

2) Now fix $\mathsf{C}_\lambda = 256 \geq \triangle_2/4$ (see proposition 3.7) and fix

$$\lambda^{1+2r+1/\alpha} = \mathsf{C}_\lambda\frac{\mathsf{Q}^2}{\mathsf{L}^2}\ \frac{1}{n} \iff \lambda = \left(\mathsf{C}_\lambda\frac{\mathsf{Q}^2}{\mathsf{L}^2}\ \frac{1}{n}\right)^\beta.$$

where $\beta = 1/(1+2r+1/\lambda) \in [1/2, 1)$.

Using our restatement of Theorem 3.4, we have that with probability at least $1-2\delta$,

$$L(\widehat{\theta}^\star_\lambda) - L(\theta^\star) \leq \left(\mathsf{C}_{\text{bias}} + \frac{1}{\mathsf{C}_\lambda}\mathsf{C}_{\text{var}}\log\frac{2}{\delta}\right)\ \mathsf{L}^2\lambda^{1+2r} \leq \mathsf{K}\ \log\frac{2}{\delta}\ \mathsf{L}^2\lambda^{1+2r},$$

where we have set $\mathsf{K} = \left(\mathsf{C}_{\text{bias}} + \frac{1}{256}\mathsf{C}_{\text{var}}\right) \leq 8$ (see proposition 3.7).
This result holds provided

$$0 < \lambda \leq \mathsf{B}^\star_2 \wedge \lambda_0 \wedge \lambda_1,\ n \geq \triangle_1\frac{\mathsf{B}^\star_2}{\lambda}\log\frac{8\square_1^2\mathsf{B}^\star_2}{\lambda\delta}. \tag{3.47}$$

Indeed, we have shown in the previous point that since $\mathsf{C}_\lambda \geq \frac{\triangle_2}{4}$, $\mathsf{r}_\lambda(\theta^\star)^2\lambda^{1/\alpha} \geq \triangle_2\ \mathsf{Q}^2\frac{1}{n}\ \log\frac{2}{\delta}$.

3) Let us now work to guarantee the conditions in Eq. (3.47).
First, to guarantee $n \geq \triangle_1\frac{\mathsf{B}^\star_2}{\lambda}\log\frac{8\square_1^2\mathsf{B}^\star_2}{\lambda\delta}$, bound

$$\frac{\mathsf{B}^\star_2}{\lambda} = \frac{\mathsf{B}^\star_2\mathsf{L}^{2\beta}n^\beta}{\mathsf{C}^\beta_\lambda\mathsf{Q}^{2\beta}\ \log^\beta\frac{2}{\delta}} \leq \frac{2}{\mathsf{C}^\beta_\lambda}\ \frac{\mathsf{B}^\star_2\mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}n^\beta.$$

Then apply Lemma 3.15 with $a_1 = \frac{2\triangle_1}{\mathsf{C}^\beta_\lambda}$, $a_2 = \frac{16\square_1^2}{\mathsf{C}^\beta_\lambda}$, $A = \frac{\mathsf{B}^\star_2\mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$. Since $\beta \geq 1/2$, using the bounds in proposition 3.7, we find $a_1 \leq 648$ and $a_2 \leq 4$, hence the following sufficient condition:

$$n \geq \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)}.$$

Then, to guarantee the condition

$$\lambda \leq \mathsf{B}^\star_2 \wedge \lambda_0 \wedge \lambda_1,$$

we simply need

$$n \geq \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}^\star_2 \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta}.$$

Hence, defining

$$N = \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}^\star_2 \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta}\ \vee\ \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)},$$

where $A = \frac{\mathsf{B}^\star_2\mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$, we see that as soon as $n \geq N$, Eq. (3.47) holds.

□

We now state the following corollary, for $r > 0$. We define N in the following way:

$$N = \frac{256\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta} \vee \left(1296\frac{1}{1-\beta}A\log\left(5184\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)} \tag{3.48}$$

where $A = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$, $\lambda_0 = (2\mathsf{L}R\log\frac{2}{\delta})^{-1/r} \wedge 1$ and $\lambda_1 = \frac{\mathsf{Q}^{2\alpha}}{(\mathsf{Q}^\star)^{2\alpha}}$.

Corollary 3.4. *Assume $\rho \in \mathcal{P}_{\alpha,r}$ with $r > 0$. Let $\delta \in (0, 0.5]$ and $n \geq N$, where N is defined in Eq. (3.48). For*

$$\lambda = \left(256\ \left(\frac{\mathsf{Q}}{\mathsf{L}}\right)^2\ \frac{1}{n}\right)^\beta,$$

with probability at least $1 - 2\delta$,

$$L(\widehat{\theta}_\lambda^\star) - L(\theta^\star) \leq 8\ (256)^\gamma\ \left(\mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma}\right)^2\ \frac{1}{n^\gamma}\log\frac{2}{\delta},$$

Moreover, $N = O\left(\text{poly}\left(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, \mathsf{Q}, R, \log\frac{1}{\delta}\right)\right)$, which means that N is bounded by a rational function of the arguments of poly.

Proof. **of Cor. 3.2** We simply apply Cor. 3.4 for $\alpha = 1$ and $\mathsf{Q} = \mathsf{B}_1^\star$. □

3.G Additional lemmas

3.G .1 Self-concordance, sufficient conditions to define L and related quantities

In this section, we will consider an arbitrary probability measure μ on $\mathcal{Z}$. We assume that ℓ_z satisfies Assumption 3.8 with a certain given function φ. Recall that $R^\mu = \sup_{z\in\text{supp}(\mu)}\sup_{g\in\varphi(z)}\|g\|$. In this section, we will also assume that $R^\mu < \infty$.

Lemma 3.10 (Gronwall lemma)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that*

$$\forall t \in \mathbb{R},\ \varphi'(t) \leq C\varphi(t).$$

Then

$$\forall (t_0, t_1) \in \mathbb{R}^2,\ \varphi(t_1) \leq e^{C|t_1 - t_0|}\varphi(t_0).$$

Lemma 3.11. *Assume that there exists θ_0 such that $\sup_{z\in\text{supp}(\mu)}\text{Tr}\left(\nabla^2\ell_z(\theta_0)\right) < \infty$*

- *$\sup_{z\in\text{supp}(\mu)}\text{Tr}\left(\nabla^2\ell_z(\theta)\right) < \infty$ for any $\theta \in \mathcal{H}$;*
- *For any given radius $T > 0$, and any $\|\theta_0\| \leq T$, we have*

$$\forall \|\theta\| \leq T, \forall z \in \mathcal{Z},\ \text{Tr}\left(\nabla^2\ell_z(\theta)\right) \leq \exp\left(2R^\mu T\right)\text{Tr}\left(\nabla^2\ell_z(\theta_0)\right) < \infty.$$

Proof. Let $z \in \text{supp}(\mu)$ be fixed. Using the same reasoning as in the proof of Eq. (3.27), we can show

$$\forall \theta_0, \theta_1 \in \mathcal{H},\ \nabla^2\ell_z(\theta_1) \preceq \exp\left(\sup_{g\in\varphi(z)}|g\cdot(\theta_1 - \theta_0)|\right)\nabla^2\ell_z(\theta_0) \preceq \exp\left(R^\mu\|\theta_1 - \theta_0\|\right)\nabla^2\ell_z(\theta_0)$$

Where we have used the fact that $R^\mu = \sup_{z\in\text{supp}(\mu)}\sup_{g\in\varphi(z)}\|g\| < \infty$ Thus, in particular

$$\forall z \in \text{supp}(\mu),\ \forall \theta_0, \theta_1 \in \mathcal{H},\ \text{Tr}\left(\nabla^2\ell_z(\theta_1)\right) \leq \exp\left(R^\mu\|\theta_1 - \theta_0\|\right)\text{Tr}\left(\nabla^2\ell_z(\theta_0)\right),$$

which leads to the desired bounds.

□

Lemma 3.12. *Assume that there exists θ_0 such that*

$$\sup_{z\in\text{supp}(\mu)} \text{Tr}\left(\nabla^2\ell_z(\theta_0)\right) < \infty, \qquad \sup_{z\in\text{supp}(\mu)} \|\nabla\ell_z(\theta_0)\| < \infty.$$

Then

- $\sup_{z\in\text{supp}(\mu)} \|\nabla\ell_z(\theta)\| < \infty$ *for any* $\theta \in \mathcal{H}$
- *For any* $T > 0$ *and any* $\|\theta_0\|, \|\theta\| \le T, z \in \text{supp}(\mu)$,

$$\begin{aligned}\|\nabla\ell_z(\theta)\| \le \|\nabla\ell_z(\theta_0)\| &+ 2T\ \text{Tr}\left(\nabla^2\ell_z(\theta_0)\right)\\ &+ 4R^\mu\ \psi(2R^\mu T)\ \ \text{Tr}\left(\nabla^2\ell_z(\theta_0)\right)\ \ R^2.\end{aligned}$$

Proof. Fix $z \in \mathcal{Z}$, $\theta_0, \theta_1 \in \mathcal{H}$ and $h \in \mathcal{H}$. Let us look at the function

$$f : t \in [0,1] \mapsto \left(\nabla\ell_z(\theta_t) - \nabla\ell_z(\theta_0) - t\nabla^2\ell_z(\theta_0)(\theta_1-\theta_0)\right)\cdot h.$$

We have $f''(t) = \nabla^3\ell_z(\theta_t)[\theta_1-\theta_0, \theta_1-\theta_0, h]$. By the self-concordant assumption, we have

$$\begin{aligned}\left|f''(t)\right| &\le \sup_{g\in\varphi(z)} |g\cdot h|\nabla^2\ell_z(\theta_t)[\theta_1-\theta_0, \theta_1-\theta_0]\\ &\le \sup_{g\in\varphi(z)} |g\cdot h| \exp(t \sup_{g\in\varphi(z)} |g\cdot\theta_1 - \theta_0|)\|\theta_1-\theta_0\|^2_{\nabla^2\ell_z(\theta_0)}.\end{aligned}$$

Integrating this knowing $f'(0) = f(0) = 0$ yields

$$|f(1)| \le \sup_{g\in\varphi(z)} |g\cdot h|\ \psi(\sup_{g\in\varphi(z)} |g\cdot(\theta_1-\theta_0)|)\|\theta_1-\theta_0\|^2_{\nabla^2\ell_z(\theta_0)}.$$

Hence :

$$\|\nabla\ell_z(\theta_1)-\nabla\ell_z(\theta_0)\| \le \|\nabla^2\ell_z(\theta_0)\|\ \|\theta_1-\theta_0\| + \|\varphi(z)\|\ \psi(\sup_{g\in\varphi(z)} |g\cdot(\theta_1-\theta_0)|)\ \|\nabla^2\ell_z(\theta_0)\|\ \|\theta_1-\theta_0\|^2$$

where $\psi(t) = (e^t - t - 1)/t^2$. Then, noting that $\|\nabla^2\ell_z(\theta)\| \le \text{Tr}(\nabla^2\ell_z(\theta))$, we have proved our lemma.

□

Lemma 3.13. *Assume that there exists θ_0 such that*

$$\sup_{z\in\text{supp}(\mu)} \text{Tr}\left(\nabla^2\ell_z(\theta_0)\right) < \infty, \qquad \sup_{z\in\text{supp}(\mu)} \|\nabla\ell_z(\theta_0)\| < \infty, \qquad \sup_{z\in\text{supp}(\mu)} |\ell_z(\theta_0)| < \infty.$$

Then

- *For any* $\theta \in \mathcal{H}$, $\sup_{z\in\text{supp}(\mu)} |\ell_z(\theta)| < \infty$
- *For any* $\theta_0 \in \mathcal{H}$, $T \ge \|\theta_0\|, \|\theta\| \le T,\ z \in \text{supp}(\mu)$, *we have:*

$$|\ell_z(\theta)| \le |\ell_z(\theta_0)| + 2\|\nabla\ell_z(\theta_0)\|T + \psi(2R^\mu\ T)\ \ \text{Tr}(\nabla^2\ell_z(\theta_0))\ T^2.$$

Proof. Proceeding as in the proof of Eq. (3.30), we get

$$\forall z \in \mathcal{Z},\ \forall \theta_0, \theta_1 \in \mathcal{H},\ 0 \leq \ell_z(\theta_1) - \ell_z(\theta_0) - \nabla \ell_z(\theta_0)(\theta_1 - \theta_0) \leq \psi\big(\sup_{g \in \varphi(z)} |g \cdot (\theta_1 - \theta_0)|\big) \|\theta_1 - \theta_0\|^2_{\nabla^2 \ell_z(\theta_0)}$$

where $\psi(t) = (e^t - t - 1)/t^2$.

□

To conclude, we give the following result.

Proposition 3.8. *Let $\lambda \geq 0$. If a probability measure μ and ℓ satisfy Assumptions 3.3, 3.4 and 3.8, the function $L_{\mu,\lambda}(\theta) := \mathbb{E}_\mu[\ell_z(\theta)] + \lambda\|\theta\|^2$ and $\nabla L_{\mu,\lambda}(\theta), \nabla^2 L_{\mu,\theta}(\theta)$ are well-defined for any $\theta \in \mathcal{H}$, and we can differentiate under the expectation. Moreover,*

$$\forall \theta \in \mathcal{H}, \quad \sup_{z \in \mathrm{supp}(\rho)} |\ell_z(\theta)|, \quad \sup_{z \in \mathrm{supp}(\rho)} \|\nabla \ell_z(\theta)\|, \quad \sup_{z \in \mathrm{supp}(\rho)} \mathrm{Tr}\left(\nabla^2 \ell_z(\theta)\right) < \infty.$$

Proof. We combine the results given in Lemmas 3.11 to 3.13. □

3.G .2 Bernstein inequalities for operators

We start by proposing a slight modification of Proposition 6 in (Rudi and Rosasco, 2017). First we need to introduce the following quantitity and some notation for Hermitian operators. We denote by $\preceq$ is the partial order between positive semidefinite Hermitian operators. Let A, B be bounded Hermitian operators on $\mathcal{H}$,

$$A \preceq B \iff v \cdot (Av) \leq v \cdot (Bv),\ \forall v \in \mathcal{H} \iff B - A \text{ is positive semidefinite.}$$

Let q be a random positive semi-definite operator and let $\mathbf{Q} := \mathbb{E}[q]$, denote by $\mathcal{F}_\infty(\lambda)$ the function of λ defined as

$$\mathcal{F}_\infty(\lambda) := \operatorname{ess\,sup} \mathrm{Tr}\left(\mathbf{Q}_\lambda^{-1/2} q \mathbf{Q}_\lambda^{-1/2}\right),$$

where ess sup is the *essential support* of q.

Remark 7. *Note that if $\mathrm{Tr}(q) \leq c_0$, for a $c_0 > 0$ almost surely, then $\mathcal{F}_\infty(\lambda) \leq c_0/\lambda$. Vice versa, if $\mathcal{F}_\infty(\lambda_0) < \infty$ for a given $\lambda_0 > 0$, then $\mathrm{Tr}(q) \leq (\|Q\| + \lambda_0)\mathcal{F}_\infty(\lambda_0)$ almost surely, moreover $\mathcal{F}_\infty(\lambda) < \frac{\|\mathbf{Q}\|+\lambda_0}{\|\mathbf{Q}\|+\lambda}\mathcal{F}(\lambda_0)$ for any $\lambda > 0$.*

Proposition 3.9 (Prop. 6 of (Rudi and Rosasco, 2017))**.** *Let $q_1, ..., q_n$ be identically distributed random positive semi-definite operators on a separable Hilbert space $\mathcal{H}$ such that the q are trace class and $\mathbf{Q} = \mathbb{E}[q]$. Let $\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^n q_i$ and take $0 < \lambda \leq \|\mathbf{Q}\|$ and assume $\mathcal{F}_\infty(\lambda) < \infty$. For any $\delta > 0$, the following holds with probability at least $1 - \delta$:*

$$\|\mathbf{Q}_\lambda^{-1/2}(\mathbf{Q} - \mathbf{Q}_n)\mathbf{Q}_\lambda^{-1/2}\| \leq \frac{2\beta(1 + \mathcal{F}_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}},\ \beta = \log\frac{8\mathcal{F}_\infty(\lambda)}{\delta}$$

Proof. Use Proposition 3 of (Rudi and Rosasco, 2017) and proceed as in the proof of Proposition 6 of (Rudi and Rosasco, 2017) except that we bound $\mathrm{Tr}(\mathbf{Q}_\lambda^{-1}\mathbf{Q}) \leq \mathcal{F}_\infty(\lambda)$ instead of bounding $\mathrm{Tr}(\mathbf{Q}_\lambda^{-1}\mathbf{Q}) \leq \frac{\mathrm{Tr}(\mathbf{Q})}{\lambda}$, we find this result. □

Here we slightly extend the results of Prop. 8 and Prop. 6 of (Rudi and Rosasco, 2017), to extend the range of λ for which the result on the partial order between operators holds, from $0 < \lambda < \|\mathbf{Q}\|$ to $\lambda > 0$.

Proposition 3.10 (Prop. 8 together with Prop. 6 of (Rudi and Rosasco, 2017)). *Let $q_1, ..., q_n$ be identically distributed random positive semi-definite operators on a separable Hilbert space $\mathcal{H}$ such that the q are trace class and $\mathbf{Q} = \mathbb{E}[q]$. Let $\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^n q_i$. Let any $\delta \in (0,1]$, $t > 0, 0 < \lambda \leq \|\mathbf{Q}\|$ and assume $\mathcal{F}_\infty(\lambda) < \infty$, when*

$$n \geq 8\mathcal{F}_\infty(\lambda) \log \frac{8\mathcal{F}_\infty(\lambda)}{\delta} \left(\frac{1}{4t^2} + \frac{1}{t}\right) \tag{3.49}$$

then the following holds with probability at least $1 - \delta$:

$$\|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\| \leq t. \tag{3.50}$$

Moreover let $\lambda > 0, \delta \in (0,1]$ and Eq. (3.49) is satisfied for $t \leq 1/2$, then the following holds with probability at least $1 - \delta$,

$$\mathbf{Q}_\lambda \preceq 2\mathbf{Q}_{n,\lambda}, \iff \|\mathbf{Q}_{n,\lambda}^{-1/2}\mathbf{Q}_\lambda^{1/2}\|^2 \leq 2. \tag{3.51}$$

Finally, let $\lambda > 0, \delta \in (0,1]$, Eq. (3.49) is satisfied for $t \leq 1/2$ and

$$n \geq 16 \frac{c_0^2}{\|\mathbf{Q}\|^2} \log \frac{2}{\delta},$$

with $c_0 = ess\sup \mathrm{Tr}(q)$, then the following holds with probability at least $1 - \delta$,

$$\mathbf{Q}_{n,\lambda} \preceq \frac{3}{2}\mathbf{Q}_\lambda, \iff \|\mathbf{Q}_{n,\lambda}^{1/2}\mathbf{Q}_\lambda^{-1/2}\|^2 \leq 3/2. \tag{3.52}$$

Proof. **Point 1)** Let $\delta \in (0,1]$ and $0 < \lambda \leq \mathbf{Q}$. Using proposition 3.9, we have that with probability at least $1 - \delta$,

$$\|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\| \leq \frac{2\beta(1 + \mathcal{F}_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}, \ \beta = \log \frac{8\mathcal{F}_\infty(\lambda)}{\delta}.$$

Now note that if $\lambda \leq \|\mathbf{Q}\|$, we have

$$\frac{1}{2} \leq \frac{\|\mathbf{Q}\|}{\|\mathbf{Q}\| + \lambda} = \|\mathbf{Q}_\lambda^{-1}\mathbf{Q}\| \leq \mathrm{Tr}\left(\mathbf{Q}_\lambda^{-1}\mathbf{Q}\right) \leq \mathcal{F}_\infty(\lambda).$$

Thus we can bound $1 + \mathcal{F}_\infty(\lambda) \leq 3\mathcal{F}_\infty(\lambda)$, and we rewrite the previous bound

$$\|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\| \leq \frac{2\beta\mathcal{F}_\infty(\lambda)}{n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}, \ \beta = \log \frac{8\mathcal{F}_\infty(\lambda)}{\delta}.$$

Point 2) Now let $t > 0$, $\delta \in (0,1]$ and $0 < \lambda \leq \|\mathbf{Q}\|$. If

$$n \geq 8\mathcal{F}_\infty(\lambda)\beta \left(\frac{1}{4t^2} + \frac{1}{t}\right),$$

then

$$\|\mathbf{Q}_\lambda^{-1/2} (\mathbf{Q} - \mathbf{Q}_n) \mathbf{Q}_\lambda^{-1/2}\| \leq t.$$

Indeed, assume we want to find $n_0 > 0$ for which for all $n \geq n_0$, $\frac{A}{n} + \sqrt{\frac{B}{n}} \leq \frac{1}{2}$ where $A, B \geq 0$. setting $x = \sqrt{n}$, this is equivalent to finding x_0 such that $\forall x \geq x_0$, $\frac{x^2}{2} - \sqrt{B}x - A \geq 0$. A

sufficient condition for this is that $x \geq \sqrt{B} + \sqrt{B+2A}$. Thus, since $A, B \geq 0$, the condition $x \geq 2\sqrt{B+2A}$ is sufficient, hence the condition $n \geq 4(B+2A)$. Then we apply this to the following A and B to obtain the condition.

$$A = \frac{\beta \mathcal{F}_\infty(\lambda)}{t}, \;\; B = \frac{\beta \mathcal{F}_\infty(\lambda)}{2t^2}.$$

Point 3) When $\lambda > \|\mathbf{Q}\|$, the result is obtained noting that

$$\|\mathbf{Q}_\lambda^{1/2}\mathbf{Q}_{n,\lambda}^{-1/2}\|^2 \leq \frac{\|\mathbf{Q}\| + \lambda}{\lambda} = 1 + \frac{\|\mathbf{Q}\|}{\lambda} \leq 2.$$

When, on the other hand $0 < \lambda \leq \|\mathbf{Q}\|$, the final result is obtained by applying Prop. 6 and Prop. 8 of (Rudi and Rosasco, 2017), or equivalently applying Eq. (3.50), with $t = 1/2$, for which the following holds with probability $1-\delta$: $\|\mathbf{Q}_\lambda^{-1/2}(\mathbf{Q} - \mathbf{Q}_n)\mathbf{Q}_\lambda^{-1/2}\| \leq t$ and noting that,

$$\|\mathbf{Q}_\lambda^{1/2}\mathbf{Q}_{n,\lambda}^{-1/2}\|^2 \leq \frac{1}{1 - \|\mathbf{Q}_\lambda^{-1/2}(\mathbf{Q} - \mathbf{Q}_n)\mathbf{Q}_\lambda^{-1/2}\|} \leq 2.$$

To conclude this point, we recall that, given two Hermitian operators A, B and $t > 0$, the inequality $A \preceq tB$ is equivalent to $B^{-1/2}AB^{-1/2} \preceq tI$, when B is invertible. Since $B^{-1/2}AB^{-1/2}$ and tI are commutative, then $B^{-1/2}AB^{-1/2} \preceq tI$ is equivalent to $v \cdot (B^{-1/2}AB^{-1/2}v) \leq t\|v\|^2$ for any $v \in \mathcal{H}$, which in turn is equivalent to $\|B^{-1/2}AB^{-1/2}\| \leq t$. So

$$\|A^{1/2}B^{-1/2}\|^2 \leq t \iff A \preceq tB.$$

Point 4) First note that

$$\|\mathbf{Q}_\lambda^{-1/2}\mathbf{Q}_{n,\lambda}^{1/2}\|^2 \leq 1 + \|\mathbf{Q}_\lambda^{-1/2}(\mathbf{Q} - \mathbf{Q}_n)\mathbf{Q}_\lambda^{-1/2}\|. \tag{3.53}$$

When $0 < \lambda \leq \|\mathbf{Q}\|$, by applying Eq. (3.50) with $t = 1/2$, we have with probability $1-\delta$: $\|\mathbf{Q}_\lambda^{-1/2}(\mathbf{Q} - \mathbf{Q}_n)\mathbf{Q}_\lambda^{-1/2}\| \leq t$, moreover by Eq. (3.53) we have

$$\|\mathbf{Q}_\lambda^{-1/2}\mathbf{Q}_{n,\lambda}^{1/2}\|^2 \leq 1 + t \leq 3/2.$$

When instead $\lambda > \|\mathbf{Q}\|$, we consider the following decomposition

$$\|\mathbf{Q}_\lambda^{-1/2}(\mathbf{Q} - \mathbf{Q}_n)\mathbf{Q}_\lambda^{-1/2}\| \leq \frac{1}{\lambda}\|\mathbf{Q} - \mathbf{Q}_n\| \leq \frac{1}{\lambda}\|\mathbf{Q} - \mathbf{Q}_n\|_{HS},$$

where we denote by $\|\cdot\|_{HS}$, the Hilbert-Schmidt norm (i.e. $\|A\|_{HS}^2 = \mathrm{Tr}(A^*A)$) and $\|\mathbf{Q} - \mathbf{Q}_n\|_{HS}$ is well defined since both $\mathbf{Q}, \mathbf{Q}_n$ are trace class. Now since the space of Hilbert-Schmidt operators on a separable Hilbert space is itself a separable Hilbert space and q are bounded almost surely by $c_0 := \operatorname{ess\,sup} \mathrm{Tr}(q)$, we can concentrate $\|\mathbf{Q} - \mathbf{Q}_n\|_{HS}$ via Bernstein inequality for random vectors (e.g. Thm. 3.3.4 of Yurinsky, 1995), obtaining with probability at least $1-\delta$

$$\|\mathbf{Q} - \mathbf{Q}_n\|_{HS} \leq \frac{2c_0 \log\frac{2}{\delta}}{n} + \sqrt{\frac{2c_0^2 \log\frac{2}{\delta}}{n}} \leq \|\mathbf{Q}\|/2,$$

where the last step is due to the fact that we require $n \geq 16c_0^2(\log\frac{2}{\delta})/\|\mathbf{Q}\|^2$, and the fact that by construction $\|\mathbf{Q}\| \leq B$. Then,

$$\|\mathbf{Q}_\lambda^{-1/2}\mathbf{Q}_{n,\lambda}^{1/2}\|^2 \leq 1 + \frac{\|\mathbf{Q}\|}{2\lambda} \leq 3/2.$$

The final result on $\preceq$ is obtained as for Point 5. □

Remark 8. *Let* $\mathrm{Tr}(q) \le c_0$ *almost surely, for a* $c_0 > 0$. *Then* $\mathcal{F}_\infty(\lambda) \le c_0/\lambda$. *So Eq.* (3.49) *is satisfied when*

$$n \ge \frac{8c_0}{\lambda} \log \frac{8c_0}{\lambda\, \delta} \left(\frac{1}{4t^2} + \frac{1}{t} \right),$$

since $\mathcal{F}_\infty(\lambda) \le c_0/\lambda$ *as observed in Remark 7. In particular, when* $t = 1/2$, *Eq.* (3.49) *is satisfied when*

$$n \ge \frac{24c_0}{\lambda} \log \frac{8c_0}{\lambda\, \delta}.$$

3.G .3 Last technical lemmas

Lemma 3.14. *Let* $a_1, a_2, A \ge 0$ *and* $\delta > 0$. *If*

$$n \ge 4a_1^2 A^2 \log^2 \left(\frac{2a_1 a_2 A^2}{\delta} \right),$$

then $n \ge a_1 A n^{1/2} \log \frac{a_2 A n^{1/2}}{\delta}$.

Proof. Indeed, note that

$$n \ge a_1 A n^{1/2} \log \frac{a_2 A n^{1/2}}{\delta} \iff \frac{a_1 A}{n^{1/2}} \log \frac{a_2 A n^{1/2}}{\delta} \le 1.$$

Now use the fact that for $A, B \ge 0$, $k \ge 2A \log(2AB)$ implies $\frac{A}{k} \log(Bk) \le 1$. Indeed, $\log(Bk) = \log(2AB) + \log \frac{Bk}{2AB} = \log(2AB) + \log \frac{k}{2A} \le \log(2AB) + \frac{k}{2A}$. Hence, multiplying by $\frac{A}{k}$, we get the result.

We apply this to $A = a_1 A, B = \frac{a_2 A}{\delta}$ and $k = n^{1/2}$ to get the bound. □

Lemma 3.15. *Let* $a_1, a_2, A \ge 0$ *and* $\delta > 0$. *Let* $p \in [\frac{1}{2}, 1)$. *If*

$$n^{1-p} \ge 2 \frac{1}{1-p} a_1 A \log \left(2a_1 (a_2 \vee 1) \frac{1}{1-p} A^2 \frac{1}{\delta} \right),$$

then

$$n \ge a_1 A n^p \log \frac{a_2 A n^p}{\delta}.$$

Proof. **1)** Let $C_1, C_2 \ge 0$, and $p \in [0, 1)$. Then

$$n \ge C_1 n^p \log(C_2 n^p) \iff \frac{C_1 \frac{p}{1-p}}{n^{1-p}} \log \left(C_2^{(1-p)/p} n^{1-p} \right) \le 1.$$

Now use the fact that for $A, B \ge 0$, $k \ge 2A \log(2AB)$ implies $\frac{A}{k} \log(Bk) \le 1$ (see proof of Lemma 3.14).

Thus, $n^{1-p} \ge 2C_1 \frac{p}{1-p} \log \left(2C_1 \frac{p}{1-p} C_2^{(1-p)/p} \right)$ is a sufficient condition.

2) Now taking $C_1 = a_1 A$ and $C_2 = \frac{a_2 A}{\delta}$, we find that

$$n^{1-p} \ge 2 \frac{p}{1-p} a_1 A \log \left(2a_1 a_2^{(1-p)/p} \frac{p}{1-p} A^{1/p} (\frac{1}{\delta})^{(1-p)/p} \right).$$

Since $0.5 \leq p \leq 1$, we see that $\frac{1-p}{p} \leq 1$ and $\frac{1}{p} \leq 2$ and thus we get our final sufficient condition.

$$n^{1-p} \geq 2\frac{1}{1-p}a_1 A \log\left(2a_1(a_2 \vee 1)\frac{1}{1-p}A^2\frac{1}{\delta}\right).$$

□

Chapter 4

Globally convergent newton methods for ill-conditioned generalized self-concordant losses

This chapter is a verbatim of the work :

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Newton methods for ill-conditioned generalized self-concordant losses. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/60495b4e033e9f60b32a6607b587aadd-Paper.pdf

Contents

4.1 Introduction . . . 138
4.2 Backround . . . 141
4.3 Globally convergent scheme . . . 143
4.4 Application to kernel methods . . . 145
4.5 Experiments . . . 147
4.A Generalized self-concordance . . . 150
4.B Approximate Newton methods . . . 155
4.C Globalization scheme . . . 165
4.D Non-parametric learning . . . 173
4.E Algorithm . . . 187
4.F Experiments . . . 189
4.G Projected problem . . . 193
4.H Expected versus empirical risk . . . 198
4.I Bounds for Hermitian operators . . . 213

4.1 Introduction

Minimization algorithms constitute a crucial algorithmic part of many machine learning methods, with algorithms available for a variety of situations Bottou, Curtis, and Nocedal (2018). In this paper, we focus on *finite sum* problems of the form

$$\min_{x\in\mathcal{H}} f_\lambda(x) = f(x) + \frac{\lambda}{2}\|x\|^2, \text{ with } f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

where $\mathcal{H}$ is a Euclidean or a Hilbert space, and each function is convex and smooth. The running-time of minimization algorithms classically depends on the number of functions n, the explicit (for Euclidean spaces) or implicit (for Hilbert spaces) dimension d of the search space, and the condition number of the problem, which is upper bounded by $\kappa = L/\lambda$, where L characterizes the smoothness of the functions f_i, and λ the regularization parameter.

In the last few years, there has been a strong focus on problems with large n and d, leading to *first-order* (i.e., gradient-based) stochastic algorithms, culminating in a sequence of linearly convergent algorithms whose running time is favorable in n and d, but scale at best in $\sqrt{\kappa}$ Defazio, Bach, and Lacoste-Julien (2014); Lin, Mairal, and Harchaoui (2015); Defazio (2016); Allen-Zhu (2017). However, modern problems lead to objective functions with very large condition numbers, i.e., in many learning problems, the regularization parameter that is optimal for test predictive performance may be so small that the scaling above in $\sqrt{\kappa}$ is not practical anymore (see examples in Sect. 4.5).

These ill-conditioned problems are good candidates for *second-order methods* (i.e., that use the Hessians of the objective functions) such as Newton method. These methods are traditionally discarded within machine learning for several reasons: (1) they are usually adapted to high precision results which are not necessary for generalization to unseen data for machine learning problems Bottou and Bousquet (2008), (2) computing the Newton step $\Delta_\lambda(x) = \nabla^2 f_\lambda(x)^{-1}\nabla f_\lambda(x)$ requires to form the Hessian and solve the associated linear system, leading to complexity which is at least quadratic in d, and thus prohibitive for large d, and (3) the global convergence properties are not applicable, unless the function is very special, i.e., self-concordant Nesterov and Nemirovskii (1994) (which includes only few classical learning problems), so they often are only shown to converge in a small area around the optimal x.

In this paper, we argue that the three reasons above for not using Newton method can be circumvented to obtain competitive algorithms: (1) high absolute precisions are indeed not needed for machine learning, but faced with strongly ill-conditioned problems, even a low-precision solution requires second-order schemes; (2) many approximate Newton steps have been designed for approximating the solution of the associated large linear system A. Erdogdu and Montanari (2015); Roosta-Khorasani and Mahoney (2019); Pilanci and Wainwright (2017); Bollapragada, Byrd, and Nocedal (2018); (3) we propose a novel second-order method which is globally convergent and which is based on performing approximate Newton methods for a certain class of so-called *generalized self-concordant functions* which includes logistic regression Bach (2010). For these functions, the conditioning of the problem is also characterized by a more *local* quantity: $\kappa_\ell = R^2/\lambda$, where R characterizes the local evolution of Hessians. This leads to second-order algorithms which are competitive with first-order algorithms for well-conditioned problems, while being superior for ill-conditioned problems which are common in practice.

Contributions. We make the following contributions:

(*a*) We build a global second-order method for the minimization of f_λ, which relies only on computing approximate Newton steps of the functions $f_\mu, \mu \geq \lambda$. The number of such steps will be of order $O(c \log \kappa_\ell + \log \frac{1}{\epsilon})$ where ϵ is the desired precision, and c is an explicit constant. In the parametric setting ($\mathcal{H} = \mathbb{R}^d$), c can be as bad as $\sqrt{\kappa_\ell}$ in the worst-case but much smaller in theory and practice. Moreover in the non-parametric/kernel machine learning setting ($\mathcal{H}$ infinite dimensional), c does not depend on the local condition number κ_ℓ.

(*b*) Together with the appropriate quadratic solver to compute approximate Newton steps, we obtain an algorithm with the same scaling as regular first-order methods but with an improved behavior, in particular in ill-conditioned problems. Indeed, this algorithm matches the performance of the best quadratic solvers but covers any generalized self-concordant function, up to logarithmic terms.

(*c*) In the non-parametric/kernel machine learning setting we provide an explicit algorithm combining the previous scheme with Nyström projections techniques. We prove that it achieves optimal generalization bounds with $O(n\mathsf{df}_\lambda)$ in time and $O(\mathsf{df}_\lambda^2)$ in memory, where n is the number of observations and df_λ is the associated degrees of freedom. In particular, this is the first large-scale algorithm to solve logistic and softmax regression in the non-parametric setting with large condition numbers and theoretical guarantees.

4.1 .1 Comparison to related work

We consider two cases for $\mathcal{H}$ and the functions f_i that are common in machine learning: $\mathcal{H} = \mathbb{R}^d$ with linear (in the parameter) models with explicit feature maps, and $\mathcal{H}$ infinite-dimensional, corresponding in machine learning to learning with kernels Shawe-Taylor and Cristianini (2004). Moreover in this section we first consider the quadratic case, for example the squared loss in machine learning (i.e., $f_i(x) = \frac{1}{2}(x^\top z_i - y_i)^2$ for some $z_i \in \mathcal{H}, y_i \in \mathbb{R}$). We first need to introduce the Hessian of the problem, for any $\lambda > 0$, define

$$\mathbf{H}(x) := \nabla^2 f(x), \qquad \mathbf{H}_\lambda(x) := \nabla^2 f_\lambda(x) = \mathbf{H}(x) + \lambda \mathbf{I},$$

in particular we denote by $\mathbf{H}$ (and analogously $\mathbf{H}_\lambda$) the Hessian at optimum (which in case of squared loss corresponds to the covariance matrix of the inputs).

Quadratic problems and $\mathcal{H} = \mathbb{R}^d$ (ridge regression). The problem then consists in solving a (ill-conditioned) positive semi-definite symmetric linear system of dimension $d \times d$. Methods based on *randomized linear algebra*, *sketching* and suitable *subsampling* Drineas, Mahoney, Muthukrishnan, and Sarlós (2011); Drineas, Magdon-Ismail, Mahoney, and Woodruff (2012); Boutsidis and Gittens (2013) are able to find the solution with precision ϵ in time that is $O((nd + \min(n, d)^3) \log(L/\lambda\epsilon))$, so essentially independently of the condition number, because of the logarithmic complexity in λ.

Quadratic problems and $\mathcal{H}$ infinite-dimensional (kernel ridge regression). Here the problem corresponds to solving a (ill-conditioned) infinite-dimensional linear system in a reproducing kernel Hilbert space Shawe-Taylor and Cristianini (2004). Since however the sum defining f is finite, the problem can be projected on a subspace of dimension at most n Aronszajn (1950), leading to a linear system of dimension $n \times n$. Solving it with the techniques above would lead to a complexity of the order $O(n^2)$, which is not feasible on massive learning problems (e.g., $n \approx 10^7$). Interestingly these problems are usually approximately low-rank, with the rank

represented by the so called *effective-dimension* df_λ Caponnetto and De Vito (2007), counting essentially the eigenvalues of the problem larger than λ,

$$\mathsf{df}_\lambda = \mathrm{Tr}(\mathbf{H}\mathbf{H}_\lambda^{-1}). \tag{4.1}$$

Note that df_λ is bounded by $\min\{n, L/\lambda\}$ and in many cases $\mathsf{df}_\lambda \ll \min(n, L/\lambda)$. Using suitable projection techniques, like *Nyström* Williams and Seeger (2001) or *random features* Rahimi and Recht (2008) it is possible to further reduce the problem to dimension df_λ, for a total cost to find the solution of $O(n\mathsf{df}_\lambda^2)$. Finally recent methods Rudi, Carratino, and Rosasco (2017), combining suitable projection methods with refined preconditioning techniques, are able to find the solution with precision compatible with the optimal statistical learning error Caponnetto and De Vito (2007) in time that is $O(n\mathsf{df}_\lambda \log(L/\lambda))$, so being essentially independent of the condition number of the problem.

Convex problems and explicit features (logistic regression). When the loss function is *self-concordant* it is possible to leverage the fast techniques for linear systems in approximate Newton algorithms Pilanci and Wainwright (2017) (see more in Sec. 4.2), to achieve the solution in essentially $O(nd + \min(n, d)^3)$ time, modulo logarithmic terms. However only few loss functions of interest are self-concordant, in particular the widely used logistic and soft-max losses are not self-concordant, but *generalized-self-concordant* Bach (2010). In such cases we need to use (accelerated/stochastic) first order optimization methods to enter in the quadratic convergence region of Newton methods Agarwal, Bullins, and Hazan (2017), which leads to a solution in $O(dn + d\sqrt{nL/\lambda} + \min(n, d)^3)$ time, which does not present any improvement on a simple accelerated first-order method. Globally convergent second-order methods have also been proposed to solve such problems Karimireddy, Stich, and Jaggi (2018), but the number of Newton steps needed being bounded only by L/λ, they lead to a solution in $O(L/\lambda\ (nd + \min(n, d)^3))$. With λ that could be as small as 10^{-12} in modern machine learning problems, this makes both these kind of approaches expensive from a computational viewpoint for ill-conditioned problems. For such problems, with our new global second-order scheme, the algorithm we propose achieves instead a complexity of essentially $O((nd + \min(n, d)^3)\log(R^2/\lambda\epsilon))$ (see Theorem 4.1).

Convex problems and $\mathcal{H}$ infinite-dimensional (kernel logistic regression). Analogously to the case above, it is not possible to use Newton methods profitably as global optimizers on losses that are not self-concordant as we see in Sec. 4.3 . In such cases by combining projecting techniques developped in Sec. 4.4 and accelerated first-order optimization methods, it is possible to find a solution in $O(n\mathsf{df}_\lambda + \mathsf{df}_\lambda\sqrt{nL/\lambda})$ time. This can still be prohibitive in the very small regularization scenario, since it strongly depends on the condition number L/λ. In Sec. 4.4 we suitably combine our optimization algorithm with projection techniques achieving optimal statistical learning error Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) in essentially $O(n\mathsf{df}_\lambda \log(R^2/\lambda))$.

First-order algorithms for finite sums. In dimension d, accelerated algorithms for strongly-convex smooth (not necessarily self-concordant) finite sums, such as K-SVRG Allen-Zhu (2017), have a running time proportional $O((n + \sqrt{nL/\lambda})d)$. This can be improved with preconditioning to $O((n + \sqrt{dL/\lambda})d)$ for large n Agarwal, Bullins, and Hazan (2017). Quasi-Newton methods can also be used Gower, Hanzely, Richtárik, and Stich (2018), but typically without the guarantees that we provide in this paper (which are logarithmic in the condition number in natural scenarios).

4.2 Background: Newton methods and generalized self concordance

In this section we start by recalling the definition of generalized self concordant functions and motivate it with examples. We then recall basic facts about Newton and approximate Newton methods, and present existing techniques to efficiently compute approximate Newton steps. We start by introducing the definition of generalized self-concordance, that here is an extension of the one in Bach (2010).

Definition 4.1 (generalized self-concordant (GSC) function)**.** *Let $\mathcal{H}$ be a Hilbert space. We say that f is a generalized self-concordant function on $\mathcal{G} \subset \mathcal{H}$, when $\mathcal{G}$ is a bounded subset of $\mathcal{H}$ and f is a convex and three times differentiable mapping on $\mathcal{H}$ such that*

$$\forall x \in \mathcal{H},\ \forall h, k \in \mathcal{H},\ \nabla^{(3)} f(x)[h,k,k] \leq \sup_{g \in \mathcal{G}} |g \cdot h|\ \nabla^2 f(x)[k,k].$$

We will usually denote by R the quantity $\sup_{g \in \mathcal{G}} \|g\| < \infty$ and often omit $\mathcal{G}$ when it is clear from the context (for simplicity think of $\mathcal{G}$ as the ball in $\mathcal{H}$ centered in zero and with radius $R > 0$, then $\sup_{g \in \mathcal{G}} |g \cdot h| = R\|h\|$). The globally convergent second-order scheme we present in Sec. 4.3 is specific to losses which satisfy this generalized self-concordance property. The following loss functions, which are widely used in machine learning, are generalized-self-concordant, and motivate this work.

Example 4.1 (Application to finite-sum minimization)**.** *The following loss functions are generalized self-concordant functions, but not self-concordant:*
(a) Logistic regression: $f_i(x) = \log(1 + \exp(-y_i w_i^\top x))$, where $x, w_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$.
(b) Softmax regression: $f_i(x) = \log\left(\sum_{j=1}^k \exp(x_j^\top w_i)\right) - x_{y_i}^\top w_i$, where now $x \in \mathbb{R}^{d \times k}$ and $y_i \in \{1, \dots, k\}$ and x_j denotes the j-th column of x.
(c) Generalized linear models with bounded features (see details in Bach (2014, Sec. 2.1)), which include conditional random fields Sutton and McCallum (2012).
(d) Robust regression: $f_i(x) = \varphi(y_i - w_i^\top x)$ with $\varphi(u) = \log(e^u + e^{-u})$.

Note that these losses are not *self-concordant* in the sense of Pilanci and Wainwright (2017). Moreover, even if the losses f_i are self-concordant, the objective function f is not necessarily self-concordant, making any attempt to prove the self-concordance of the objective function f almost impossible.

Newton method (NM). Given $x_0 \in \mathcal{H}$, the Newton method consists in doing the following update:

$$x_{t+1} = x_t - \Delta_\lambda(x_t), \qquad \Delta_\lambda(x_t) := \mathbf{H}_\lambda^{-1}(x_t) \nabla f_\lambda(x_t). \tag{4.2}$$

The quantity $\Delta_\lambda(x) := \mathbf{H}_\lambda^{-1}(x) \nabla f_\lambda(x)$ is called the Newton step at point x, and $x - \Delta_\lambda(x)$ is the minimizer of the second order approximation of f_λ around x. Newton methods enjoy the following key property: if x_0 is close enough to the optimum, the convergence to the optimum is quadratic and the number of iterations required to a given precision is independent of the condition number of the problem Boyd and Vandenberghe (2004).

However Newton methods have two main limitations: (a) the region of quadratic convergence can be quite small and reaching the region can be computationally expensive, since it is usually done via first order methods Agarwal, Bullins, and Hazan (2017) that converge linearly depending on the condition number of the problem, (b) the cost of computing the Hessian can be really

expensive when n, d are large, and also (c) the cost of computing $\Delta_\lambda(x_t)$ can be really prohibitive. In the rest of the section we recall some ways to deal with (b) and (c). Our main result of Sec. 4.3 is to provide globalization scheme for the Newton method to tackle problem (a), which is easily integrable with approximate techniques to deal with (b) ans (c), to make second-order technique competitive.

Approximate Newton methods (ANM) and approximate solutions to linear systems. Computing exactly the Newton increment $\Delta_\lambda(x_t)$, which corresponds essentially to the solution of a linear system, can be too expensive when n, d are large. A natural idea is to approximate the Newton iteration, leading to *approximate Newton methods*,

$$x_{t+1} = x_t - \widetilde{\Delta}_\lambda(x_t), \qquad \widetilde{\Delta}_\lambda \approx \Delta_\lambda(x_t). \tag{4.3}$$

In this paper, more generally we consider any technique to compute $\widetilde{\Delta}_\lambda(x_t)$ that provides a *relative approximation* Deuflhard (2011) of $\Delta_\lambda(x_t)$ defined as follows.

Definition 4.2 (relative approximation)**.** *Let $\rho < 1$, let $\mathbf{A}$ be an invertible positive definite Hermitian operator on $\mathcal{H}$ and b in $\mathcal{H}$. We denote by* $\mathrm{LinApprox}(\mathbf{A}, b, \rho)$ *the set of all ρ-relative approximations of $z^* = \mathbf{A}^{-1}b$, i.e.,* $\mathrm{LinApprox}(\mathbf{A}, b, \rho) = \{z \in \mathcal{H} \mid \|z - z^*\|_{\mathbf{A}} \leq \rho \|z^*\|_{\mathbf{A}}\}$.

Sketching and subsampling for approximate Newton methods. Many techniques for approximating linear systems have been used to compute $\widetilde{\Delta}_\lambda$, in particular *sketching* of the Hessian matrix via fast transforms and *subsampling* (see Pilanci and Wainwright (2017); Bollapragada, Byrd, and Nocedal (2018); Agarwal, Bullins, and Hazan (2017) and references therein). Assuming for simplicity that $f_i = \ell_i(w_i^\top x)$, with $\ell_i : \mathbb{R} \to \mathbb{R}$ and $w_i \in \mathcal{H}$, it holds:

$$\mathbf{H}(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i^{(2)}(w_i^\top x) w_i w_i^\top = V_x^\top V_x, \tag{4.4}$$

with $V_x \in \mathbb{R}^{n\times d} = D_x W$, where $D_x \in \mathbb{R}^{n\times n}$ is a diagonal matrix defined as $(D_x)_{ii} = (\ell_i^{(2)}(w_i^\top x))^{1/2}$ and $W \in \mathbb{R}^{n\times d}$ defined as $W = (w_1, \dots, w_n)^\top$.

Both sketching and subsampling methods approximate $z^* = \mathbf{H}_\lambda(x)^{-1}\nabla f_\lambda(x)$ with $\tilde{z} = \widetilde{\mathbf{H}}_\lambda(x)^{-1}\nabla f_\lambda(x)$, in particular, in the case of subsampling $\widetilde{\mathbf{H}}(x) = \sum_{j=1}^{Q} p_j w_{i_j} w_{i_j}^\top$ where $Q \ll \min(n, d)$, $(p_j)_{j=1}^n$ are suitable weights and $(i_j)_{j=1}^Q$ are indices selected at random from $\{1, \dots, n\}$ with suitable probabilities. Sketching methods instead use $\widetilde{\mathbf{H}}(x) = \widetilde{V}_x^\top \widetilde{V}_x$, with $\widetilde{V}_x = \Omega V_x$ with $\Omega \in \mathbb{R}^{Q\times n}$ a structured matrix such that computing $\widetilde{V}_x$ has a cost in the order of $O(nd \log n)$; to this end usually Ω is based on fast Fourier or Hadamard transforms Pilanci and Wainwright (2017). Note that essentially all the techniques used in approximate Newton methods guarantee relative approximation. In particular the following results can be found in the literature (see Lemmas 4.28 and 4.29 in Sec. 4.I and Pilanci and Wainwright (2017), Lemma 2 for more details).

Lemma 4.1. *Let $x, b \in \mathcal{H}$ and assume that $\ell_i^{(2)} \leq a$ for $a > 0$. With probability $1-\delta$ the following methods output an element in* $\mathrm{LinApprox}(\mathbf{H}_\lambda(x), b, \rho)$*, in $O(Q^2 d + Q^3 + c)$ time, $O(Q^2 + d)$ space:*
(a) Subsampling with uniform sampling (see Roosta-Khorasani and Mahoney (2019); Rudi, Camoriano, and Rosasco (2015)), where $Q = O(\rho^{-2} a/\lambda \log \frac{1}{\lambda\delta})$ and $c = O(1)$.
(b) Subsampling with approximate leverage scores Roosta-Khorasani and Mahoney (2019); Alaoui and Mahoney (2015); Rudi, Camoriano, and Rosasco (2015)), where $Q = O(\rho^{-2}\bar{\mathsf{df}}_\lambda \log 1/\lambda\delta), c = O(\min(n, a/\lambda)\bar{\mathsf{df}}_\lambda^{\,2})$ and $\bar{\mathsf{df}}_\lambda = \mathrm{Tr}(W^\top W(W^\top W + \lambda/aI)^{-1})$ Rudi, Calandriello, Carratino, and Rosasco (2018). Note that $\bar{\mathsf{df}}_\lambda \leq \min(n, d)$.

(c) Sketching with fast Hadamard transform *Pilanci and Wainwright (2017), where* $Q = O(\rho^{-2}\mathsf{df}_\lambda \log a/\lambda\delta), c = O(nd \log n)$.

4.3 Globally convergent scheme for ANM algorithms on GSC functions

The algorithm is based on the observation that when f_λ is generalized self concordant, there exists a region where t steps of ANM converge as fast as 2^{-t}. Our idea is to start from a very large regularization parameter λ_0, such that we are sure that x_0 is in the convergence region and perform some steps of ANM such that the solution enters in the convergence region of f_{λ_1}, with $\lambda_1 = q\lambda_0$ with $q < 1$, and to iterate this procedure until we enter the convergence region of f_λ. First we define the region of interest and characterize the behavior of NM and ANM in the region, then we analyze the globalization scheme.

Preliminary results: the Dikin ellipsoid. We consider the following region that we prove to be contained in the region of quadratic convergence for the Newton method and that will be useful to build the globalization scheme. Let $c, R > 0$ and f_λ be generalized self-concordant with coefficient R, we call *Dikin ellipsoid* and denote by $\mathsf{D}_\lambda(\mathsf{c})$ the region

$$\mathsf{D}_\lambda(\mathsf{c}) := \{x \mid \nu_\lambda(x) \leq \mathsf{c}\sqrt{\lambda}/R\}, \quad \text{with} \quad \nu_\lambda(x) := \|\nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)},$$

where $\nu_\lambda(x)$ is usually called the *Newton decrement* and $\|x\|_{\mathbf{A}}$ stands for $\|\mathbf{A}^{1/2}x\|$.

Lemma 4.2. *Let* $\lambda > 0, \mathsf{c} \leq 1/7$, *let* f_λ *be generalized self-concordant and* $x \in \mathsf{D}_\lambda(\mathsf{c})$. *Then it holds:* $\frac{1}{4}\nu_\lambda(x)^2 \leq f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \nu_\lambda(x)^2$. *Moreover Newton method starting from* x_0 *has quadratic convergence, i.e., let* x_t *be obtained via* $t \in \mathbb{N}$ *steps of Newton method in Eq. (4.2), then* $\nu_\lambda(x_t) \leq 2^{-(2^t-1)}\nu_\lambda(x_0)$. *Finally, approximate Newton methods starting from* x_0 *have a linear convergence rate, i.e., let* x_t *given by Eq. (4.3), with* $\widetilde{\Delta}_t \in \mathrm{LinApprox}(\mathbf{H}_\lambda(x_t), \nabla f_\lambda(x_t), \rho)$ *and* $\rho \leq 1/7$, *then* $\nu_\lambda(x_t) \leq 2^{-t}\nu_\lambda(x_0)$.

This result is proved in Lemma 4.11 in Sec. 4.B .3. The crucial aspect of the result above is that when $x_0 \in \mathsf{D}_\lambda(\mathsf{c})$, the convergence of the approximate Newton method is linear and does not depend on the condition number of the problem. However $\mathsf{D}_\lambda(\mathsf{c})$ itself can be very small depending on $\sqrt{\lambda}/R$. In the next subsection we see how to enter in $\mathsf{D}_\lambda(\mathsf{c})$ in an efficient way.

Entering the Dikin ellipsoid using a second-order scheme. The lemma above shows that $\mathsf{D}_\lambda(\mathsf{c})$ is a good region where to use the approximate Newton algorithm on GSC functions. However the region itself is quite small, since it depends on $\sqrt{\lambda}/R$. Some other globalization schemes arrive to regions of interest by first-order methods or back-tracking schemes Agarwal, Bullins, and Hazan (2017); A. Erdogdu and Montanari (2015). However such approaches require a number of steps that is usually proportional to $\sqrt{L/\lambda}$ making them non-beneficial in machine learning contexts. Here instead we consider the following simple scheme where $\mathtt{ANM}_\rho(f_\lambda, x, t)$ is the result of a ρ-relative approximate Newton method performing t steps of optimization starting from x.

The main ingredient to guarantee the scheme to work is the following lemma (see Lemma 4.13 in Sec. 4.C .1 for a proof).

Lemma 4.3. *Let* $\mu > 0$, $\mathsf{c} < 1$ *and* $x \in \mathcal{H}$. *Let* $s = 1 + R\|x\|/\mathsf{c}$, *then for* $q \in [1 - 2/(3s), 1)$

$$\mathsf{D}_\mu(\mathsf{c}/3) \subseteq \mathsf{D}_{q\mu}(\mathsf{c}).$$

Now we are ready to show that we can guarantee the loop invariant $x_k \in \mathsf{D}_{\mu_k}(\mathsf{c})$. Indeed assume that $x_{k-1} \in \mathsf{D}_{\mu_{k-1}}(\mathsf{c})$. Then $\nu_{\mu_{k-1}}(x_{k-1}) \leq \mathsf{c}\sqrt{\mu_{k-1}}/R$. By taking $t = 2, \rho = 1/7$, and performing $x_k = \texttt{ANM}_\rho(f_{\mu_{k-1}}, x_{k-1}, t)$, by Lemma 4.2, $\nu_{\mu_{k-1}}(x_k) \leq 1/4\nu_{\mu_{k-1}}(x_{k-1}) \leq \mathsf{c}/4\ \sqrt{\mu_{k-1}}/R$, i.e., $x_k \in \mathsf{D}_{\mu_{k-1}}(\mathsf{c}/4)$. If q_k is large enough, this implies that $x_k \in \mathsf{D}_{q_k\mu_{k-1}}(\mathsf{c}) = \mathsf{D}_{\mu_k}(\mathsf{c})$, by Lemma 4.3. Now we are ready to state our main theorem of this section.

Proposed Globalization Scheme

Phase I: Getting in the Dikin ellispoid of f_λ

Start with $x_0 \in \mathcal{H}, \mu_0 > 0$, $t, T \in \mathbb{N}$ and $(q_k)_{k\in\mathbb{N}} \in (0, 1]$.

For $k \in \mathbb{N}$

$x_{k+1} \leftarrow \texttt{ANM}_\rho(f_{\mu_k}, x_k, t)$

$\mu_{k+1} \leftarrow q_{k+1}\mu_k$

Stop when $\mu_{k+1} < \lambda$ and set $x_{last} \leftarrow x_k$.

Phase II: reach a certain precision starting from inside the Dikin ellipsoid

Return $\widehat{x} \leftarrow \texttt{ANM}_\rho(f_\lambda, x_{last}, T)$

Fully adaptive method. The scheme presented above converges with the following parameters.

Theorem 4.1. *Let $\epsilon > 0$. Set $\mu_0 = 7R\|\nabla f(0)\|$, $x_0 = 0$, and perform the globalization scheme above for $\rho \leq 1/7, t = 2$, and $q_k = \frac{1/3+7R\|x_k\|}{1+7R\|x_k\|}$, $T = \lceil \log_2 \sqrt{1 \vee (\lambda\epsilon^{-1}/R^2)} \rceil$. Then denoting by K the number of steps performed in the Phase I, it holds:*

$$f_\lambda(\widehat{x}) - f_\lambda(x_\lambda^\star) \leq \epsilon, \qquad K \leq \lfloor (3 + 11R\|x_\lambda^\star\|) \log(7R\|\nabla f(0)\|/\lambda) \rfloor .$$

Note that the theorem above (proven in Sec. 4.C .3) guarantees a solution with error ϵ with K steps of ANM each performing 2 iterations of approximate linear system solving, plus a final step of ANM which performs T iterations of approximate linear system solving. In case of $f_i(x) = \ell_i(w_i^\top x)$, with $\ell_i : \mathbb{R} \to \mathbb{R}$, $w_i \in \mathcal{H}$ with $\ell_i^{(2)} \leq a$, for $a > 0$, the final runtime cost of the proposed scheme to achieve precision ϵ, when combined with of the methods for approximate linear system solving from Lemma 4.1 (i.e. sketching), is $O(Q^2 + d)$ in memory and

$$O\Big((nd \log n + dQ^2 + Q^3)\Big(R\|x_\lambda^\star\| \log \frac{R}{\lambda} + \log \frac{\lambda}{R\epsilon}\Big)\Big) \text{ in time}, \quad Q = O\Big(\bar{\mathsf{df}}_\lambda \log \frac{1}{\lambda\delta}\Big),$$

where $\bar{\mathsf{df}}_\lambda$, defined in Lemma 4.1, measures the *effective dimension* of the correlation matrix $W^\top W$ with $W = (w_1, \ldots, w_n)^\top \in \mathbb{R}^{n\times d}$, corresponding essentially to the number of eigenvalues of $W^\top W$ larger than λ/a. In particular note that $\bar{\mathsf{df}}_\lambda \leq \min(n, d, \text{rank}(W), ab^2/\lambda)$, with $b := \max_i \|w_i\|$, and usually way smaller than such quantities.

Remark 9. *The proposed method does not depend on the condition number of the problem L/λ, but on the term $R\|x_\lambda^\star\|$ which can be in the order of $R/\sqrt{\lambda}$ in the worst case, but usually way smaller. For example, it is possible to prove that this term is bounded by an absolute constant not depending on λ, if at least one minimum for f exists. In the appendix (see proposition 4.7), we show a variant of this adaptive method which can leverage the regularity of the solution with respect to the Hessian, i.e., depending on the smaller quantity $R\sqrt{\lambda}\|x_\lambda^\star\|_{\mathbf{H}_\lambda^{-1}(x_\lambda^\star)}$ instead of $R\|x_\lambda^\star\|$.*

Finally note that it is possible to use $q_k = q$ fixed for all the iterations and way smaller than the one in Theorem 4.1, depending on some regularity properties of $\mathbf{H}$ (see proposition 4.8 in Sec. 4.C .2).

4.4 Application to the non-parametric setting: Kernel methods

In supervised learning the goal is to predict well on future data, given the observed training dataset. Let $\mathcal{X}$ be the input space and $\mathcal{Y} \subseteq \mathbb{R}^p$ be the output space. We consider a probability distribution P over $\mathcal{X} \times \mathcal{Y}$ generating the data and the goal is to estimate $g^* : \mathcal{X} \to \mathcal{Y}$ solving the problem

$$g^* = \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathcal{Y}} \mathcal{L}(g), \quad \mathcal{L}(g) = \mathbb{E}[\ell(g(x), y)], \tag{4.5}$$

for a given loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Note that P is not known, and accessible only via the dataset $(x_i, y_i)_{i=1}^n$, with $n \in \mathbb{N}$, independently sampled from P. A prototypical estimator for g^* is the regularized minimizer of the empirical risk $\widehat{\mathcal{L}}(g) = \frac{1}{n}\sum_{i=1}^n \ell(g(x_i), y_i)$ over a suitable space of functions $\mathcal{G}$. Given $\phi : \mathcal{X} \to \mathcal{H}$ a common choice is to select $\mathcal{G}$ as the set of linear functions of $\phi(x)$, that is, $\mathcal{G} = \{w^\top \phi(\cdot) \mid w \in \mathcal{H}\}$. Then the regularized minimizer of $\widehat{\mathcal{L}}$, denoted by $\widehat{g}_\lambda$, corresponds to

$$\widehat{g}_\lambda(x) = \widehat{w}_\lambda^\top \phi(x), \quad \widehat{w}_\lambda = \operatorname*{arg\,min}_{w\in\mathcal{H}} \tfrac{1}{n}\textstyle\sum_{i=1}^n f_i(w) + \lambda\|w\|^2, \quad f_i(w) = \ell(w^\top\phi(x_i), y_i). \tag{4.6}$$

Learning theory guarantees how fast $\widehat{g}_\lambda$ converges to the best possible estimator g^* with respect to the number of observed examples, in terms of the so called *excess risk* $\mathcal{L}(\widehat{g}_\lambda) - \mathcal{L}(g^*)$. The following theorem recovers the minimax optimal learning rates for squared loss and extend them to any generalized self-concordant loss function.

Note on df_λ. In this section, we always denote with df_λ the effective dimension of the problem in Eq. (4.5). When the loss belongs to the family of generalized linear models (see Example 4.1) and if the model is well-specified, then df_λ is defined exactly as in Eq. (4.1) otherwise we need a more refined definition (see Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) or Eq. (4.30) in Sec. 4.D).

Theorem 4.2 (from Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), Thm. 4)**.** *Let $\lambda > 0, \delta \in (0, 1]$. Let ℓ be generalized self-concordant with parameter $R > 0$ and $\sup_{x\in X} \|\phi(x)\| \le C < \infty$. Assume that there exists g^* minimizing $\mathcal{L}$. Then there exists c_0 not depending on $n, \lambda, \delta, \mathsf{df}_\lambda, C, g^*$, such that if $\sqrt{\mathsf{df}_\lambda/n}, \mathsf{b}_\lambda \le \lambda^{1/2}/R$, and $n \ge C/\lambda \log(\delta^{-1}C/\lambda)$ the following holds with probability $1 - \delta$:*

$$\mathcal{L}(\widehat{g}_\lambda) - \mathcal{L}(g^*) \le c_0\Big(\frac{\mathsf{df}_\lambda}{n} + \mathsf{b}_\lambda^2\Big)\log(1/\delta), \qquad \mathsf{b}_\lambda := \lambda\|g^*\|_{\mathbf{H}_\lambda^{-1}(g^*)}. \tag{4.7}$$

Under standard regularity assumptions of the learning problems Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), i.e., (a) the *capacity condition* $\sigma_j(\mathbf{H}(g^*)) \le Cj^{-\alpha}$, for $\alpha \ge 1, C > 0$ (i.e., a decay of eigenvalues $\sigma_j(\mathbf{H}(g^*))$ of the Hessian at the optimum), and (b) the *source condition* $g^* = \mathbf{H}(g^*)^r v$, with $v \in \mathcal{H}$ and $r > 0$ (i.e., the control of the optimal g^* for a specific Hessian-dependent norm), $\mathsf{df}_\lambda \le C'\lambda^{-1/\alpha}$ and $\mathsf{b}_\lambda^2 \le C''\lambda^{1+2r}$, leading to the following optimal learning rate,

$$\mathcal{L}(\widehat{g}_\lambda) - \mathcal{L}(g^*) \le c_1 n^{-\frac{1+2r\alpha}{1+\alpha+2r\alpha}} \log(1/\delta), \quad \text{when} \quad \lambda = n^{-\frac{\alpha}{1+\alpha+2r\alpha}}. \tag{4.8}$$

Now we propose an algorithmic scheme to compute efficiently an approximation of $\widehat{g}_\lambda$ that achieves the same optimal learning rates. First we need to introduce the technique we are going to use.

Nyström projection. It consists in suitably selecting $\{\bar{x}_1, \dots, \bar{x}_M\} \subset \{x_1, \dots, x_n\}$, with $M \ll n$ and computing $\bar{g}_{M,\lambda}$, i.e., the solution of Eq. (4.6) over $\mathcal{H}_M = \text{span}\{\phi(\bar{x}_1), \dots, \phi(\bar{x}_M)\}$ instead of $\mathcal{H}$. In this case the problem can be reformulated as a problem in $\mathbb{R}^M$ as

$$\bar{g}_{M,\lambda} = \bar{\alpha}_{M,\lambda}^\top \mathbf{T}^{-1} v(x), \qquad \bar{\alpha}_{M,\lambda} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^M} \bar{f}_\lambda(\alpha), \qquad \bar{f}(\alpha) = \frac{1}{n} \sum_{i=1}^n \bar{f}_i(\alpha) + \lambda \|\alpha\|^2, \tag{4.9}$$

where $\bar{f}_i(\alpha) = \ell(v(x_i)^\top \mathbf{T}^{-1} \alpha,\ y_i)$ and $v(x) \in \mathbb{R}^M$, $v(x) = (k(x, \bar{x}_1), \dots, k(x, \bar{x}_M))$ with $k(x, x') = \phi(x)^\top \phi(x')$ the associated positive-definite kernel Shawe-Taylor and Cristianini (2004), while $\mathbf{T}$ is the upper triangular matrix such that $\mathbf{K} = \mathbf{T}^\top \mathbf{T}$, with $\mathbf{K} \in \mathbb{R}^{M \times M}$ with $\mathbf{K}_{ij} = k(\bar{x}_i, \bar{x}_j)$. In the next theorem we characterize the sufficient M to achieve minimax optimal rates, for two standard techniques of choosing the Nyström points $\{\bar{x}_1, \dots, \bar{x}_M\}$.

Theorem 4.3 (Optimal rates for learning with Nyström)**.** *Let $\lambda > 0, \delta \in (0, 1]$. Assume the conditions of Theorem 4.2. Then the excess risk of $\bar{g}_{M,\lambda}$ is bounded with prob. $1 - 2\delta$ as in Eq. (4.7) (with $c_1' \propto c_1$), when*

(1) *Uniform Nyström method Rudi, Camoriano, and Rosasco (2015); Rudi, Carratino, and Rosasco (2017) is used and $M \geq C_1/\lambda \log(C_2/\lambda\delta)$.*

(2) *Approximate leverage score method Alaoui and Mahoney (2015); Rudi, Camoriano, and Rosasco (2015); Rudi, Carratino, and Rosasco (2017) is used and $M \geq C_3\, \mathsf{df}_\lambda \log(C_4/\lambda\delta)$.*

Here C, C_1, C_2, C_4 do not depend on $\lambda, n, M, \mathsf{df}_\lambda, \delta$.

Theorem 4.3 generalizes results for learning with Nyström and squared loss Rudi, Camoriano, and Rosasco (2015), to GSC losses. It is proved in Theorem 4.6, in Sec. 4.D .4. As in Rudi, Camoriano, and Rosasco (2015), Theorem 4.3 shows that Nyström is a valid technique for dimensionality reduction. Indeed it is essentially possible to project the learning problem on a subspace $\mathcal{H}_M$ of dimension $M = O(c/\lambda)$ or even as small as $M = O(\mathsf{df}_\lambda)$ and still achieve the optimal rates of Theorem 4.2. Now we are ready to introduce our algorithm.

Proposed algorithm. The algorithm conceptually consists in (a) performing a projection step with Nyström, and (b) solving the resulting optimization problem with the globalization scheme proposed in Sec. 4.3 based on ANM in Eq. (4.3). In particular, we want to avoid to apply explicitly $\mathbf{T}^{-1}$ to each $v(x_i)$ in Eq. (4.9), which would require $O(nM^2)$ time. Then we will use the following approximation technique based only on matrix vector products, so we can just apply $\mathbf{T}^{-1}$ to α at each iteration, with a total cost proportional only to $O(nM + M^2)$ per iteration. Given $\alpha, \nabla \bar{f}_\lambda(\alpha)$, we approximate $z^* = \bar{\mathbf{H}}_\lambda(\alpha)^{-1} \nabla \bar{f}_\lambda(\alpha)$, where $\bar{\mathbf{H}}_\lambda$ is the Hessian of $\bar{f}_\lambda(\alpha)$, with $\tilde{z}$ defined as

$$\tilde{z} = \texttt{prec-conj-grad}_t(\bar{\mathbf{H}}_\lambda(\alpha), \nabla \bar{f}_\lambda(\alpha)),$$

where $\texttt{prec-conj-grad}_t$ corresponds to performing t steps of preconditioned conjugate gradient Golub and Van Loan (2012) with preconditioner computed using a subsampling approach for the Hessian among the ones presented in Sec. 4.2 , in the paragraph starting with Eq. (4.4). The pseudocode for the whole procedure is presented in Alg. 1, Sec. 4.E . This technique of approximate linear system solving has been studied in Rudi, Carratino, and Rosasco (2017) in the context of empirical risk minimization for squared loss.

Lemma 4.4 (Rudi, Carratino, and Rosasco (2017))**.** *Let $\lambda > 0, \alpha, b \in \mathbb{R}^M$. The previous method, applied with $t = O(\log 1/\rho)$, outputs an element of $\text{LinApprox}(\bar{\mathbf{H}}_\lambda(\alpha), b, \rho)$, with probability $1 - \delta$ with complexity $O((nM + M^2 Q + M^3 + c)t)$ in time and $O(M^2 + n)$ in space, with $Q = O(C_1/\lambda \log(C_1/\lambda\delta)), c = O(1)$ if uniform sub-sampling is used or $Q = O(C_2 \mathsf{df}_\lambda \log(C_1/\lambda\delta)), c =$*

$O(\mathsf{df}_\lambda^2 \min(n, \frac{1}{\lambda}))$ *if sub-sampling with leverage scores is used Rudi, Calandriello, Carratino, and Rosasco (2018).*

A more complete version of this lemma is shown in proposition 4.12 in Sec. 4.D .5. We conclude this section with a result proving the learning properties of the proposed algorithm.

Theorem 4.4 (Optimal rates for the proposed algorithms)**.** *Let $\lambda > 0$ and $\epsilon < \lambda/R^2$. Under the hypotheses of Theorem 4.3, if we set M as in Theorem 4.3, Q as in Lemma 4.4 and setting the globalization scheme as in Theorem 4.1, then the proposed algorithm (Alg. 1, Sec. 4.E) finishes in a finite number of newton steps $N_{ns} = O(R\|g^*\| \log(C/\lambda) + \log(C/\epsilon))$ and returns a predictor $g_{Q,M,\lambda}$ of the form $g_{Q,M,\lambda} = \alpha^\top \mathbf{T}^{-1} v(x)$. With probability at least $1-\delta$, this predictor satisfies:*

$$\mathcal{L}(g_{Q,M,\lambda}) - \mathcal{L}(g^*) \leq c_0 \Big(\frac{\mathsf{df}_\lambda}{n} + \mathsf{b}_\lambda^2 + \epsilon \Big) \log(1/\delta), \qquad \mathsf{b}_\lambda := \lambda \|g^*\|_{\mathbf{H}_\lambda^{-1}(g^*)}. \tag{4.10}$$

The theorem above (see proposition 4.14, Sec. 4.D .6 for exacts quantifications) shows that the proposed algorithm is able to achieve the same learning rates of plain empirical risk minimization as in Theorem 4.2. The total complexity of the procedure, including the cost of computing the preconditioner, the selection of the Nyström points via approximate leverage scores and also the computation of the leverage scores Rudi, Calandriello, Carratino, and Rosasco (2018) is then

$$O\left(R\|g^*\| \log(R^2/\lambda) \left(n\ \mathsf{df}_\lambda \log(C\lambda^{-1}\delta^{-1})\ c_X + \ + \mathsf{df}_\lambda^3 \log^3(C\lambda^{-1}\delta^{-1}) + \min(n, C/\lambda)\ \mathsf{df}_\lambda^2\right)\right)$$

in time and $O(\mathsf{df}_\lambda^2 \log^2(C\lambda^{-1}\delta^{-1}))$ in space, where c_X is the cost of computing the inner product $k(x, x')$ (in the kernel setting assumed when the input space X is $X = \mathbb{R}^p$ it is $c = O(p)$). As noted in Rudi, Calandriello, Carratino, and Rosasco (2018), under the standard regularity assumptions on the learning problem seen above, $\mathsf{df}_\lambda^2 \leq \mathsf{df}_\lambda/\lambda \leq n$ when the optimal λ is chosen. So the total computational complexity is

$$O\left(R \log(R^2/\lambda)\ \log^3(C\lambda^{-1}\delta^{-1})\ \|g^*\| \cdot n \cdot \mathsf{df}_\lambda \cdot c_X\right) \text{ in time}, \quad O(\mathsf{df}_\lambda^2 \cdot \log^2(C\lambda^{-1}\delta^{-1})) \text{ in space}.$$

First note, the fact that due to the statistical properties of the problem the complexity does not depend even implicitly on $\sqrt{C/\lambda}$, but only on $\log(C/\lambda)$, so the algorithm runs in essentially $O(n\mathsf{df}_\lambda)$, compared to $O(\mathsf{df}_\lambda \sqrt{nC/\lambda})$ of the accelerated first-order methods we develop in Sec. 4.F and the $O(n\mathsf{df}_\lambda \sqrt{C/\lambda})$ of other Newton schemes (see Sec. 4.1 .1). To our knowledge, this is the first algorithm to achieve optimal statistical learning rates for generalized self-concordant losses and with complexity only $\widetilde{O}(n\mathsf{df}_\lambda)$. This generalizes similar results for squared loss Rudi, Carratino, and Rosasco (2017); Rudi, Calandriello, Carratino, and Rosasco (2018).

4.5 Experiments

The code necessary to reproduce the following experiments is available on GitHub at `https://github.com/umarteau/Newton-Method-for-GSC-losses-`.

We compared the performances of our algorithm for kernel logistic regression on two large scale classification datasets ($n \approx 10^7$), Higgs and Susy, pre-processed as in Rudi, Carratino, and Rosasco (2017). We implemented the algorithm in pytorch and performed the computations on 1 Tesla P100-PCIE-16GB GPU. For Susy ($n = 5 \times 10^6, p = 18$): we used Gaussian kernel with $k(x, x') = e^{-\|x-x'\|^2/(2\sigma^2)}$, with $\sigma = 5$, which we obtained through a grid search (in Rudi, Carratino, and Rosasco (2017), $\sigma = 4$ is taken for the ridge regression); $M = 10^4$ Nyström

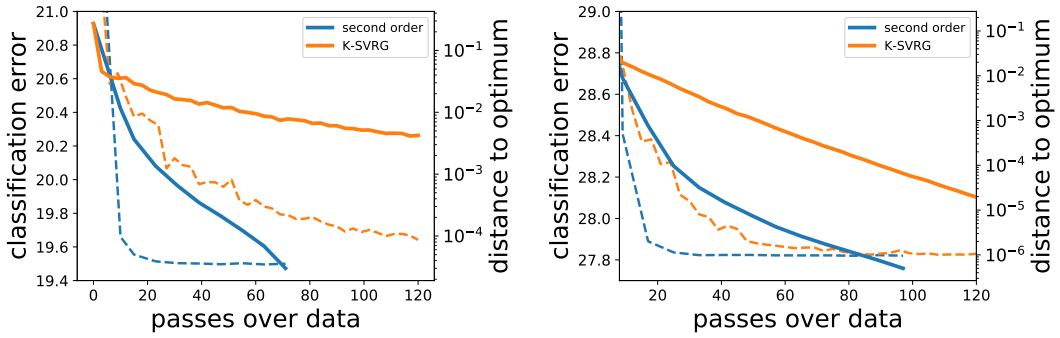

Figure 4.1: Training loss and test error as as function of the number of passes on the data for our algorithm vs. K-SVRG. on the **(left)** Susy and **(right)** Higgs data sets.

centers and a subsampling $Q = M$ for the preconditioner, both obtained with uniform sampling. Analogously for Higgs ($n = 1.1 \times 10^7, p = 28$): , we used a Gaussian kernel with $\sigma = 5$ and $M = 2.5 \times 10^4$ and $Q = M$, using again uniform sampling. To find reasonable λ for supervised learning applications, we cross-validated λ finding the minimum test error at $\lambda = 10^{-10}$ for Susy and $\lambda = 10^{-9}$ for Higgs (see Figs. 4.2 and 4.3 in Sec. 4.F) for such values our algorithm and the competitor achieve an error of 19.5% on the test set for Susy, comparable to the state of the art (19.6% Rudi, Carratino, and Rosasco (2017)) and analogously for Higgs (see Sec. 4.F). We then used such λ's as regularization parameters and compared our algorithm with a well known accelerated stochastic gradient technique *Katyusha SVRG* (K-SVRG) Allen-Zhu (2017), tailored to our problem using mini batches. In Fig. 4.1 we show the convergence of the training loss and classification error with respect to the number of passes on the data, of our algorithm compared to K-SVRG. It is possible to note our algorithm is order of magnitude faster in achieving convergence, validating empirically the fact that the proposed algorithm scales as $O(n\mathsf{df}_\lambda)$ in learning settings, while accelerated first order methods go as $O((n + \sqrt{nL/\lambda})\mathsf{df}_\lambda)$. Moreover, as mentioned in the introduction, this highlights the fact that precise optimization is necessary to achieve a good performance in terms of test error. Finally, note that since a pass on the data is much more expensive for K-SVRG than for our second order method (see Sec. 4.F for details), the difference in computing time between the second order scheme and K-SVRG is even more in favour of our second order method (see Figs. 4.4 and 4.5 in Sec. 4.F).

Organization of the Appendix

4.A . Main results on generalized self-concordant functions

Notations, definitions and basic results concerning generalized self-concordant functions.

4.B . Results on approximate Newton methods

In this section, the interaction between the notion of Dikin ellipsoid, approximate Newton methods and generalized self-concordant functions is studied. The results needed in the main paper are all concentrated in Sec. 4.B .3. In particular the results in Lemma 4.2 are proven in a more general form in Lemma 4.11.

4.C . Proof of bounds for the globalization scheme

In this section, we leverage the results of the previous two sections to analyze the globalization scheme.

4.C .1. Main technical lemmas

We start by proving the result on the inclusion of Dikin ellipsoids (Lemma 4.3).

4.C .2. Proof of main theorems

In particular, a general version of Theorem 4.1 is proven. Moreover Remark 9 is proven in proposition 4.7, while the fixed scheme to choose $(q_k)_{k\in\mathbb{N}}$ is proven in proposition 4.8.

4.C .3. Proof of Thm. 1

Finally, we prove the properties of the globalization schemes presented in Theorem 4.1.

4.D . Non-parametric learning with generalized self-concordant functions

In this section, some basic results about non-parametric learning with generalized self-concordant functions are recalled and the main results of Sec. 4.4 are proven.

4.D .1. General setting and assumptions, statistical result for regularized ERM.

More details about the generalization properties of empirical risk minimization as well as the optimal rates in Theorem 4.2 are recalled.

4.D .2. Reducing the dimension: projecting on a subspace using Nyström sub-sampling.

4.D .3. Sub-sampling techniques.

The basics of uniform sub-sampling and sub-sampling with approximate leverage scores are recalled.

4.D .4. Selecting the M Nyström points

Theorem 4.3 is proven in a more general version in Theorem 4.6.

4.D .5 Performing the globalization scheme to approximate $\beta_{M,\lambda}$

A general scheme is proposed to solve the projected problem approximately using the globalization scheme.

4.D .5. Performing approximate Newton steps

We start by describing the way of computing approximate Newton steps. A generalized version of Lemma 4.4 is proven in proposition 4.12.

4.D .5. Applying the globalization scheme to control $\widehat{\nu}_{M,\lambda}(\beta)$
We then completely analyse the approximating of $\beta_{M,\lambda}$ from an optimization point of view (see proposition 4.13).

4.D .6. Final algorithm and results
Finally, the proof of Theorem 4.4 is provided, using the results of the previous subsections.

4.E . Algorithm

In this section, the pseudocode for the algorithm presented in Sec. 4.4 and analyzed in Theorem 4.7 is provided.

4.F . Experiments

In this section, more details about the experiments are provided.

4.G . Solving a projected problem to reduce dimension

In this section, more details about the problem of randomized projections are provided.

4.G .2. Relating the projected to the original problem
In particular, results to relate the ERM with the projected ERM in terms of excess risk are provided for generalized self-concordant functions.

4.H . Relations between statistical problems and empirical problem.

In this section, we provide results to relate excess expected risk with excess empirical risk for generalized self-concordant functions.

4.I . Multiplicative approximations for Hermitian operators

In this section, some general analytic results on multiplicative approximations for Hermitian operators are derived. Moreover they are used to provide a simplified proof for the results in Lemma 4.1. See in particular Lemmas 4.28 and 4.29 and Pilanci and Wainwright (2017), Lemma 2.

4.A Main results on generalized self-concordant functions

In this section, we start by introducing a few notations. We define the key notion of generalized self-concordance in Sec. 4.A .1, and present the main results concerning generalized self-concordant functions. In Sec. 4.A .2, we describe how generalized self-concordance behaves with respect to an expectation or to certain relaxations.

Notations Let $\lambda \geq 0$ and $\mathbf{A}$ be a bounded positive semidefinite Hermitian operator on $\mathcal{H}$. We denote with $\mathbf{I}$ the identity operator, and

$$\|x\|_{\mathbf{A}} := \|\mathbf{A}^{1/2}x\|, \tag{4.11}$$

$$\mathbf{A}_\lambda := \mathbf{A} + \lambda\mathbf{I}. \tag{4.12}$$

Let f be a twice differentiable convex function on a Hilbert space $\mathcal{H}$. We adopt the following notation for the Hessian of f:

$$\forall x \in \mathcal{H},\ \mathbf{H}_f(x) := \nabla^2 f(x) \in \mathcal{L}(\mathcal{H}).$$

For any $\lambda > 0$, we define the λ-regularization of f:

$$f_\lambda := f + \frac{\lambda}{2}\|\cdot\|^2.$$

f_λ is λ-strongly convex and has a unique minimizer which we denote with $x_\star^{f,\lambda}$. Moreover, define

$$\forall x \in \mathcal{H},\ \mathbf{H}_{f,\lambda}(x) := \nabla^2 f_\lambda(x) = \mathbf{H}_f(x) + \lambda \mathbf{I}, \qquad \nu_{f,\lambda}(x) := \|\nabla f_\lambda(x)\|_{\mathbf{H}_{f,\lambda}^{-1}(x)}.$$

The quantity $\nu_{f,\lambda}(x)$ is called the **Newton decrement** at point x and will play a significant role.

When the function f is clear from the context, we will omit the subscripts with f and use $\mathbf{H}, \mathbf{H}_\lambda, \nu_\lambda$....

4.A .1 Definitions and results on generalized self-concordant functions

In this section, we introduce the main definitions and results for self-concordant functions. These results are mainly the same as in appendix B of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019).

Definition 4.3 (generalized self-concordant function)**.** *Let $\mathcal{H}$ be a Hilbert space. Formally, a generalized self-concordant function on $\mathcal{H}$ is a couple $(f, \mathcal{G})$ where:*

i $\mathcal{G}$ is a bounded subset of $\mathcal{H}$; we will usually denote $\|\mathcal{G}\|$ or R the quantity $\sup_{g\in\mathcal{G}} \|g\| < \infty$;

ii f is a convex and three times differentiable mapping on $\mathcal{H}$ such that

$$\forall x \in \mathcal{H},\ \forall h, k \in \mathcal{H},\ \nabla^{(3)} f(x)[h,k,k] \leq \sup_{g\in\mathcal{G}} |g \cdot h|\ \nabla^2 f(x)[k,k].$$

To make notations lighter, we will often omit $\mathcal{G}$ from the notations and simply say that f stands both for the mapping and the couple $(f, \mathcal{G})$.

Definition 4.4 (Definitions)**.** *Let f be a generalized self-concordant function. We define the following quantities.*

- $\forall h \in \mathcal{H},\ \mathsf{t}_f(h) := \sup_{g\in\mathcal{G}} |h \cdot g|$;
- $\forall x \in \mathcal{H},\ \forall \lambda > 0,\ \mathsf{r}_{f,\lambda}(x) := \frac{1}{\sup_{g\in\mathcal{G}} \|g\|_{\mathbf{H}_{f,\lambda}^{-1}(x)}}$;
- $\forall \mathsf{c} \geq 0,\ \forall \lambda > 0,\ \mathsf{D}_{f,\lambda}(\mathsf{c}) := \{x\ :\ \nu_{f,\lambda}(x) \leq \mathsf{c}\mathsf{r}_{f,\lambda}(x)\}$.

We also define the following functions:

$$\psi(t) = \frac{e^t - t - 1}{t^2}, \quad \underline{\phi}(t) = \frac{1 - e^{-t}}{t}, \quad \overline{\phi}(t) = \frac{e^t - 1}{t}. \tag{4.13}$$

Note that ψ, $\overline{\phi}$ are increasing functions and that $\underline{\phi}$ is a decreasing function. Moreover, $\frac{\overline{\phi}(t)}{\underline{\phi}(t)} = e^t$. Once again, if f is clear, we will often omit the reference to f in the quantities above, keeping only $\mathsf{t}, \mathsf{r}_\lambda, \mathsf{D}_\lambda$...

We condense results obtained in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) under a slightly different form. The proofs, however, are exactly the same.

While in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), only the regularized case is dealt with, the proof techniques are exactly the same to obtain proposition 4.1. proposition 4.2 is proved explicitly in Proposition 4 of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) and Lemma 4.5 is proved in Proposition 5.
Omitting the subscript f, we get the following results.

Proposition 4.1 (Bounds for the non-regularized function f)**.** *Let f be a generalized self-concordant function. Then the following bounds hold (we omit f in the subscripts):*

$$\forall x \in \mathcal{H},\ \forall h \in \mathcal{H},\ e^{-\mathsf{t}(h)}\mathbf{H}(x) \preceq \mathbf{H}(x+h) \preceq e^{\mathsf{t}(h)}\mathbf{H}(x), \tag{4.14}$$

$$\forall x, h \in \mathcal{H},\ \forall \lambda > 0,\ \|\nabla f(x+h) - \nabla f(x)\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \overline{\phi}(\mathsf{t}(h))\|h\|_{\mathbf{H}_\lambda(x)}, \tag{4.15}$$

$$\forall x, h \in \mathcal{H},\ \psi(-\mathsf{t}(h))\|h\|^2_{\mathbf{H}(x)} \leq f(x+h) - f(x) - \nabla f(x).h \leq \psi(\mathsf{t}(h))\|h\|^2_{\mathbf{H}(x)}. \tag{4.16}$$

We get the analoguous bounds in the regularized case.

Proposition 4.2 (Bounds for the regularized function f_λ)**.** *Let f be a generalized self-concordant function and $\lambda > 0$ be a regularizer. Then the following bounds hold:*

$$\forall x, h \in \mathcal{H},\ e^{-\mathsf{t}(h)}\mathbf{H}_\lambda(x) \preceq \mathbf{H}_\lambda(x+h) \preceq e^{\mathsf{t}(h)}\mathbf{H}_\lambda(x), \tag{4.17}$$

$$\forall x, h \in \mathcal{H},\ \underline{\phi}(\mathsf{t}(h))\|h\|_{\mathbf{H}_\lambda(x)} \leq \|\nabla f_\lambda(x+h) - \nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \overline{\phi}(\mathsf{t}(h))\|h\|_{\mathbf{H}_\lambda(x)}, \tag{4.18}$$

$$\forall x, h \in \mathcal{H},\ \psi(-\mathsf{t}(h))\|h\|^2_{\mathbf{H}_\lambda(x)} \leq f_\lambda(x+h) - f_\lambda(x) - \nabla f_\lambda(x).h \leq \psi(\mathsf{t}(h))\|h\|^2_{\mathbf{H}_\lambda(x)}. \tag{4.19}$$

Corollary 4.1. *Let f be a $\mathcal{G}$ generalized self-concordant function and $\lambda > 0$ be a regularizer, and $x^\star_\lambda$ the unique minimizer of f_λ. Then the following bounds hold for any $x \in \mathcal{H}$:*

$$\underline{\phi}(\mathsf{t}(x - x^\star_\lambda))\|x - x^\star_\lambda\|_{\mathbf{H}_\lambda(x)} \leq \underbrace{\|\nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)}}_{\nu_\lambda(x)} \leq \overline{\phi}(\mathsf{t}(x - x^\star_\lambda))\|x - x^\star_\lambda\|_{\mathbf{H}_\lambda(x)}, \tag{4.20}$$

$$\psi(-\mathsf{t}(x - x^\star_\lambda))\|x - x^\star_\lambda\|^2_{\mathbf{H}_\lambda(x^\star_\lambda)} \leq f_\lambda(x) - f_\lambda(x^\star_\lambda) \leq \psi(\mathsf{t}(x - x^\star_\lambda))\|x - x^\star_\lambda\|^2_{\mathbf{H}_\lambda(x^\star_\lambda)}. \tag{4.21}$$

Moreover, the following localization lemma holds.

Lemma 4.5 (localization)**.** *Let $\lambda > 0$ be fixed. If $\frac{\nu_\lambda(x)}{\mathsf{r}_\lambda(x)} < 1$, then*

$$\mathsf{t}(x - x^\star_\lambda) \leq -\log\left(1 - \frac{\nu_\lambda(x)}{\mathsf{r}_\lambda(x)}\right). \tag{4.22}$$

In particular, this shows:

$$\forall \mathsf{c} < 1,\ \forall \lambda > 0,\ x \in \mathsf{D}_\lambda(\mathsf{c}) \implies \mathsf{t}(x - x^\star_\lambda) \leq -\log(1 - \mathsf{c}).$$

We now state a Lemma which shows that the difference to the optimum in function values is equivalent to the squared newton decrement in a small Dikin ellipsoid. We will use this result in the main paper.

Lemma 4.6 (Equivalence of norms)**.** *Let $\lambda > 0$ and $x \in \mathsf{D}_\lambda(\frac{1}{7})$. Then the following holds:*

$$\frac{1}{4}\nu_\lambda(x)^2 \leq f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \nu_\lambda(x)^2.$$

Proof. Apply Lemma 4.5 knowing $x \in \mathsf{D}_\lambda(\frac{1}{7})$ to get $\mathsf{t}(x - x_\lambda^\star) \leq \log(7/6)$. Then apply Eq. (4.19) and Eq. (4.18) to get:

$$\begin{aligned} f_\lambda(x) - f_\lambda(x_\lambda^\star) &\leq \psi(\mathsf{t}(x - x_\lambda^\star))\|x - x_\lambda^\star\|^2_{\mathbf{H}_\lambda(x_\lambda^\star)} \\ &\leq e^{\mathsf{t}(x-x_\lambda^\star)}\psi(\mathsf{t}(x - x_\lambda^\star))\|x - x_\lambda^\star\|^2_{\mathbf{H}_\lambda(x)} \\ &\leq \frac{e^{\mathsf{t}(x-x_\lambda^\star)}\psi(\mathsf{t}(x - x_\lambda^\star))}{\underline{\phi}(\mathsf{t}(x - x_\lambda^\star))^2}\nu_\lambda(x)^2. \end{aligned}$$

Replacing with the bound above, we get

$$\forall \lambda > 0,\ \forall x \in \mathsf{D}_\lambda(\frac{1}{7}),\ f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \nu_\lambda(x)^2.$$

For the lower bound, proceed in exactly the same way. □

4.A .2 Comparison between generalized self-concordant functions

The following result is straightforward.

Lemma 4.7 (Comparison between generalized self-concordant functions)**.** *Let $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{H}$ be two bounded subsets. If $(f, \mathcal{G}_1)$ is generalized self-concordant, then $(f, \mathcal{G}_2)$ is also generalized self-concordant. Moreover,*

$$\forall x \in \mathcal{H},\ \forall \lambda > 0,\ \mathsf{r}_{(f,\mathcal{G}_1),\lambda}(x) \geq \mathsf{r}_{(f,\mathcal{G}_2),\lambda}(x).$$

In particular, we will often use the following fact. If $(f, \mathcal{G})$ is generalized self-concordant, and $\mathcal{G}$ is bounded by R, then $(f, B_\mathcal{H}(R))$ is also generalized self-concordant. Moreover,

$$\mathsf{r}_{(f,B_\mathcal{H}(R)),\lambda}(x) = \frac{\sqrt{\lambda + \lambda_{\min}(\mathbf{H}_f(x))}}{R} \geq \frac{\sqrt{\lambda}}{R}.$$

We now state a result which shows that, given a family of generalized self-concordant functions, the expectancy of that family is also generalized self-concordant. This can be seen as a reformulation of Proposition 2 of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019).

Proposition 4.3 (Expectation)**.** *Let $\mathcal{Z}$ be a polish space equipped with its Borel sigma-algebra, and $\mathcal{H}$ be a Hilbert space. Let $((f_z, \mathcal{G}_z))_{z\in\mathcal{Z}}$ be a family of generalized self-concordant functions such that the mapping $(z, x) \mapsto f_z(x)$ is measurable.*

Assume we are given a random variable Z on $\mathcal{Z}$, whose support we denote with $\operatorname{supp}(Z)$, such that

- *the random variables $\|f_Z(0)\|, \|\nabla f_Z(0)\|, \operatorname{Tr}(\nabla^2 f_Z(0))$ are are bounded;*
- *$\mathcal{G} := \bigcup_{z\in\operatorname{supp}(Z)} \mathcal{G}_z$ is a bounded subset of $\mathcal{H}$.*

Then the mapping $f : x \in \mathcal{H} \mapsto \mathbb{E}\left[f_Z(x)\right]$ *is well defined,* $(f, \mathcal{G})$ *is generalized self-concordant, and we can differentiate under the expectation.*

Corollary 4.2. *Let* $n \in \mathbb{N}$ *and* $(f_i, \mathcal{G}_i)_{1 \leq i \leq n}$ *be a family of generalized self-concordant functions. Define*

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x), \ \mathcal{G} = \bigcup_{i=1}^{n} \mathcal{G}_i.$$

Then $(f, \mathcal{G})$ *is generalized self-concordant.*

4.B Results on approximate Newton methods

In this section, we assume we are given a generalized self-concordant function f in the sense of Sec. 4.A . As f will be fixed throughout this part, we will omit it from the notations. Recall the definitions from definition 4.4:

$$\nu_\lambda(x) := \|\nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)}, \ \frac{1}{\mathsf{r}_\lambda(x)} := \sup_{g\in\mathcal{G}} \|g\|_{\mathbf{H}_\lambda^{-1}(x)}, \ \mathsf{D}_\lambda(\mathsf{c}) := \left\{ x \ : \ \frac{\nu_\lambda(x)}{\mathsf{r}_\lambda(x)} \leq \mathsf{c} \right\}.$$

Define the following quantities:

- the true Newton step at point x for the λ-regularized problem:

$$\Delta_\lambda(x) := \mathbf{H}_\lambda^{-1}(x)\nabla f_\lambda(x).$$

- the renormalized Newton decrement $\widetilde{\nu}_\lambda(x)$:

$$\widetilde{\nu}_\lambda(x) := \frac{\nu_\lambda(x)}{\mathsf{r}_\lambda(x)}.$$

Moreover, note that a direct application of Eq. (4.17) yields the following equation which relates the radii at different points:

$$\forall \lambda > 0, \ \forall x \in \mathcal{H}, \ \forall h \in \mathcal{H}, \ e^{-\mathsf{t}(h)}\mathsf{r}_\lambda(x) \leq \mathsf{r}_\lambda(x+h) \leq e^{\mathsf{t}(h)}\mathsf{r}_\lambda(x). \tag{4.23}$$

In this appendix, we develop a complete analysis of so-called approximate Newton methods in the case of generalized self-concordant losses. By "approximate Newton method", we mean that instead of performing the classical update $x_{t+1} = x_t - \Delta_\lambda(x_t)$, we perform an update of the form $x_{t+1} = x_t - \widetilde{\Delta}_t$ where $\widetilde{\Delta}_t$ is an approximation of the real Newton step. We will characterize this approximation by measuring its distance to the real Newton step using two parameters ρ and ϵ_0:

$$\|\widetilde{\Delta}_t - \Delta_\lambda(x_t)\| \leq \rho\nu_\lambda(x_t) + \epsilon_0.$$

We start by presenting a few technical results in Sec. 4.B .1. We continue by proving that an approximate Newton method has linear convergence guarantees in the right Dikin ellipsoid in Sec. 4.B .2. In Sec. 4.B .3, we adapt these results to a certain way of computing approximate Newton steps, which will be the one we use in the core of the paper. In Sec. 4.B .4, we mention ways to reduce the computational burden of these methods by showing that since all Hessians are equivalent in Dikin ellipsoids, one can actually sketch the Hessian at one given point in that ellipsoid instead of re-sketching it at each Newton step. For the sake of simplicity, this is not mentioned in the core paper, but works very well in practice.

4.B .1 Main technical results

We start with a technical decomposition of the Newton decrement at point $x - \widetilde{\Delta}$ for a given $\widetilde{\Delta} \in \mathcal{H}$.

Lemma 4.8 (Technical decomposition)**.** *Let $\lambda > 0$, $x \in \mathcal{H}$ be fixed. Assume we perform a step of the form $x - \widetilde{\Delta}$ for a certain $\widetilde{\Delta} \in \mathcal{H}$. Define*

$$\delta := \|\widetilde{\Delta} - \Delta_\lambda(x)\|_{\mathbf{H}_\lambda(x)}, \qquad \widetilde{\delta} := \frac{\delta}{\mathsf{r}_\lambda(x)}.$$

The following holds:

$$\widetilde{\nu}_\lambda(x - \widetilde{\Delta}) \leq e^{\widetilde{\nu}_\lambda(x)+\widetilde{\delta}} \left[\psi(\widetilde{\nu}_\lambda(x) + \widetilde{\delta})(\widetilde{\nu}_\lambda(x) + \widetilde{\delta})^2 + \widetilde{\delta}\right]; \tag{4.24}$$

$$\nu_\lambda(x - \widetilde{\Delta}_\lambda(x)) \leq e^{\widetilde{\nu}_\lambda(x)+\widetilde{\delta}} \left[\psi(\widetilde{\nu}_\lambda(x) + \widetilde{\delta})(\widetilde{\nu}_\lambda(x) + \widetilde{\delta})(\nu_\lambda(x) + \delta) + \delta\right]. \tag{4.25}$$

Proof. Note that by definition, $\nabla f_\lambda(x) = \mathbf{H}_\lambda(x)\Delta_\lambda(x)$. Hence

$$\begin{aligned}
\|\nabla f^\lambda(x - \widetilde{\Delta})\|_{\mathbf{H}_\lambda^{-1}(x)} &= \|\nabla f^\lambda(x - \widetilde{\Delta}) - \nabla f^\lambda(x) + \mathbf{H}_\lambda(x)\Delta_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)} \\
&\leq \|\nabla f^\lambda(x - \widetilde{\Delta}) - \nabla f^\lambda(x) + \mathbf{H}_\lambda(x)\widetilde{\Delta}\|_{\mathbf{H}_\lambda^{-1}(x)} \\
&+ \|\mathbf{H}_\lambda(x)(\Delta_\lambda(x) - \widetilde{\Delta})\|_{\mathbf{H}_\lambda^{-1}(x)} \\
&= \|\int_0^1 [\mathbf{H}_\lambda(x - s\widetilde{\Delta}) - \mathbf{H}_\lambda(x)]\widetilde{\Delta} ds\|_{\mathbf{H}_\lambda^{-1}(x)} + \delta \\
&\leq \int_0^1 \|\mathbf{H}_\lambda^{-1/2}(x)\mathbf{H}_\lambda(x - s\widetilde{\Delta})\mathbf{H}_\lambda^{-1/2}(x) - \mathbf{I}\| ds \; \|\widetilde{\Delta}\|_{\mathbf{H}_\lambda(x)} + \delta.
\end{aligned}$$

Now using Eq. (4.17), one has $\|\mathbf{H}_\lambda^{-1/2}(x)\mathbf{H}_\lambda(x - s\widetilde{\Delta})\mathbf{H}_\lambda^{-1/2}(x) - \mathbf{I}\| \leq e^{s\mathsf{t}(\widetilde{\Delta})} - 1$, whose integral on s is $\psi(\mathsf{t}(\widetilde{\Delta}))\mathsf{t}(\widetilde{\Delta})$ where ψ is defined in definition 4.4. Morever, bounding

$$\|\widetilde{\Delta}\|_{\mathbf{H}_\lambda(x)} \leq \|\widetilde{\Delta} - \Delta_\lambda(x)\|_{\mathbf{H}_\lambda(x)} + \|\Delta_\lambda(x)\|_{\mathbf{H}_\lambda(x)} = \delta + \nu_\lambda(x),$$

it holds

$$\|\nabla f^\lambda(x - \widetilde{\Delta})\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \psi(\mathsf{t}(\widetilde{\Delta}))\mathsf{t}(\widetilde{\Delta}) \; (\nu_\lambda(x) + \delta) + \delta.$$

1. Now note that using Eq. (4.17), it holds: $\nu_\lambda(x - \widetilde{\Delta}) \leq e^{\mathsf{t}(\widetilde{\Delta})/2}\|\nabla f^\lambda(x - \widetilde{\Delta})\|_{\mathbf{H}_\lambda^{-1}(x)}$ and hence:

$$\nu_\lambda(x - \widetilde{\Delta}) \leq e^{\mathsf{t}(\widetilde{\Delta})/2} \left(\psi(\mathsf{t}(\widetilde{\Delta}))\mathsf{t}(\widetilde{\Delta}) \; (\nu_\lambda(x) + \delta) + \delta\right). \tag{4.26}$$

2. Moreover, using Eq. (4.23),

$$\widetilde{\nu}_\lambda(x - \widetilde{\Delta}) \leq e^{\mathsf{t}(\widetilde{\Delta})} \left(\psi(\mathsf{t}(\widetilde{\Delta}))\mathsf{t}(\widetilde{\Delta}) \; (\widetilde{\nu}_\lambda(x) + \widetilde{\delta}) + \widetilde{\delta}\right). \tag{4.27}$$

Noting that

$$\mathsf{t}(\widetilde{\Delta}) \leq \frac{\|\widetilde{\Delta}\|_{\mathbf{H}_\lambda(x)}}{\mathsf{r}_\lambda(x)} \leq \widetilde{\nu}_\lambda(x) + \widetilde{\delta},$$

and bounding Eq. (4.26) simply by taking $e^{\mathsf{t}(\widetilde{\Delta})/2} \leq e^{\mathsf{t}(\widetilde{\Delta})}$, we get the two bounds in the lemma.

□

We now place ourselves in the case where we are given an approximation of the Newton step of the following form. Assume λ and x are fixed, and that we approximate $\Delta_\lambda(x)$ with $\widetilde{\Delta}$ such that there exists $\rho \geq 0$ and $\epsilon_0 \geq 0$ such that it holds:

$$\|\widetilde{\Delta} - \Delta_\lambda(x)\|_{\mathbf{H}_\lambda(x)} \leq \rho\nu_\lambda(x) + \epsilon_0.$$

We define/prove the three different following regimes.

Lemma 4.9 (3 regimes)**.** *Let $x \in \mathsf{D}_\lambda\left(\frac{1}{7}\right)$ and $\lambda > 0$ be fixed. Let*

$$0 \leq \rho \leq \frac{1}{7},\ \epsilon_0 \geq 0 \text{ s.t. } \tilde{\varepsilon}_0 := \frac{\epsilon_0}{\mathsf{r}_\lambda(x)} \leq \frac{1}{21}.$$

Let $\widetilde{\Delta}$ be an approximation of the Newton steps satisfying $\|\widetilde{\Delta} - \Delta_\lambda(x)\|_{\mathbf{H}_\lambda(x)} \leq \rho\nu_\lambda(x) + \epsilon_0$. The three following regimes appear.

- *If $\widetilde{\nu}_\lambda(x) \geq \rho$ and $\widetilde{\nu}_\lambda(x)^2 \geq \tilde{\varepsilon}_0$, then we are in the **quadratic regime**, i.e.*

$$\frac{10\widetilde{\nu}_\lambda(x - \widetilde{\Delta}_\lambda(x))}{3} \leq \left(\frac{10\widetilde{\nu}_\lambda(x)}{3}\right)^2,\ \nu_\lambda(x - \widetilde{\Delta}_\lambda(x)) \leq \frac{10}{3}\widetilde{\nu}_\lambda(x)\nu_\lambda(x).$$

- *If $\rho \geq \widetilde{\nu}_\lambda(x)$ and $\rho\widetilde{\nu}_\lambda(x) \geq \widetilde{\epsilon}_0$, then we are in the **linear regime**, i.e.*

$$\frac{10}{3}\widetilde{\nu}_\lambda(x - \widetilde{\Delta}_\lambda(x)) \leq \left(\frac{10\rho}{3}\right)\left(\frac{10}{3}\widetilde{\nu}_\lambda(x)\right),\ \nu_\lambda(x - \widetilde{\Delta}_\lambda(x)) \leq \frac{10}{3}\widetilde{\nu}_\lambda(x)\nu_\lambda(x).$$

- *If $\widetilde{\epsilon}_0 \geq \widetilde{\nu}_\lambda(x)^2, \rho\ \widetilde{\nu}_\lambda(x)$, then the **maximal precision** of the approximation is reached, and it holds:*

$$\widetilde{\nu}_\lambda(x - \widetilde{\Delta}_\lambda(x)) \leq 3\widetilde{\epsilon}_0 \leq \frac{1}{7},\ \nu_\lambda(x - \widetilde{\Delta}_\lambda(x)) \leq 3\epsilon_0.$$

Proof. Using the previous lemma,

$$\begin{aligned}\widetilde{\nu}_\lambda(x - \widetilde{\Delta}_\lambda(x)) &\leq e^{(1+\rho)\widetilde{\nu}_\lambda(x)+\widetilde{\epsilon}_0}\left[\psi((1+\rho)\widetilde{\nu}_\lambda(x) + \widetilde{\epsilon}_0)((1+\rho)\widetilde{\nu}_\lambda(x) + \widetilde{\epsilon}_0)^2 + \rho\widetilde{\nu}_\lambda(x) + \widetilde{\epsilon}_0\right] \\ &\leq \square_1(\widetilde{\nu}_\lambda(x), \rho, \widetilde{\epsilon}_0)\ \widetilde{\nu}_\lambda(x)^2 + \square_2(\widetilde{\nu}_\lambda(x), \rho, \widetilde{\epsilon}_0)\ \rho\widetilde{\nu}_\lambda(x) + \square_3(\widetilde{\nu}_\lambda(x), \rho, \widetilde{\epsilon}_0)\ \widetilde{\epsilon}_0,\end{aligned}$$

and

$$\nu_\lambda(x - \widetilde{\Delta}_\lambda(x)) \leq \square_1(\widetilde{\nu}_\lambda(x), \rho, \widetilde{\epsilon}_0)\ \widetilde{\nu}_\lambda(x)\nu_\lambda(x) + \square_2(\widetilde{\nu}_\lambda(x), \rho, \widetilde{\epsilon}_0)\ \rho\nu_\lambda(x) + \square_3(\widetilde{\nu}_\lambda(x), \rho, \widetilde{\epsilon}_0)\ \epsilon_0,$$

where the following defintions are used:

$$\begin{aligned}\square_1(\widetilde{\nu}, \rho, \widetilde{\epsilon}_0) &:= e^{(1+\rho)\widetilde{\nu}+\widetilde{\epsilon}_0}\psi((1+\rho)\widetilde{\nu} + \widetilde{\epsilon}_0)(1+\rho)^2, \\ \square_2(\widetilde{\nu}, \rho, \widetilde{\epsilon}_0) &:= e^{(1+\rho)\widetilde{\nu}+\widetilde{\epsilon}_0}, \\ \square_3(\widetilde{\nu}, \rho, \widetilde{\epsilon}_0) &:= e^{(1+\rho)\widetilde{\nu}+\widetilde{\epsilon}_0}\left[2\psi((1+\rho)\widetilde{\nu} + \widetilde{\epsilon}_0)(1+\rho)\widetilde{\nu} + 1\right].\end{aligned}$$

Now assume $\widetilde{\epsilon}_0 \leq \frac{1}{21}$, $\widetilde{\nu}_\lambda(x), \rho \leq \frac{1}{7}$. Replacing these values in the functions above bounds $\square_1, \square_2$ and $\square_3$, and using the case distinction, we get the result. □

4.B .2 General analysis of an approximate Newton method

The following proposition describes the behavior of an approximate newton method where ρ and ϵ_0 are fixed a priori.

Proposition 4.4 (General approximate Newton scheme results)**.** *Let $\mathsf{c} \leq \frac{1}{7}$ be fixed and $x_0 \in \mathsf{D}_\lambda(\mathsf{c})$ be a given starting point.*
Let $\rho \leq \frac{1}{7}$ and ϵ_0 such that $\epsilon_0 \leq \frac{\mathsf{c}}{4}\,\mathsf{r}_\lambda(x_0)$.
Define the following approximate Newton scheme:

$$\forall t \geq 0,\ x_{t+1} = x_t - \widetilde{\Delta}_t, \qquad \|\widetilde{\Delta}_t - \Delta_\lambda(x_t)\|_{\mathbf{H}_\lambda(x_t)} \leq \rho\nu_\lambda(x_t) + \epsilon_0.$$

The following guarantees hold.

- $\forall t \geq 0,\ x_t \in \mathsf{D}_\lambda(\mathsf{c})$.
- *Let* $t_c = \left\lfloor \log_2 \log_2 \frac{3}{10\rho} \right\rfloor + 1$.

$$\forall t \leq t_c,\ \frac{10\widetilde{\nu}_\lambda(x_t)}{3} \leq \max\left(\frac{12\epsilon_0}{\mathsf{r}_\lambda(x_0)}, 2^{-2^t}\right),$$

$$\forall t \geq t_c,\ \frac{10\widetilde{\nu}_\lambda(x_t)}{3} \leq \max\left(\frac{12\epsilon_0}{\mathsf{r}_\lambda(x_0)}, \left(\frac{10\rho}{3}\right)^{t-t_c+1}\right).$$

- *We can bound the relative decrease for both the Newton decrement and the renormalized Newton decrement:*

$$\begin{aligned}
\forall t \leq t_c, \qquad & \nu_\lambda(x_t) \leq \max\left(3\epsilon_0, \left(\frac{1}{2}\right)^{2^t-1} \nu_\lambda(x_0)\right), \\
& \widetilde{\nu}_\lambda(x_t) \leq \max\left(\frac{18\epsilon_0}{5\mathsf{r}_\lambda(x_0)}, \left(\frac{1}{2}\right)^{2^t-1} \widetilde{\nu}_\lambda(x_0)\right). \\
\forall t \geq t_c, \qquad & \nu_\lambda(x_t) \leq \max\left(3\epsilon_0, \left(\frac{10\rho}{3}\right)^{t-t_c+1} \nu_\lambda(x_0)\right), \\
& \widetilde{\nu}_\lambda(x_t) \leq \max\left(\frac{18\epsilon_0}{5\mathsf{r}_\lambda(x_0)}, \left(\frac{10\rho}{3}\right)^{t-t_c+1} \widetilde{\nu}_\lambda(x_0)\right).
\end{aligned}$$

Proof. Start by noting, using Eq. (4.23),

$$\forall x \in \mathsf{D}_\lambda\left(\frac{1}{7}\right),\ \varepsilon \leq \frac{\mathsf{r}_\lambda(x)}{21},\ \frac{6}{7}\mathsf{r}_\lambda(x_0) \leq \mathsf{r}_\lambda(x) \leq \frac{7}{6}\mathsf{r}_\lambda(x_0). \tag{4.28}$$

In particular, this holds for any $x \in \mathsf{D}_\lambda(\mathsf{c}),\ \mathsf{c} \leq \frac{1}{7}$. Thus,

$$\forall \mathsf{c} \leq \frac{1}{7},\ \forall x_0 \in \mathsf{D}_\lambda(\mathsf{c}),\ \frac{\epsilon_0}{\mathsf{r}_\lambda(x_0)} \leq \frac{\mathsf{c}}{4} \implies \forall x \in \mathsf{D}_\lambda(\mathsf{c}),\ \frac{\epsilon_0}{\mathsf{r}_\lambda(x)} \leq \frac{\mathsf{c}}{3}.$$

1. Proving the first point is simple by induction. Indeed, assume $\widetilde{\nu}_\lambda(x_t) \leq \mathsf{c}$. We can apply Lemma 4.9 since the conditions on ε and ρ guarantee that the conditions of this lemma are satisfied.

If we are in either the linear or quadratic regime, the fact that $\frac{10\rho}{3}, \frac{10\widetilde{\nu}_\lambda(x_t)}{3} \leq \frac{10}{21}$ show that $\widetilde{\nu}_\lambda(x_{t+1}) \leq \frac{10}{21}\widetilde{\nu}_\lambda(x_t) \leq \mathsf{c}$.

If we are in the last case, $\widetilde{\nu}_\lambda(x_{t+1}) \leq \frac{3\epsilon_0}{\mathsf{r}_\lambda(x_t)} \leq \mathsf{c}$.

2. Let us prove the second bullet point by induction. Start by assuming the property holds at t. By the previous point, the hypothesis of Lemma 4.9 are satisfied at x_t with ρ and ε. Assume we are in the limiting case; we easily show that in this case,

$$\frac{10\widetilde{\nu}_\lambda(x_{t+1})}{3} \leq \frac{10}{3}\, 3\frac{\epsilon_0}{\mathsf{r}_\lambda(x_t)} \leq \frac{35\epsilon_0}{3\mathsf{r}_\lambda(x_0)}.$$

Here, the last inequality comes from Eq. (4.28). If we are not in the limiting case, let us distinguish between the two following cases.

If $t \leq t_c - 1$,

$$\begin{aligned}\frac{10\widetilde{\nu}_\lambda(x_{t+1})}{3} &\leq \frac{10\widetilde{\nu}_\lambda(x_t)}{3}\max\left(\frac{10\widetilde{\nu}_\lambda(x_t)}{3}, \frac{10\rho}{3}\right)\\ &\leq \max\left(\frac{35\epsilon_0}{3\mathsf{r}_\lambda(x_0)}, \frac{10\widetilde{\nu}_\lambda(x_t)}{3}\max\left(\left(\frac{1}{2}\right)^{2^t}, \frac{10\rho}{3}\right)\right),\end{aligned}$$

where the last inequality comes from using the induction hypothesis and the fact that $\frac{10\widetilde{\nu}_\lambda(x_t)}{3} \leq 1$. Using once again the induction hypotheses and the fact that $t \leq \left\lfloor \log_2 \log_2 \frac{3}{10\rho} \right\rfloor$ which implies $\frac{10\rho}{3} \leq \left(\frac{1}{2}\right)^{2^t}$, we finally get

$$\frac{10\widetilde{\nu}_\lambda(x_{t+1})}{3} \leq \max\left(\frac{35\epsilon_0}{3\mathsf{r}_\lambda(x_0)}, \left(\frac{1}{2}\right)^{2^{t+1}}\right).$$

The fact that the second property holds for $t = t_c$ is trivial Now consider the case where $t \geq t_c$. Using the same technique as before but noting that in this case

$$\frac{10\widetilde{\nu}_\lambda(x_t)}{3} \leq \max\left(\frac{35\epsilon_0}{3\mathsf{r}_\lambda(x_0)}, \left(\frac{10\rho}{3}\right)^{t-t_c+1}\right) \leq \max\left(\frac{35\epsilon_0}{3\mathsf{r}_\lambda(x_0)}, \frac{10\rho}{3}\right),$$

We easily use Lemma 4.9 to reach the desired conclusion.

3. Let $t < t_c$. Then using Lemma 4.9:

$$\forall s \leq t,\ \nu_\lambda(x_{s+1}) \leq \max\left(3\epsilon_0, \max(\frac{10\rho}{3}, \frac{10\widetilde{\nu}_\lambda(x_s)}{3})\nu_\lambda(x_s)\right).$$

Using the fact that for any $s \leq t$, $\frac{10\widetilde{\nu}_\lambda(x_s)}{3} \leq \max(\frac{35\epsilon_0}{3\mathsf{r}_\lambda(x_0)}, \left(\frac{1}{2}\right)^{2^s})$:

$$\forall s \leq t,\ \nu_\lambda(x_{s+1}) \leq \max\left(3\epsilon_0, \frac{35\epsilon_0}{3}\frac{\nu_\lambda(x_s)}{\mathsf{r}_\lambda(x_0)}, \max(\frac{10\rho}{3}, \left(\frac{1}{2}\right)^{2^s})\nu_\lambda(x_s)\right).$$

Now using the fact that for any $s \leq t$, $\widetilde{\nu}_\lambda(x_s) \leq \frac{1}{7}$, we see that $\frac{\nu_\lambda(x_s)}{\mathsf{r}_\lambda(x_0)} \leq \frac{7}{6}\widetilde{\nu}_\lambda(x_s) \leq \frac{1}{6}$ and hence $\frac{35\epsilon_0}{3}\frac{\nu_\lambda(x_s)}{\mathsf{r}_\lambda(x_0)} \leq 3\epsilon_0$. Moreover, since $s \leq t < t_c$, $\max(\frac{10\rho}{3}, \left(\frac{1}{2}\right)^{2^s}) = \left(\frac{1}{2}\right)^{2^s}$. Thus:

$$\forall s \leq t,\ \nu_\lambda(x_{s+1}) \leq \max\left(3\epsilon_0, \left(\frac{1}{2}\right)^{2^s} \nu_\lambda(x_s)\right).$$

Combining these results yields:

$$\nu_\lambda(x_{t+1}) \leq \max\left(3\epsilon_0, \left(\frac{1}{2}\right)^{2^{t+1}-1} \nu_\lambda(x_0)\right).$$

This shows the first equation, that is:

$$\forall t \leq t_c,\ \nu_\lambda(x_t) \leq \max\left(3\epsilon_0, \left(\frac{1}{2}\right)^{2^{t}-1} \nu_\lambda(x_0)\right).$$

The case for $t \geq t_c$ is completely analogous. We can also reproduce the same proof to get the same bounds for $\widetilde{\nu}$, since the bounds in Lemma 4.9 are the same for both.

□

4.B .3 Main results in the paper

In the main paper, we mention two types of Newton method. First, we present a result of convergence on the full Newton method:

Lemma 4.10 (Quadratic convergence of the full Newton method)**.** *Let* $\mathsf{c} \leq \frac{1}{7}$ *and* $x_0 \in \mathsf{D}_\lambda(\mathsf{c})$. *Define*

$$x_{t+1} = x_t - \Delta_\lambda(x_t).$$

Then this scheme converges quadratically, i.e.:

$$\forall t \in \mathbb{N},\ \frac{\nu_\lambda(x_t)}{\nu_\lambda(x_0)}, \frac{\widetilde{\nu}_\lambda(x_t)}{\widetilde{\nu}_\lambda(x_0)} \leq 2^{-(2^t-1)}.$$

Thus :

- $\forall t \in \mathbb{N},\ x_t \in \mathsf{D}_\lambda(\mathsf{c})$.
- *For any* $\tilde{\mathsf{c}} \leq \mathsf{c}$ *then* $\forall t \geq \left\lceil \log_2\left(1 + \log_2 \frac{\mathsf{c}}{\tilde{\mathsf{c}}}\right)\right\rceil,\ x_t \in \mathsf{D}_\lambda(\tilde{\mathsf{c}})$.
- *For any* $\varepsilon > 0$, $\forall t \geq \left\lceil \log_2\left(1 + \log_2 \frac{\nu_\lambda(x_0)}{\sqrt{\varepsilon}}\right)\right\rceil,\ \nu_\lambda(x_t) \leq \sqrt{\varepsilon},\ f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \varepsilon$.
- *If we perform the Newton method and return the first* x_t *such that* $\nu_\lambda(x_t) \leq \sqrt{\varepsilon}$, *then the number of Newton steps computations is at most* $1 + \left\lceil \log_2\left(1 + \log_2 \frac{\nu_\lambda(x_0)}{\sqrt{\varepsilon}}\right)\right\rceil$.

Proof. A full Newton method is an approximate Newton method where $\rho, \epsilon_0 = 0$. Thus apply proposition 4.4; note that in this case $t_c = +\infty$. The last point shows that if $\mathsf{c} \leq \frac{1}{7}$, and if we perform the Newton method with a full Newton step, then

$$\forall t \geq 0,\ \widetilde{\nu}_\lambda(x_t) \leq 2^{-(2^t-1)}\nu_\lambda(x_0),\ \widetilde{\nu}_\lambda(x_t) \leq 2^{-(2^t-1)}\nu_\lambda(x_0).$$

This shows the quadratic convergence, and the first two points directly follow. For the third point, the result for $\nu_\lambda(x_t)$ directly follows from the previous equation, and the one on function

values is a direct consequence of Lemma 4.6 and the fact that $x_t \in \mathsf{D}_\lambda(1/7)$.

For the last point, note that $\nu_t(x_t) = \nabla f_\lambda(x_t) \cdot \Delta_\lambda(x_t)$ is accessible. Moreover, the bound on t is given in the point before, and since one has to compute $\Delta_\lambda(x_s)$ for $0 \leq s \leq t$, there are at most $t+1$ computations. □

In the main paper, we compute approximate Newton steps by considering methods which naturally yield only a relative error ρ and no absolute error ϵ_0. Indeed, we take the following notation.

Approximate solutions to linear problems. Let $\mathbf{A}$ be a positive definite Hermitian operator on $\mathcal{H}$, b in $\mathcal{H}$, and a wanted relative precision ρ.

We say that x is a ρ-relative approximation to the linear problem $\mathbf{A}x = b$ and write $x \in \mathrm{LinApprox}(\mathbf{A}, b, \rho)$ if the following holds:

$$\|\mathbf{A}^{-1}b - x\|_{\mathbf{A}} \leq \rho \|b\|_{\mathbf{A}^{-1}} = \rho \|\mathbf{A}^{-1}b\|_{\mathbf{A}}.$$

Note that if $x \in \mathrm{LinApprox}(\mathbf{A}, b, \rho)$ for $\rho < 1$, then

$$(1-\rho)\|b\|_{\mathbf{A}^{-1}} \leq x \cdot b \leq (1+\rho)\|b\|_{\mathbf{A}^{-1}}.$$

The following lemma shows that if, instead of computing the exact Newton step, we compute a relative approximation of the Newton step belonging to $\mathrm{LinApprox}(\mathbf{H}_\lambda(x), \nabla f_\lambda(x), \rho)$ for a given $\rho < 1$, then one has linear convergence. Moreover, we show that we can still perform a method which automatically stops.

Proposition 4.5 (relative approximate Newton method)**.** *Let $\lambda > 0$, $\rho \leq \frac{1}{7}$, $\mathsf{c} \leq \frac{1}{7}$ and a starting point $x_0 \in \mathsf{D}_\lambda(\mathsf{c})$. Assume we perform the following Newton scheme:*

$$\forall t \geq 0,\ x_{t+1} = x_t - \widetilde{\Delta}_t, \qquad \widetilde{\Delta}_t \in \mathrm{LinApprox}(\mathbf{H}_\lambda(x_t), \nabla f_\lambda(x_t), \rho).$$

Then the scheme converges linearly, i.e.

$$\forall t \in \mathbb{N},\ \frac{\nu_\lambda(x_t)}{\nu_\lambda(x_0)}, \frac{\widetilde{\nu}_\lambda(x_t)}{\widetilde{\nu}_\lambda(x_0)} \leq 2^{-t}.$$

Thus,

- $\forall t \in \mathbb{N},\ x_t \in \mathsf{D}_\lambda(\mathsf{c})$.
- *For any $\tilde{\mathsf{c}} \leq \mathsf{c}$ then $\forall t \geq \left\lceil \log_2 \frac{\mathsf{c}}{\tilde{\mathsf{c}}} \right\rceil$, $x_t \in \mathsf{D}_\lambda(\tilde{\mathsf{c}})$.*
- *For any $\varepsilon > 0$, $\forall t \geq \left\lceil \log_2 \frac{\nu_\lambda(x_0)}{\sqrt{\varepsilon}} \right\rceil$, $\nu_\lambda(x_t) \leq \sqrt{\varepsilon}$, $f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \varepsilon$*
- *If the method is performed and returns the first x_t such that $x_t \cdot \widetilde{\Delta}_t \leq \frac{6}{7}\varepsilon$, then at most $2 + \left\lfloor \log_2 \left(\sqrt{\frac{4}{3}} \frac{\nu_\lambda(x_0)}{\sqrt{\varepsilon}} \right) \right\rfloor$ approximate Newton steps computations have been performed, and $\nu_\lambda(x_t) \leq \sqrt{\varepsilon}$, $f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \varepsilon$.*

Proof. Apply proposition 4.4 with $\epsilon_0 = 0$ and $\rho = \frac{1}{7}$, since if $\rho \leq \frac{1}{7}$, then a fortiori the approximation satisfies the condition for $\rho = \frac{1}{7}$. The last point clearly states that

$$\forall t \in \mathbb{N},\ \frac{\nu_\lambda(x_t)}{\nu_\lambda(x_0)}, \frac{\widetilde{\nu}_\lambda(x_t)}{\widetilde{\nu}_\lambda(x_0)} \leq \left(\frac{10}{21}\right)^t \leq 2^{-t}.$$

From this, using Lemma 4.6 for the third point, the first three points are easily proven.
For the last point, note that since $\widetilde{\Delta}_t \in \text{LinApprox}(\mathbf{H}_\lambda(x_t), \nabla f_\lambda(x_t), \rho)$, the following holds: $\nabla f_\lambda(x_t) \cdot \widetilde{\Delta}_t = \nu_\lambda(x_t)^2 + \nabla f_\lambda(x_t) \cdot \left(\widetilde{\Delta}_t - \mathbf{H}_\lambda^{-1}(x_t)\nabla f_\lambda(x_t)\right)$. Now bound

$$|\nabla f_\lambda(x_t) \cdot \left(\widetilde{\Delta}_t - \mathbf{H}_\lambda^{-1}(x_t)\nabla f_\lambda(x_t)\right)| \leq \nu_\lambda(x_t) \; \|\widetilde{\Delta}_t - \mathbf{H}_\lambda^{-1}(x_t)\nabla f_\lambda(x_t)\|_{\mathbf{H}_\lambda(x_t)} \leq \rho\nu_\lambda(x_t)^2.$$

Thus:

$$(1-\rho)\nu_\lambda(x_t)^2 \leq \nabla f_\lambda(x_t) \cdot \widetilde{\Delta}_t \leq (1+\rho)\nu_\lambda(x_t)^2.$$

Since $\rho = \frac{1}{7}$, we see that if $\nabla f_\lambda(x_t) \cdot \widetilde{\Delta}_t \leq \frac{6}{7}\varepsilon$, then $\nu_\lambda(x_t)^2 \leq \varepsilon$. Moreover, since we stop at the first t where $\nabla f_\lambda(x_t) \cdot \widetilde{\Delta}_t \leq \frac{6}{7}\varepsilon$, then if t denotes the time at which we stop,

$$\frac{6}{7}\varepsilon < \nabla f_\lambda(x_{t-1}) \cdot \widetilde{\Delta}_{t-1} \leq \frac{8}{7}\nu_\lambda(x_{t-1})^2 \implies \nu_\lambda(x_{t-1})^2 \geq \frac{3}{4}\varepsilon.$$

Since $\nu_\lambda(x_{t-1})^2 \leq 2^{-2(t-1)}\nu_\lambda(x_0)^2$, this implies in turn that $t - 1 \leq \log_2\left(\sqrt{\frac{4}{3}}\frac{\nu_\lambda(x_0)}{\sqrt{\varepsilon}}\right)$. Thus, necessarily, $t \leq 1 + \left\lfloor \log_2\left(\sqrt{\frac{4}{3}}\frac{\nu_\lambda(x_0)}{\sqrt{\varepsilon}}\right)\right\rfloor$, and since we compute approximate Newton steps for $s = 0, ..., t$, we finally have that the number of approximate Newton steps is bounded by

$$2 + \left\lfloor \log_2\left(\sqrt{\frac{4}{3}}\frac{\nu_\lambda(x_0)}{\sqrt{\varepsilon}}\right)\right\rfloor.$$

□

Last but not least, we summarize all these theorem in the following simple result.

Lemma 4.11. *Let $\lambda > 0, \mathsf{c} \leq 1/7$, let f_λ be generalized self-concordant and $x \in \mathsf{D}_\lambda(\mathsf{c})$. It holds: $\frac{1}{4}\nu_\lambda(x)^2 \leq f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \nu_\lambda(x)^2$. Moreover, the full Newton method starting from x_0 has quadratic convergence, i.e. if x_t is obtained via $t \in \mathbb{N}$ steps of the Newton method Eq. (4.2), then $\nu_\lambda(x_t) \leq 2^{-(2^t-1)}\nu_\lambda(x_0)$. Finally, the approximate Newton method starting from x_0 has linear convergence, i.e. if x_t is obtained via $t \in \mathbb{N}$ steps of Eq. (4.3), with $\widetilde{\Delta}_t \in \text{LinApprox}(\mathbf{H}_\lambda(x_t), \nabla f_\lambda(x_t), \rho)$ and $\rho \leq 1/7$, then $\nu_\lambda(x_t) \leq 2^{-t}\nu_\lambda(x_0)$.*

Proof. The three points are obtained in the following lemmas, assuming $x \in \mathsf{D}_\lambda(1/7)$.

- For $\frac{1}{4}\nu_\lambda(x)^2 \leq f_\lambda(x) - f_\lambda(x_\lambda^\star) \leq \nu_\lambda(x)^2$, see Lemma 4.6 in Sec. 4.A .1.
- The convergence rate of the full Newton method starting in $\mathsf{D}_\lambda(1/7)$ is obtained in Lemma 4.10.
- The convergence rate of the approximate Newton method starting in $\mathsf{D}_\lambda(1/7)$ is obtained in proposition 4.5.

□

4.B .4 Sketching the Hessian only once in each Dikin ellispoid

In this section, we provide a lemma which shows in essence that if we are in a small Dikin ellipsoid, then we can keep the Hessian of the starting point and compute approximations of $\mathbf{H}_\lambda^{-1}(x_0)\nabla f_\lambda(x_t)$; they will be good approximations to $\mathbf{H}_\lambda^{-1}(x_t)\nabla f_\lambda(x_t)$ as well.

Lemma 4.12. *Let* $\mathsf{c} < 1$ *and* $x_0 \in \mathsf{D}_\lambda(\mathsf{c})$ *be fixed.*

Let $\widetilde{\mathbf{H}}$ *be an approximation of the Hessian at* x_0*, approximation wich we quantify with*

$$t := \|\mathbf{H}_\lambda^{-1/2}(x_0)\left(\mathbf{H}_\lambda(x_0) - \widetilde{\mathbf{H}}\right)\mathbf{H}_\lambda^{-1/2}(x_0)\|.$$

Assume

$$1 + t < 2(1-\mathsf{c})^2.$$

Let $b \in \mathcal{H}$*. If* $\widetilde{\Delta} \in \mathrm{LinApprox}(\widetilde{\mathbf{H}}_\lambda, b, \tilde{\rho})$*, then*

$$\forall x \in \mathsf{D}_\lambda(\mathsf{c}),\ \widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{H}_\lambda(x), b, \rho),\ \rho = \frac{(\tilde{\rho}-1)(1-\mathsf{c})^2 + (1+t)}{2(1-\mathsf{c})^2 - (1+t)}.$$

In particular, if $\mathsf{c} \leq \frac{1}{30}$*,* $x_0 \in \mathsf{D}_\lambda(\mathsf{c})$*,*

$$\forall x \in \mathsf{D}_\lambda(\mathsf{c}),\ \forall b \in \mathcal{H},\ \widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{H}_\lambda(x_0), b, \frac{1}{20}) \implies \widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{H}_\lambda(x), b, \frac{1}{7}).$$

Proof. First, start with a general theoretical result.

1. Let $\mathbf{A}$ and $\mathbf{B}$ be two positive semi-definite hermitian operators. Let $\lambda > 0$, $b \in \mathcal{H}$ and $\widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{B}_\lambda, b, \tilde{\rho})$. Decompose

$$\begin{aligned}\|\mathbf{A}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{A}_\lambda} &\leq \|\mathbf{A}_\lambda^{-1}b - \mathbf{B}_\lambda^{-1}b\|_{\mathbf{A}_\lambda} + \|\mathbf{B}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{A}_\lambda} \\ &\leq \|\mathbf{A}_\lambda^{1/2}(\mathbf{A}_\lambda^{-1} - \mathbf{B}_\lambda^{-1})\mathbf{A}_\lambda^{1/2}\|\ \|b\|_{\mathbf{A}_\lambda^{-1}} + \|\mathbf{A}_\lambda^{1/2}\mathbf{B}_\lambda^{-1/2}\|\ \|\mathbf{B}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{B}_\lambda}.\end{aligned}$$

Now using the fact that $\mathbf{A}_\lambda^{-1} - \mathbf{B}_\lambda^{-1} = \mathbf{B}_\lambda^{-1}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1}$,

$$\begin{aligned}\|\mathbf{A}_\lambda^{1/2}(\mathbf{A}_\lambda^{-1} - \mathbf{B}_\lambda^{-1})\mathbf{A}_\lambda^{1/2}\| &\leq \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|\ \|\mathbf{A}_\lambda^{1/2}\mathbf{B}_\lambda^{-1}\mathbf{A}_\lambda^{1/2}\| \\ &= \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|\ \|\mathbf{A}_\lambda^{1/2}\mathbf{B}_\lambda^{-1/2}\|^2.\end{aligned}$$

Moreover,

$$\|\mathbf{B}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{B}_\lambda} \leq \tilde{\rho}\|b\|_{\mathbf{B}_\lambda^{-1}} \leq \|\mathbf{A}^{1/2}\mathbf{B}^{-1/2}\|\ \|b\|_{\mathbf{A}_\lambda^{-1}}.$$

Putting things together, and noting that from Lemma 4.21, $\|\mathbf{A}^{1/2}\mathbf{B}^{-1/2}\|^2 \leq \frac{1}{1-\|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|}$ as soon as $\|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\| < 1$, it holds:

$$\widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{A}_\lambda, b, \rho),\ \rho = \frac{\tilde{\rho} + \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|}{1 - \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|}.$$

The aim is now to apply this lemma to $\mathbf{A} = \mathbf{H}(x)$ and $\mathbf{B} = \widetilde{\mathbf{H}}$.

2. Let $x, x_0 \in \mathsf{D}_\lambda(\mathsf{c})$. Using Lemma 4.22, we see that

$$1 + \|\mathbf{H}_\lambda^{-1/2}(x)(\widetilde{\mathbf{H}} - \mathbf{H}(x))\mathbf{H}_\lambda^{-1/2}(x)\| \leq (1+t)(1 + \|\mathbf{H}_\lambda^{-1/2}(x)(\mathbf{H}(x_0) - \mathbf{H}(x))\mathbf{H}_\lambda^{-1/2}(x)\|).$$

Using Eq. (4.17), it holds:

$$(e^{-\mathsf{t}(x-x_0)} - 1)\mathbf{I} \preceq \mathbf{H}_\lambda^{-1/2}(x)(\mathbf{H}(x_0) - \mathbf{H}(x))\mathbf{H}_\lambda^{-1/2}(x) \preceq (e^{\mathsf{t}(x_0-x)} - 1)\mathbf{I}.$$

Thus,

$$\|\mathbf{H}_\lambda^{-1/2}(x)(\mathbf{H}(x_0) - \mathbf{H}(x))\mathbf{H}_\lambda^{-1/2}(x)\| \leq \max(1 - e^{-\mathsf{t}(x-x_0)}, e^{\mathsf{t}(x-x_0)} - 1) = e^{\mathsf{t}(x-x_0)} - 1.$$

Finally, using the fact that $x_0, x \in \mathsf{D}_\lambda(\mathsf{c})$ for $\mathsf{c} < 1$ yields $\mathsf{t}(x - x_0) \leq 2\log\frac{1}{1-\mathsf{c}}$. Hence

$$1 + \|\mathbf{H}_\lambda^{-1/2}(x)(\mathbf{H}(x_0) - \mathbf{H}(x))\mathbf{H}_\lambda^{-1/2}(x)\| \leq \frac{1}{(1-\mathsf{c})^2}.$$

Thus,

$$\|\mathbf{H}_\lambda^{-1/2}(x)(\widetilde{\mathbf{H}} - \mathbf{H}(x))\mathbf{H}_\lambda^{-1/2}(x)\| \leq \frac{1+t}{(1-\mathsf{c})^2} - 1.$$

The result then follows. □

4.C Proof of bounds for the globalization scheme

In this section, we prove that the scheme of decreasing μ towards λ converges.

4.C .1 Main technical lemmas

Lemma 4.13 (Next μ)**.** *Let $\mu > 0$, $\mathsf{c} < 1$.*

$$\nu_\mu(x) \leq \frac{\mathsf{c}}{3}\,\frac{\sqrt{\mu}}{R} \implies \nu_{\widetilde{\mu}}(x) \leq \mathsf{c}\,\frac{\sqrt{\widetilde{\mu}}}{R}, \qquad \widetilde{\mu} := q\,\mu, \qquad q \geq \frac{\frac{1}{3} + \frac{R\sqrt{\mu}\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}}}{1 + \frac{R\sqrt{\mu}\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}}}.$$

$$x \in \mathsf{D}_\mu\left(\frac{\mathsf{c}}{3}\right) \implies x \in \mathsf{D}_{\widetilde{\mu}}(\mathsf{c}), \qquad \widetilde{\mu} := q\,\mu, \qquad q \geq \frac{\frac{1}{3} + \frac{\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}\ \mathsf{r}_\mu(x)}}{1 + \frac{\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}\ \mathsf{r}_\mu(x)}}.$$

Proof. For any $\widetilde{\mu} < \mu$, note that

$$\forall x \in \mathcal{H},\ \|\mathbf{H}_{\widetilde{\mu}}^{-1/2}(x)\mathbf{H}_\mu^{1/2}(x)\| = \sqrt{\frac{\lambda_{\min}(\mathbf{H}(x)) + \mu}{\lambda_{\min}(\mathbf{H}(x)) + \widetilde{\mu}}} \leq \sqrt{\mu/\widetilde{\mu}}.$$

This shows that $\|\cdot\|_{\mathbf{H}_{\widetilde{\mu}}^{-1}(x)} \leq \sqrt{\frac{\mu}{\widetilde{\mu}}}\,\|\cdot\|_{\mathbf{H}_\mu^{-1}(x)}$, and in particular that $\frac{1}{\mathsf{r}_{\widetilde{\mu}}(x)} \leq \sqrt{\mu/\widetilde{\mu}}\frac{1}{\mathsf{r}_\mu(x)}$.
Using this fact, it holds:

$$\begin{aligned}
\widetilde{\nu}_{\widetilde{\mu}}(x) &= \frac{\|\nabla f_{\widetilde{\mu}}(x)\|_{\mathbf{H}_{\widetilde{\mu}}^{-1}(x)}}{\mathsf{r}_{\widetilde{\mu}}(x)} \\
&= \frac{\|\nabla f_\mu(x) - (\mu - \widetilde{\mu})x\|_{\mathbf{H}_{\widetilde{\mu}}^{-1}(x)}}{\mathsf{r}_{\widetilde{\mu}}(x)} \\
&\leq \frac{\mu}{\widetilde{\mu}}\,\frac{\|\nabla f_\mu(x)\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{r}_\mu(x)} + \left(\frac{\mu}{\widetilde{\mu}} - 1\right)\frac{\|\mu x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{r}_\mu(x)}.
\end{aligned}$$

Hence, if $\widetilde{\nu}_\mu(x) \leq \frac{\mathsf{c}}{3}$, a condition to obtain $\widetilde{\nu}_{\widetilde{\mu}}(x) \leq \mathsf{c}$ is the following:

$$\frac{\mu}{\widetilde{\mu}}\left(\frac{\mathsf{c}}{3} + t\right) \leq \mathsf{c} + t \Leftrightarrow \widetilde{\mu} \geq \mu\frac{\mathsf{c}/3 + t}{\mathsf{c} + t} \qquad t = \frac{\|\mu x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{r}_\mu(x)}.$$

This yields the second point of the lemma. The analysis is completely analoguous for the first. □

Lemma 4.14 (Useful bounds for q)**.** *Let $\mu > 0$. Then the following hold:*

$$\forall x \in \mathcal{H},\ \frac{\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{r}_\mu(x)} \leq R\sqrt{\mu}\|x\|_{\mathbf{H}_\mu^{-1}(x)} \leq R\|x\|.$$

Moreover, we can bound all of these quantities using $x_\mu^\star$:

- *For any $\mathsf{c} < 1$, $x \in \mathcal{H}$, if $x \in \mathsf{D}_\mu(\mathsf{c}/3)$, then the following holds:*

$$\frac{\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}\ \mathsf{r}_\mu(x)} \leq \frac{1}{3}\left(1 + \frac{1}{1 - \mathsf{c}/3}\right) + \frac{1}{1 - \mathsf{c}/3}\frac{\|\mu x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}}{\mathsf{c}\ \mathsf{r}_\mu(x_\mu^\star)}.$$

- *For any* $\mathsf{c} < 1$, $x \in \mathcal{H}$, *if* $\frac{R\nu_\mu(x)}{\sqrt{\mu}} \leq \frac{\mathsf{c}}{3}$, *then the following holds:*

$$\frac{R\sqrt{\mu}\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}} \leq \left(1 + \frac{1}{1-\mathsf{c}/3}\right)\frac{1}{3} + \sqrt{\frac{1}{1-\mathsf{c}/3}}\frac{R\sqrt{\mu}\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}}{\mathsf{c}}.$$

Likewise, it can be shown that under the same conditions:

$$\frac{R\|x\|}{\mathsf{c}} \leq \frac{R\|x_\mu^\star\|}{\mathsf{c}} + \frac{1}{3}\overline{\phi}(-\log(1-\mathsf{c}/3)).$$

Proof. The first bound is obvious. Moreover, the fact that $\widetilde{\nu}_\mu(x) \leq \frac{\mathsf{c}}{3}$ implies that $\mathsf{t}(x - x_\mu^\star) \leq \log\frac{1}{1-\mathsf{c}/3}$. Thus, we get the classical bounds on the Hessian using Eq. (4.14):

$$e^{-\mathsf{t}(x-x_\mu^\star)}\mathbf{H}(x) \preceq \mathbf{H}(x_\mu^\star) \preceq e^{\mathsf{t}(x-x_\mu^\star)}\mathbf{H}(x).$$

1. Bound on $\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}$. Using Eqs. (4.17) and (4.18),

$$\begin{aligned}\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)} &= \|\nabla f_\mu(x) - \nabla f(x) + \nabla f(x_\mu^\star) - \nabla f(x_\mu^\star)\|_{\mathbf{H}_\mu^{-1}(x)} \\ &\leq \nu_\mu(x) + \int_0^1 \|\mathbf{H}_\mu(x)^{-1/2}\mathbf{H}(x_t)(x - x_\mu^\star)\|\, dt + \|\nabla f(x_\mu^\star)\|_{\mathbf{H}_\mu(x)},\ x_t = tx + (1-t)x_\mu^\star.\end{aligned}$$

Now bound $\|\mathbf{H}_\mu(x)^{-1/2}\mathbf{H}(x_t)(x-x_\mu^\star)\| \leq \|\mathbf{H}_\mu(x)^{-1/2}\,\mathbf{H}_\mu(x_t)^{1/2}\|\,\|x-x_\mu^\star\|_{\mathbf{H}(x_t)}$ and use Eq. (4.17) and Eq. (4.14) to get:

$$\|\mathbf{H}_\mu(x)^{-1/2}\mathbf{H}(x_t)(x - x_\mu^\star)\| \leq e^{t\ \mathsf{t}(x-x_\mu^\star)}\|x - x_\mu^\star\|_{\mathbf{H}(x)}.$$

Integrating this yields:

$$\int_0^1 \|\mathbf{H}_\mu(x)^{-1/2}\mathbf{H}(x_t)(x - x_\mu^\star)\|\, dt \leq \overline{\phi}(\mathsf{t}(x - x_\mu^\star))\ \|x - x_\mu^\star\|_{\mathbf{H}(x)} \leq e^{\mathsf{t}(x-x_\mu^\star)}\ \nu_\mu(x).$$

Where the last inequality is obtained using the bounds between gradient and hessian distance Eq. (4.18). Finally, using the bound on $\mathsf{t}(x - x_\mu^\star)$,

$$\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)} \leq \left(1 + \frac{1}{1-\mathsf{c}/3}\right)\nu_\mu(x) + \sqrt{\frac{1}{1-\mathsf{c}/3}}\|\nabla f(x_\mu^\star)\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}.$$

2. Bound on $R\|x\|$. Start by decomposing

$$R\|x\| \leq R\|x_\mu^\star\| + R\|x - x_\mu^\star\|.$$

Now bound

$$R\|x - x_\mu^\star\| \leq \frac{R}{\sqrt{\mu}}\|x - x_\mu^\star\|_{\mathbf{H}_\mu(x)}.$$

Using Eq. (4.17), $\|x - x_\mu^\star\|_{\mathbf{H}_\mu(x)} \leq \overline{\phi}(-\log(1-\mathsf{c}/3))\nu_\mu(x)$. Hence:

$$R\|x\| \leq R\|x_\mu^\star\| + \overline{\phi}(-\log(1-\mathsf{c}/3))\frac{R\nu_\mu(x)}{\sqrt{\mu}}.$$

3. Now assume $x \in \mathsf{D}_\mu(\mathsf{c}/3)$. Using the bound on $\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}$, and noting that

$$\frac{1}{\mathsf{r}_\mu(x)} \leq e^{\mathsf{t}(x-x_\mu^\star)/2}\frac{1}{\mathsf{r}_\mu(x_\mu^\star)},$$

it holds:

$$\frac{\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}\ \mathsf{r}_\mu(x)} \leq \frac{1}{3}\left(1+\frac{1}{1-\mathsf{c}/3}\right)+\frac{1}{1-\mathsf{c}/3}\frac{\|\mu x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}}{\mathsf{c}\ \mathsf{r}_\mu(x_\mu^\star)}.$$

4. Now assume $\frac{R\nu_\mu(x)}{\sqrt{\mu}} \leq \frac{\mathsf{c}}{3}$. . We know that in particular, $x \in \mathsf{D}_\mu(\mathsf{c}/3)$ and hence:

$$\begin{aligned} R\sqrt{\mu}\|x\|_{\mathbf{H}_\mu^{-1}(x)} &\leq \left(1+\frac{1}{1-\mathsf{c}/3}\right)\frac{R\nu_\mu(x)}{\sqrt{\mu}}+\sqrt{\frac{1}{1-\mathsf{c}/3}}\frac{R\mu\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}}{\sqrt{\mu}} \\ &\leq \left(1+\frac{1}{1-\mathsf{c}/3}\right)\frac{\mathsf{c}}{3}+\sqrt{\frac{1}{1-\mathsf{c}/3}}R\sqrt{\mu}\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}. \end{aligned}$$

Hence

$$\frac{R\sqrt{\mu}\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{c}} \leq \left(1+\frac{1}{1-\mathsf{c}/3}\right)\frac{1}{3}+\sqrt{\frac{1}{1-\mathsf{c}/3}}\frac{R\sqrt{\mu}\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}}{\mathsf{c}}.$$

Likewise:

$$\frac{R\|x\|}{\mathsf{c}} \leq \frac{R\|x_\mu^\star\|}{\mathsf{c}}+\frac{1}{3}\overline{\phi}(-\log(1-\mathsf{c}/3)).$$

□

We can get the following simpler bounds.

Corollary 4.3 (Application to $\mathsf{c}=\frac{1}{7}$)**.** *Applying Lemma 4.14 to $\mathsf{c}=\frac{1}{7}$, we get the following bounds. Let $\mu>0$.*

- *For any $x \in \mathcal{H}$, if $x \in \mathsf{D}_\mu(\mathsf{c}/3)$, then the following holds:*

$$\frac{7\mu\|x\|_{\mathbf{H}_\mu^{-1}(x)}}{\mathsf{r}_\mu(x)} \leq 1+\frac{8\|\mu x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}}{\mathsf{r}_\mu(x_\mu^\star)}.$$

- *For any $\mathsf{c}<1$, $x \in \mathcal{H}$, if $\frac{R\nu_\mu(x)}{\sqrt{\mu}} \leq \frac{\mathsf{c}}{3}$, then the following hold:*

$$7R\sqrt{\mu}\|x\|_{\mathbf{H}_\mu^{-1}(x)} \leq 1+8R\sqrt{\mu}\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}.$$

$$7R\|x\| \leq 7R\|x_\mu^\star\|+1.$$

4.C .2 Proof of main theorems

In this section, we bound the number of iterations of our scheme in different cases.

Recall the proposed globalization scheme in the paper, where $\texttt{ANM}_\rho(f,x,t)$ is a method performing t successive ρ-relative approximate Newton steps of f starting at x.

Proposed Globalization Scheme

Phase I: Getting in the Dikin ellispoid of f_λ

Start with $x_0 \in \mathcal{H}, \mu_0 > 0$, $t, T \in \mathbb{N}$ and $(q_k)_{k\in\mathbb{N}} \in (0,1]$.

For $k \in \mathbb{N}$

$x_{k+1} \leftarrow \texttt{ANM}_\rho(f_{\mu_k}, x_k, t)$

$\mu_{k+1} \leftarrow q_{k+1}\mu_k$

Stop when $\mu_{k+1} < \lambda$ and set $x_{last} \leftarrow x_k$. $K \leftarrow k$

Phase II: reach a certain precision starting from inside the Dikin ellipsoid

Return $\widehat{x} \leftarrow \texttt{ANM}_\rho(f_\lambda, x_{last}, T)$

Throughout this section, we will denote with K the value of k when the scheme stops, i.e. the first value of k such that $\mu_{k+1} < \lambda$.

Adaptive methods We start by presenting an adaptive way to select μ_{k+1} from μ_k, with theoretical guarantees. The main result is the following.

Proposition 4.6 (Adaptive, simple version)**.** *Assume that we perform phase I starting at x_0 such that*

$$\frac{R\nu_{\mu_0}(x_0)}{\sqrt{\mu_0}} \leq \frac{1}{7}.$$

Assume that at each step k, we compute x_{k+1} using $t = 2$ iterations of the ρ-relative approximate Newton method. Then if at each iteration, we set:

$$\mu_{k+1} = q_{k+1}\ \mu_k, \qquad q_{k+1} := \frac{\frac{1}{3} + 7R\|x_{k+1}\|}{1 + 7R\|x_{k+1}\|}.$$

Then the following hold:

1. $\forall k \leq K+1,\ \frac{R\nu_{\mu_k}(x_k)}{\sqrt{\mu_k}} \leq \frac{1}{7}$.

2. *The decreasing parameter q_{k+1} is bounded above before reaching K:*

$$\forall k \leq K,\ q_{k+1} \leq \frac{\frac{4}{3} + 7R\|x^\star_{\mu_k}\|}{2 + 7R\|x^\star_{\mu_k}\|} \leq \frac{\frac{4}{3} + 7R\|x^\star_\lambda\|}{2 + 7R\|x^\star_\lambda\|}.$$

3. *K is finite,*

$$K \leq \left\lfloor \frac{\log \frac{\mu_0}{\lambda}}{\log \frac{2+7R\|x^\star_\lambda\|}{\frac{4}{3}+7R\|x^\star_\lambda\|}} \right\rfloor \leq \left\lfloor (3 + 11R\|x^\star_\lambda\|) \log \frac{\mu_0}{\lambda} \right\rfloor,$$

and $\frac{R\nu_\lambda(x_{K+1})}{\sqrt{\lambda}} \leq \frac{1}{7}$.

Proof. Let us prove the three points one by one.

1. This is easily proved by induction, the keys to the induction hypothesis being:

- Using the induction hypothesis, $x_k \in \mathsf{D}_{\mu_k}(\mathsf{c})$ and hence, using proposition 4.5 shows that after two iterations of the approximate Newton scheme, $\frac{\nu_{\mu_k}(x_{k+1})}{\nu_{\mu_k}(x_k)} \leq \frac{1}{3}$ which implies $\frac{R\nu_{\mu_k}(x_{k+1})}{\sqrt{\mu_k}} \leq \frac{\mathsf{c}}{3}$.

- Now using Lemma 4.13, we see that that since

$$7R\|x_{k+1}\| = \frac{R\|x_{k+1}\|}{\mathsf{c}} \geq \frac{R\sqrt{\mu_k}\|x_{k+1}\|_{\mathbf{H}_{\mu_k}^{-1}(x_{k+1})}}{\mathsf{c}},$$

the hypotheses to guarantee the bound for q_{k+1} hold, hence

$$\frac{R\nu_{\mu_{k+1}}(x_{k+1})}{\sqrt{\mu_{k+1}}} \leq \mathsf{c}.$$

2. Using the second bullet point of Cor. 4.3, we see that the previous point implies

$$\forall k \leq K,\ 7R\|x_{k+1}\| \leq 7R\|x^\star_{\mu_k}\| + 1 \implies q_{k+1} \leq \frac{4/3 + 7R\|x^\star_{\mu_k}\|}{2 + 7R\|x^\star_{\mu_k}\|}.$$

Now using the fact that for any $k \leq K$, $\mu_k > \lambda$, we can use the simple fact that $\|x^\star_\lambda\| \geq \|x^\star_{\mu_k}\|$ to get the desired bound for q_{k+1}.

3. Using the previous point clearly shows the following bound:

$$\forall k \leq K+1,\ \mu_k \leq \left(\frac{\frac{4}{3} + 7R\|x^\star_\lambda\|}{2 + 7R\|x^\star_\lambda\|}\right)^k \mu_0.$$

As this clearly converges to 0 when k goes to infinity, K is necessarily finite. Applying this for $k = K$, we see that:

$$\lambda \leq \mu_K \leq \left(\frac{\frac{4}{3} + 7R\|x^\star_\lambda\|}{2 + 7R\|x^\star_\lambda\|}\right)^K \mu_0.$$

This shows that $K \leq \frac{\log \frac{\mu_0}{\lambda}}{\log \frac{2+7R\|x^\star_\lambda\|}{\frac{4}{3}+7R\|x^\star_\lambda\|}}$.

The final bound is obtained noting that

$$\frac{2 + 7R\|x^\star_\lambda\|}{\frac{4}{3} + 7R\|x^\star_\lambda\|} = 1 + \frac{1}{t}, \qquad t = 2 + \frac{21}{2}R\|x^\star_\lambda\|,$$

and using the classical bound:

$$\frac{1}{\log(1 + \frac{1}{t})} \leq t + 1.$$

Finally, the fact that $\frac{R\nu_\lambda(x_{K+1})}{\sqrt{\lambda}} \leq \mathsf{c}$ is just a consequence of the fact that $\mu_{K+1} \leq \lambda \leq \mu_K$ and thus that $\lambda = q\mu_K$ with $q \geq q_{K+1}$, which is shown to satisfy the condition in Lemma 4.13. Hence, the lemma holds not only for μ_{K+1} but also for λ. □

Remark 10 (μ_0)**.** *In the previous proposition, we assume start at x_0, μ_0 such that*

$$\frac{R\nu_{\mu_0}(x_0)}{\sqrt{\mu_0}} \leq \frac{1}{7}.$$

A simple way to have such a pair is simply to select:

$$x_0 = 0,\ \mu_0 = 7R\|\nabla f(0)\|,$$

since $\frac{R\nu_{\mu_0}(x_0)}{\sqrt{\mu_0}} = \frac{R\|\nabla f(0)\|_{\mathbf{H}_{\mu_0}^{-1}(0)}}{\sqrt{\mu_0}} \leq \frac{R\|\nabla f(0)\|}{\mu_0}$.

Alternatively, if one can approximately compute $\|x\|_{\mathbf{H}_\mu^{-1}(x)}$, one can propose the following variant, whose proof is completely analogous.

Proposition 4.7 (Adaptive, small variant version)**.** *Assume that we perform phase I starting at x_0 such that*

$$\frac{R\nu_{\mu_0}(x)}{\sqrt{\mu_0}} \leq \frac{1}{7}.$$

Then if at each iteration, we set:

$$t_{k+1} = 7\sqrt{\frac{7}{6}}R\sqrt{\mu_k}\sqrt{x_{k+1}\cdot s_{k+1}}, s_{k+1} \in \text{LinApprox}(\mathbf{H}_{\mu_k}(x_{k+1}), x_{k+1}, \frac{1}{7}),$$

and

$$\mu_{k+1} = q_{k+1}\ \mu_k, \qquad q_{k+1} := \frac{\frac{1}{3}+t_{k+1}}{1+t_{k+1}}.$$

Then the following hold:

1. $\forall k \leq K,\ \frac{R\nu_{\mu_k}(x_k)}{\sqrt{\mu_k}} \leq \frac{1}{7}$.

2. *The decreasing parameter q_{k+1} is bounded above before reaching K:*

$$\forall k \leq K,\ q_{k+1} \leq \sup_{\mu_0 \geq \mu \geq \lambda} \frac{\frac{7}{3}+10R\sqrt{\mu}\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}}{3+10R\sqrt{\mu}\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}} \leq \frac{\frac{7}{3}+10R\|x_\lambda^\star\|}{3+10R\|x_\lambda^\star\|}.$$

3. *K is finite,*

$$K \leq \left(\frac{9}{2}+15\sup_{\lambda\leq\mu\leq\mu_0} R\sqrt{\mu}\|x_\mu^\star\|_{\mathbf{H}_\mu^{-1}(x_\mu^\star)}\right)\log\frac{\mu_0}{\lambda},$$

and $\frac{R\nu_\lambda(x_{K+1})}{\sqrt{\lambda}} \leq \frac{1}{7}$.

Proof. The main thing to note is that because of the properties of $\frac{1}{7}$-approximations, if $s_{k+1} \in \text{LinApprox}(\mathbf{H}_{\mu_k}(x_{k+1}), x_{k+1}, \frac{1}{7})$,

$$(1-\frac{1}{7})\|x_{k+1}\|^2_{\mathbf{H}^{-1}_{\mu_k}(x_{k+1})} \leq x_{k+1}\cdot s_{k+1} \leq (1+\frac{1}{7})\|x_{k+1}\|^2_{\mathbf{H}^{-1}_{\mu_k}(x_{k+1})}.$$

Hence,

$$\|x_{k+1}\|_{\mathbf{H}^{-1}_{\mu_k}(x_{k+1})} \leq \sqrt{\frac{7}{6}}\sqrt{x_{k+1}\cdot s_{k+1}} \leq \sqrt{\frac{4}{3}}\|x_{k+1}\|_{\mathbf{H}^{-1}_{\mu_k}(x_{k+1})}.$$

Hence, $t_{k+1} \geq 7R\sqrt{\mu_k}\|x_{k+1}\|_{\mathbf{H}^{-1}_{\mu_k}(x_{k+1})}$, and we can apply Lemma 4.13 to get the first point. To get the second point, we bound t_{k+1} above:

$$t_{k+1} \leq 7\sqrt{\frac{4}{3}}R\sqrt{\mu_k}\|x_{k+1}\|_{\mathbf{H}^{-1}_{\mu_k}(x_{k+1})}.$$

Now use Cor. 4.3 to find:

$$t_{k+1} \leq \sqrt{\frac{4}{3}}\left(1+8R\sqrt{\mu_k}\|x^\star_{\mu_k}\|_{\mathbf{H}^{-1}_{\mu_k}(x^\star_{\mu_k})}\right) \leq 2+10R\sqrt{\mu_k}\|x^\star_{\mu_k}\|_{\mathbf{H}^{-1}_{\mu_k}(x^\star_{\mu_k})}.$$

Thus,

$$q_{k+1} \leq \frac{\frac{7}{3} + 10R\sqrt{\mu_k}\|x^\star_{\mu_k}\|_{\mathbf{H}^{-1}_{\mu_k}(x^\star_{\mu_k})}}{3 + 10R\sqrt{\mu_k}\|x^\star_{\mu_k}\|_{\mathbf{H}^{-1}_{\mu_k}(x^\star_{\mu_k})}}.$$

Note that as long as $k \geq K$,

$$q_{k+1} \leq \sup_{\mu \geq \lambda} \frac{\frac{7}{3} + 10R\sqrt{\mu}\|x^\star_{\mu}\|_{\mathbf{H}^{-1}_{\mu}(x^\star_{\mu})}}{3 + 10R\sqrt{\mu}\|x^\star_{\mu}\|_{\mathbf{H}^{-1}_{\mu}(x^\star_{\mu})}} \leq \frac{\frac{7}{3} + 10R\|x^\star_{\lambda}\|}{3 + 10R\|x^\star_{\lambda}\|}.$$

This guarantees convergence.

For the last point, the proof is exactly the same as in the previous proposition.

□

General non-adaptive result. As mentioned in the core of the article, in practice, we do not select q_{k+1} at each iteration using a safe adaptative value, but rather decrease $\mu_{k+1} = q\mu_k$ with a constant q, which we see as a parameter to tune. The following result shows that for q large enough, this is justified, and that the lower bound we get for q depends on the radius of the Dikin ellipsoid $\mathsf{r}_\mu(x)$, instead of $\frac{\sqrt{\mu}}{R}$ in the previous bounds, which is somewhat finer and shows that if the data is structured such that this radius is very big, then q might actually be very small.

Proposition 4.8 (Fixed q)**.** *Assume that we perform phase I starting at x_0 such that*

$$x_0 \in \mathsf{D}_{\mu_0}(\frac{1}{7}).$$

Assume we perform the method with a fixed $q_{k+1} = q$, satisfying

$$q \geq \sup_{\lambda \leq \mu \leq \mu_0} \frac{\frac{4}{3} + 8\frac{\mu\|x^\star_\mu\|_{\mathbf{H}^{-1}_\mu(x^\star_\mu)}}{\mathsf{r}_\mu(x^\star_\mu)}}{2 + 8\frac{\mu\|x^\star_\mu\|_{\mathbf{H}^{-1}_\mu(x^\star_\mu)}}{\mathsf{r}_\mu(x^\star_\mu)}}.$$

Then the following hold:

1. $\forall k \leq K+1,\ x_k \in \mathsf{D}_{\mu_k}(\frac{1}{7})$.

2. *K is finite,*

$$K \leq \frac{1}{1-q}\log\frac{\mu_0}{\lambda},$$

and $x_{K+1} \in \mathsf{D}_\lambda(\frac{1}{7})$.

Proof. Let us prove the two points.

1. Let us prove the result by induction. The initialization is trivial. Now assume $x_k \in \mathsf{D}_{\mu_k}(\frac{1}{7})$. Performing two iterations of the approximate Newton method guarantees that

$$x_{k+1} \in \mathsf{D}_{\mu_k}(\frac{1}{21}),$$

as show in proposition 4.5. Now using Lemma 4.13, we see that $x_{k+1} \in \mathsf{D}_{q\mu_k}(\frac{1}{7})$, provided that

$$q \geq \frac{\frac{1}{3} + \frac{7\mu_k \|x_{k+1}\|_{\mathbf{H}_{\mu_k}^{-1}(x_{k+1})}}{\mathsf{r}_{\mu_k}(x_{k+1})}}{1 + \frac{7\mu_k \|x_{k+1}\|_{\mathbf{H}_{\mu_k}^{-1}(x_{k+1})}}{\mathsf{r}_{\mu_k}(x_{k+1})}}.$$

Now using Cor. 4.3, we get that

$$\frac{7\mu_k \|x_{k+1}\|_{\mathbf{H}_{\mu_k}^{-1}(x_{k+1})}}{\mathsf{r}_{\mu_k}(x_{k+1})} \leq 1 + \frac{8\mu_k \|x^\star_{\mu_k}\|_{\mathbf{H}_{\mu_k}^{-1}(x^\star_{\mu_k})}}{\mathsf{r}_{\mu_k}(x^\star_{\mu_k})} \leq 1 + 8 \sup_{\lambda \leq \mu \leq \mu_0} \frac{\mu \|x^\star_\mu\|_{\mathbf{H}_\mu^{-1}(x^\star_\mu)}}{\mathsf{r}_\mu(x^\star_\mu)}.$$

Hence the result.

2. This point just follows, using the bound $\frac{1}{\log \frac{1}{q}} \leq \frac{1}{1-q}$.

□

4.C .3 Proof of Theorem 4.1

Using Remark 10, the fact that $x_0 = 0$ and $\mu_0 = 7R\|\nabla f(0)\|$, as well as the hypotheses of the theorem, we can apply proposition 4.6, and show that the number of steps K performed in the first phase is bounded:

$$K \leq \lfloor (3 + 11R\|x^\star_\lambda\|) \log(7R\|\nabla f(0)\|/\lambda) \rfloor .$$

Moreover, this proposition also shows that $R\nu_\lambda(x_{last})/\sqrt{\lambda} \leq \frac{1}{7}$. Hence, we can use proposition 4.5: if

$$t \geq T = \left\lceil \log_2 \sqrt{\frac{\lambda\varepsilon^{-1}}{R^2}} \right\rceil \geq \left\lceil \log_2 \frac{\nu_\lambda(x_{last})}{\sqrt{\varepsilon}} \right\rceil,$$

then it holds $\nu_\lambda(\hat{x}) \leq \sqrt{\varepsilon}$ and $f_\lambda(\hat{x}) - f_\lambda(x^\star_\lambda) \leq \varepsilon$. □

4.D Non-parametric learning with generalized self-concordant functions

In this section, the aim is to provide a fast algorithm in the case of Kernel methods which achieves the optimal statistical guarantees.

4.D .1 General setting and assumptions, statistical result for regularized ERM.

In this section, we consider the supervised learning problem of learning a predictor $f : \mathcal{X} \to \mathcal{Y}$ from training samples $(x_i, y_i)_{1 \leq i \leq n}$ which we assume to be realisations from a certain random variable $Z = (X, Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ whose distribution is ρ. In what follows, for simplification purposes, we assume $\mathcal{Y} = \mathbb{R}$; however, this analysis can easily be adapted (although with heavier notations) to the setting where $\mathcal{Y} = \mathbb{R}^p$. Our aim is to compute the predictor of minimal generalization error

$$\inf_{f \in \mathcal{H}} L(f) := \mathbb{E}_{z \sim \rho} \left[\ell_z(f(x)) \right], \tag{4.29}$$

where $\mathcal{H}$ is a space of candidate solutions and $\ell_z : \mathbb{R} \to \mathbb{R}$ is a loss function comparing the prediction $f(x)$ to the objective y.

Kernel methods. Kernel methods consider a space of functions $\mathcal{H}_K$ implicitly constructed from a symmetric positive semi-definite Kernel $K : \mathcal{X} \times \mathcal{X} \to$ and whose basic functions are the $K_x : t \in \mathcal{X} \mapsto K(x, t)$ and the linear combinations of such functions $f = \sum_{j=1}^m \alpha_j K_{x_j}$.

It is endowed with a scalar product such that: $\forall x_1, x_2 \in \mathcal{X},\ K_{x_1} \cdot K_{x_2} = K(x_1, x_2)$, and as a consequence, $\mathcal{H}_K$ satisfies the self-reprocucing property:

$$\forall x \in \mathcal{X},\ \forall f \in \mathcal{H},\ f(x) = \langle f, K_x \rangle_{\mathcal{H}}.$$

In order to find a good predictor for Eq. (4.29), the following estimator, called the regularized ERM estimator, is often computed:

$$\widehat{f}_\lambda := \arg\min_{f \in \mathcal{H}} \widehat{L}_\lambda(f) := \frac{1}{n} \sum_{i=1}^n \ell_{z_i}(f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2.$$

The properties of this estimator have been studied in Caponnetto and De Vito (2007) for the square loss and Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) for generalized self-concordant functions. In Sec. 4.H , we recall the full setting of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), and extend it to include the statistical properties of the projected problem.

Assumptions In this section, we will make the following assumptions, which are reformulations of the assumptions of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), which we recall in Sec. 4.H , in order to have the statistical properties of the regularized ERM. First, we assume that the (x_i, y_i) are i.i.d. samples.

Assumption 4.1 (i.i.d. data)**.** *The samples $(z_i)_{1 \leq i \leq n} = (x_i, y_i)_{1 \leq i \leq n} \in \mathcal{Z}^n$ are independently and identically distributed according to ρ.*

In the case where $\mathcal{Y} = \mathbb{R}$, we make the following assumptions on the loss, which leads to the self concordance of the mappings $f \mapsto \ell_z(f(x))$ and that of L, $\widehat{L}$...

Assumption 4.2 (Technical assumptions)**.** *The mapping $(z,t) \in \mathcal{Z} \times \mathbb{R} \mapsto \ell_z(t)$ is measurable. Moreover,*

- *there exists $\mathsf{R}_\ell < \infty$ such that for all $z \in \operatorname{supp}(Z)$,*

$$\forall t \in \mathbb{R},\ |\ell_z^{(3)}(t)| \leq \mathsf{R}_\ell \ell_z''(t),$$

- *the random variables $|\ell_Z(0)|, |\ell_Z'(0)|, |\ell_Z''(0)|$ are are bounded;*
- *The kernel is bounded, i.e. $\forall x \in \operatorname{supp}(X),\ K(x,x) \leq \kappa^2$ for a certain κ.*

Using these assumptions, we see that the following properties are satisfied. Define $L_z(f) := \ell_z(f(x))$. Then the L_z satisfy the following properties:

- For any $z \in \mathcal{Z}$, $(L_z, \{\mathsf{R}_\ell K_x\})$ is a generalized self-concordant function in the sense of definition 4.4.
- The mapping $(z,f) \in \mathcal{Z} \times \mathcal{H} \mapsto L_z(f)$ is measurable;
- the random variables $\|L_Z(0)\|, \|\nabla L_Z(0)\|, \operatorname{Tr}(\nabla^2 L_Z(0))$ are bounded by $|\ell_Z(0)|$, $\kappa|\ell_Z'(0)|$, $\kappa^2|\ell_Z''(0)|$;
- $\mathcal{G} := \{\mathsf{R}_\ell K_x \ :\ z \in \operatorname{supp}(Z)\}$ is a bounded subset of $\mathcal{H}$, bounded by $R = \mathsf{R}_\ell \kappa$.

This shows that Assumption 4.7 and Assumption 4.8 are satisfied by the L_z and hence, using proposition 4.16 in the next appendix, L is well-defined, generalized self-concordant with $\mathcal{G}$. Moreover, the empirical loss

$$\widehat{L} = \frac{1}{n}\sum_{i=1}^n L_{z_i},$$

is also generalized self-concordant with $\widehat{\mathcal{G}} := \{\mathsf{R}_\ell K_{x_i} \ :\ 1 \leq i \leq n\}$.

Finally, as in Sec. 4.H , we make an assumption on the regularity of the problem; namely, we assume that a solution to the learning problem exists in $\mathcal{H}$.

Assumption 4.3 (Existence of a minimizer)**.** *There exists $f^\star \in \mathcal{H}$ such that $L(f^\star) = \inf_{f\in\mathcal{H}} L(f)$.*

We adopt all the notations from Sec. 4.H , doing the distinction between expected an empirical problems by adding a $\widehat{\cdot}$ over the quantities related to the empirical problem. We continue using the standard notations for L: for any $f \in \mathcal{H}$ and $\lambda > 0$,

$$L_\lambda(f) = L(f) + \frac{\lambda}{2}\|f\|^2, \qquad \widehat{L}_\lambda(f) = \widehat{L}(f) + \frac{\lambda}{2}\|f\|^2$$

$$\mathbf{H}(f) = \nabla^2 L(f), \qquad \mathbf{H}_\lambda(f) = \nabla^2 L_\lambda(f) = \mathbf{H}(f) + \lambda \mathbf{I}$$

$$\widehat{\mathbf{H}}(f) = \nabla^2 \widehat{L}(f), \qquad \widehat{\mathbf{H}}_\lambda(f) = \nabla^2 \widehat{L}_\lambda(f) = \widehat{\mathbf{H}}(f) + \lambda \mathbf{I}$$

Recall that $\widehat{f}_\lambda$ is defined as the minimizer of $\widehat{L}_\lambda$.

Define the following bounds on the second order derivatives:

$$\forall f \in \mathcal{H},\ \mathsf{b}_2(f) = \sup_{z \in \operatorname{supp}(Z)} \ell_z''(f(x)).$$

Statistical properties of the estimator The statistical properties of the estimator $\widehat{f}_\lambda$ have been studied in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) in the case of generalized self concordance, an are reported in the main lines in Sec. 4.H . The statistical rates of this estimator and the optimal choice of λ is determined by two parameters, defined in proposition 4.17 and which we adapt to the Kernel problem here.

- the *bias* $\mathsf{b}_\lambda = \|\mathbf{H}_\lambda(f^\star)^{-1/2}\nabla L_\lambda(f^\star)\| = \lambda\|f^\star\|_{\mathbf{H}_\lambda^{-1}(f^\star)}$, which characterizes the regularity of the optimum. The faster b_λ decreases to zero, the more regular $f^\star$ is.
- the *effective dimension*

$$\mathsf{df}_\lambda = \mathbb{E}\left[\|\mathbf{H}_\lambda(f^\star)^{-1/2}\nabla L_Z(f^\star)\|^2\right]. \tag{4.30}$$

  This quantity characterizes the size of the space $\mathcal{H}$ with respect to the problem; the slower it explodes as λ goes to zero, the smaller the size of $\mathcal{H}$.

For more complete explanations on the meaning of these quantities, we refer to Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019).

Moreover, as mentioned in proposition 4.17, one can define

$$\mathsf{B}_1^\star := \sup_{z\in\mathrm{supp}(Z)} \|\nabla L_z(f^\star)\|,\ \mathsf{B}_2^\star := \sup_{z\in\mathrm{supp}(Z)} \mathrm{Tr}(\nabla^2 L_z(f^\star)),\ \mathsf{Q}^\star = \frac{\mathsf{B}_1^\star}{\sqrt{\mathsf{B}_2^\star}},\ \mathsf{b}_2^\star = \mathsf{b}_2(f^\star). \tag{4.31}$$

We assume the following regularity condition on the minimizer $f^\star$, in order to get statistical bounds.

Assumption 4.4 (Source condition)**.** *There exists $r > 0$ and $g \in \mathcal{H}$ such that $f^\star = \mathbf{H}^r(f^\star)g$. This implies the following decrease rate of the bias:*

$$\mathsf{b}_\lambda \leq \mathsf{L}\lambda^{1/2+r}, \qquad \mathsf{L} = \|g\|_{\mathcal{H}}.$$

This is a stronger assumption than the existence of the minimizer as $r > 0$ is crucial for our analysis.

We also quantify the effective dimension df_λ: (however, since it always holds for $\alpha = 1$, this is not, strictly speaking, an additional assumption).

Assumption 4.5 (Effective dimension)**.** *The effective dimension decreases as $\mathsf{df}_\lambda \leq \mathsf{Q}\lambda^{-1/\alpha}$.*

If these two assumptions hold, define:

$$\beta = \frac{\alpha}{1+\alpha(1+2r)}, \qquad \gamma = \frac{(1+2r)\alpha}{1+\alpha(1+2r)}.$$

Under these assumptions, one can obtain the following statistical rates (which can be found in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) or in Cor. 4.4).

Proposition 4.9. *Let $\delta \in (0, 1/2]$. Under Assumptions 4.1 to 4.5, when $n \geq N$ and $\lambda = (C_0/n)^\beta$, then with probability at least $1 - 2\delta$,*

$$L(\widehat{f}_\lambda) - L(f^\star) \leq C_1 n^{-\gamma} \log\frac{2}{\delta},$$

with $C_0 = 256(\mathsf{Q}/\mathsf{L})^2$, $C_1 = 8(256)^\gamma\ (\mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma})^2$ and N defined in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), and satisfying $N = O(\mathrm{poly}(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, \mathsf{Q}, R, \log(1/\delta)))$.

4.D .2 Reducing the dimension: projecting on a subspace using Nyström sub-sampling.

Computations Using a representer theorem, one of the key properties of Kernel spaces is that, owing to the reproducing property,

$$\widehat{f}_\lambda \in \mathcal{H}_n := \left\{ \sum_{i=1}^n \alpha_i K_{x_i} \ : \ (\alpha_i) \in \mathbb{R}^n \right\}.$$

This means that solving the regularized empirical problem can be turned into a finite dimensional problem in α. Indeed $\widehat{f}_\lambda = \sum_{i=1}^n \alpha_i K_{x_i}$ where $\alpha = (\alpha_i)_{1 \leq i \leq n}$ is the solution to the following problem:

$$\alpha = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell_{z_i}(\alpha^\top \mathbf{K}_{nn} e_i) + \frac{\lambda}{2} \alpha^\top \mathbf{K}_{nn} \alpha, \qquad \mathbf{K}_{nn} = (K(x_i, x_j))_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}.$$

The previous problem is usually too costly to solve directly for large values of n, both in time and memory, because of the operations involving $\mathbf{K}_{nn}$. A solution consists in looking for a solution in a smaller dimensional sub-space $\mathcal{H}_M$ constructed from sub-samples of the data $\{\tilde{x}_1, ..., \tilde{x}_M\} \subset \{x_1, ..., x_n\}$:

$$\mathcal{H}_M := \left\{ \sum_{j=1}^M \tilde{\alpha}_j K_{\tilde{x}_j} \ : \ \tilde{\alpha} \in \mathbb{R}^M \right\}.$$

In this case, the minimizer $\widehat{f}_{M,\lambda} = \arg\min_{f \in \mathcal{H}_M} \widehat{L}_\lambda(f)$ can be written $\widehat{f}_{M,\lambda} = \sum_{j=1}^M \tilde{\alpha}_j K_{\tilde{x}_j}$, where $\tilde{\alpha}$ is the solution to the following problem:

$$\tilde{\alpha} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^n \ell_{z_i}(\alpha^\top \mathbf{K}_{Mn} e_i) + \frac{\lambda}{2} \alpha^\top \mathbf{K}_{MM} \alpha,$$

where

$$\mathbf{K}_{nM} = (K(x_i, \tilde{x}_j))_{\substack{1 \leq i \leq n \\ 1 \leq j \leq M}}, \ \mathbf{K}_{Mn} = \mathbf{K}_{nM}^\top, \ \mathbf{K}_{MM} := (K(\tilde{x}_i, \tilde{x}_j))_{1 \leq i,j \leq M}.$$

Let $\mathbf{T}$ be an upper triangular matrix such that $\mathbf{T}^\top \mathbf{T} = \mathbf{K}_{MM}$. One can re-parametrize the previous problem in the following way. For any $\beta \in \mathbb{R}^M$, define $f_\beta = \sum_{j=1}^M [\mathbf{T}^\dagger \beta]_j \ K_{\tilde{x}_j}$. This implies in particular that $\|f_\beta\|_{\mathcal{H}} = \|\beta\|_{\mathbb{R}^M}$. Then $\widehat{f}_{M,\lambda} = f_{\beta_{M,\lambda}}$, where

$$\beta_{M,\lambda} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^M} \widehat{L}_{M,\lambda}(\beta) := \frac{1}{n} \sum_{i=1}^n \ell_{z_i}(e_i^\top \mathbf{K}_{nM} \mathbf{T}^\dagger \beta) + \frac{\lambda}{2} \|\beta\|^2.$$

Using the properties the ℓ_z, one easily shows that $\beta \mapsto \ell_{z_i}(e_i^\top \mathbf{K}_{nM} \mathbf{T}^\dagger \beta)$ is $\left\{ \mathsf{R}_\ell \mathbf{T}^{-\top} \mathbf{K}_{Mn} e_i \right\}$ generalized self-concordant, and $\|\mathsf{R}_\ell \mathbf{T}^{-\top} \mathbf{K}_{Mn} e_i\| \leq \mathsf{R}_\ell \sqrt{K(x_i, x_i)}$. Thus, $\widehat{L}_M$ is also generalized self-concordant, and the associated $\widehat{\mathcal{G}}_M$ is bounded by $R = \mathsf{R}_\ell \kappa$. It will therefore be possible to apply the second order scheme presented in this paper to approximately compute $\beta_{M,\lambda}$.

Statistics Let $\widehat{\nu}_{\lambda,M}(\beta)$ denote the Newton decrement of $\widehat{L}_{\lambda,M}$ at point β and $\mathbf{P}_M$ denote the orthogonal projection on $\mathcal{H}_M$. Then the following statistical result shows that provided β is a good enough approximation of the optimum, and provided $\mathcal{H}_M$ is large enough, then f_β has the same generalization error as the empirical risk minimizer $\widehat{f}_\lambda$.

Recall the following result proved in proposition 4.19 in Sec. 4.H .3.

Proposition 4.10 (Behavior of an approximation to the projected problem)**.** *Suppose that Assumptions 4.1 to 4.3 are satisfied. Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Whenever*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \qquad \mathsf{C}_1 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n} \log \frac{2}{\delta}} \leq \frac{\lambda^{1/2}}{R}, \qquad \mathsf{C}_1 \mathsf{b}_\lambda \leq \frac{\lambda^{1/2}}{R},$$

if

$$\|\mathbf{H}^{1/2}(f^\star)(\mathbf{I} - \mathbf{P}_M)\|^2 \leq \lambda \frac{\sqrt{2}}{480}, \; 126\widehat{\nu}_{M,\lambda}(\beta) \leq \frac{\lambda^{1/2}}{R},$$

the following holds, with probability at least $1 - 2\delta$.

$$L(f_\beta) - L(f^\star) \leq \mathsf{K}_1 \; \mathsf{b}_\lambda^2 + \mathsf{K}_2 \; \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n} \; \log \frac{2}{\delta} + \mathsf{K}_3 \; \widehat{\nu}_{M,\lambda}^2(\beta), \qquad R\|f_\beta - f^\star\|_{\mathcal{H}} \leq 10,$$

where $\mathsf{K}_1 \leq$ 6.0e4, $\mathsf{K}_2 \leq$ 6.0e6 and $\mathsf{K}_3 \leq 810$, C_1 is defined in Lemma 4.19, and the other constants are defined in Theorem 4.8.

In particular, if we apply the previous result for a fixed λ, the following theorem holds (for a proof, see Sec. 4.H .4).

Theorem 4.5 (Quantitative result with source $r > 0$)**.** *Suppose that Assumptions 4.1 to 4.5 are satisfied. Let $n \geq N$ and $\delta \in (0, \frac{1}{2}]$. If $\lambda = \left(\left(\frac{\mathsf{Q}}{\mathsf{L}}\right)^2 \; \frac{1}{n} \right)^{\frac{\alpha}{\alpha(1+2r)+1}}$, and if*

$$\|\mathbf{H}^{1/2}(f^\star)(\mathbf{I} - \mathbf{P}_M)\|^2 \leq \lambda \frac{\sqrt{2}}{480}, \; \widehat{\nu}_{M,\lambda}(\beta) \leq \mathsf{Q}^\gamma \; \mathsf{L}^{1-\gamma} n^{-\gamma/2},$$

then with probability at least $1 - 2\delta$,

$$L(f_\beta) - L(f^\star) \leq \mathsf{K} \left(\mathsf{Q}^\gamma \; \mathsf{L}^{1-\gamma}\right)^2 \; \frac{1}{n^\gamma} \log \frac{2}{\delta}, \qquad R\|f_\beta - f^\star\| \leq 10,$$

where N is defined in Eq. (4.42) and $\mathsf{K} \leq$ 7.0e6.

The proof of the previous result is quite technical and can be found in Sec. 4.H , in Theorem 4.9.

4.D .3 A note on sub-sampling techniques

Let Z be a random variable on a Polish space $\mathcal{Z}$ and $(v_z)_{z \in \mathcal{Z}}$ be a family of vectors in $\mathcal{H}$ such that $||v||_{L^\infty(Z)} := \sup_{z \in \mathrm{supp}(Z)} \|v_z\| < \infty$ is bounded. Assume that $z_1, ..., z_n$ are i.i.d. samples from Z.

Define the following trace class Hermitian operators:

$$\mathbf{A} = \mathbb{E}\left[v_Z \otimes v_Z\right], \; \widehat{\mathbf{A}} = \frac{1}{n} \sum_{i=1}^n v_{z_i} \otimes v_{z_i}.$$

Define

$$\mathcal{N}^{\mathbf{A}}(\lambda) := \mathrm{Tr}(\mathbf{A}_\lambda^{-1}\mathbf{A}), \qquad \mathcal{N}_\infty^{\mathbf{A}}(\lambda) := \sup_{z\in \mathrm{supp}(Z)} \|\mathbf{A}_\lambda^{-1/2} v_z\|^2. \tag{4.32}$$

We typically have:

$$\mathcal{N}^{\mathbf{A}}(\lambda) \leq \mathcal{N}_\infty^{\mathbf{A}}(\lambda) \leq \frac{\|v\|^2_{L^\infty(Z)}}{\lambda}.$$

We define the leverage scores associated to the points z_i and $\mathbf{A}$:

$$\forall 1 \leq i \leq n,\ \forall t > 0,\ l_i^{\mathbf{A}}(t) = \|\widehat{\mathbf{A}}_t^{-1/2} v_{z_i}\|^2 = n\left((\mathbf{G}_{nn} + tn\mathbf{I})^{-1}\mathbf{G}_{nn}\right)_{ii}, \tag{4.33}$$

where $\mathbf{G}_{nn} = (v_{z_i} \cdot v_{z_j})_{1\leq i,j\leq n}$ denotes the Gram matrix associated to the family v_{z_i}.

As in Rudi, Camoriano, and Rosasco (2015), definition 1, we give the following definition for leverage scores.

Definition 4.5 (q-approximate leverage scores)**.** *given t_0, a family $(\tilde{l}_i^{\mathbf{A}}(t))_{1\leq i\leq n}$ is said to be a family of q-approximate leverage scores with respect to $\mathbf{A}$ if*

$$\forall 1 \leq i \leq n,\ \forall t \geq t_0,\ \frac{1}{q}\, l_i^{\mathbf{A}}(t) \leq \tilde{l}_i^{\mathbf{A}}(t) \leq q\, l_i^{\mathbf{A}}(t).$$

We say that a subset of m points $\{\tilde{z}_1, ..., \tilde{z}_m\} \subset \{z_i \ :\ 1 \leq i \leq n\}$ is:

- **Sampled using q-approximate leverage scores for** t if the $\tilde{z}_j = z_{i_j}$ where the i_j are m i.i.d. samples from $\{1, ..., n\}$ using the probability vector $p_i = \frac{\tilde{l}_i^{\mathbf{A}}(t)}{\sum_{i=1}^n \tilde{l}_i^{\mathbf{A}}(t)}$. In that case, we define $\widehat{\mathbf{A}}_m := \frac{1}{m}\sum_{j=1}^m \frac{1}{np_{i_j}} v_{\tilde{z}_j} \otimes v_{\tilde{z}_j}$.
- **Sampled uniformly** if the $\{i_j \ :\ 1 \leq j \leq m\}$ is a uniformly chosen subset of $\{1, ..., n\}$ of size m. In this case, we define $\widehat{\mathbf{A}}_m := \frac{1}{m}\sum_{j=1}^m v_{\tilde{z}_j} \otimes v_{\tilde{z}_j}$.

In Sec. 4.I .1, we present technical lemmas which allow us to show that if m is large enough, the following hold:

- $\|\mathbf{A}_\eta(\mathbf{I} - \mathbf{P}_m)\|^2 \leq 3\eta$, where $\mathbf{P}_m$ is the orthogonal projection on the subspace induced by the $v_{\tilde{z}_j}$;
- $\widehat{\mathbf{A}}_{m,\lambda}$ is equivalent to $\widehat{\mathbf{A}}_\lambda$.

Remark 11 (cost of computing q-approximate leverage scores)**.** *In Rudi, Calandriello, Carratino, and Rosasco (2018), one can show that the complexity of computing q-approximate leverage scores can be achieved in: $\mathtt{c}_{samp} = O(q^2\mathcal{N}^{\mathbf{A}}(\lambda)^2 \min(n, 1/\lambda))$ time (where a unit of time is a scalar product evaluation) and $O(\mathcal{N}^{\mathbf{A}}(\lambda)^2 + n)$ in memory.*

4.D .4 Selecting the M Nyström points

In order for Theorem 4.5 to hold, we must subsample the M points such as to guarantee $\|\mathbf{H}^{1/2}(f^\star)(\mathbf{I} - \mathbf{P}_M)\|^2 \leq \frac{\sqrt{2}\lambda}{480}$.

Since we must sub-sample the M points a priori, i.e. before performing the method, it is necessary to have sub-sampling schemes which do not depend heavily on the point. Define the covariance operator:

$$\mathbf{\Sigma} = \mathbb{E}\left[K_X \otimes K_X\right].$$

Since $\mathbf{H}(f^\star) = \mathbb{E}\left[\ell''_Z(f(X))\ K_X \otimes K_X\right]$, it is easy to see that $\mathbf{H}(f^\star) \preceq \mathsf{b}_2^\star \boldsymbol{\Sigma}$. Note that for $\boldsymbol{\Sigma}$, since $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}$, the leverage scores have the following form:

$$\forall 1 \le i \le n,\ l_i^{\boldsymbol{\Sigma}}(t) = n\left((\mathbf{K}_{nn} + \lambda n \mathbf{I})^{-1}\mathbf{K}_{nn}\right)_{ii}.$$

Proposition 4.11 (Selecting Nyström points)**.** *Let $\delta > 0$. Let $\eta = \min(\|\boldsymbol{\Sigma}\|, \frac{\lambda\sqrt{2}}{1440(\mathsf{b}_2^\star \vee 1)})$. Assume the samples $\{\tilde{x}_1, ..., \tilde{x}_M\}$ are obtained with one of the following.*
1. *$n \ge M \ge \left(10 + 160\mathcal{N}_\infty^{\boldsymbol{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ using uniform sampling;*
2. *$M \ge \left(6 + 486q^2\mathcal{N}^{\boldsymbol{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ using q-approximate leverage scores with respect to $\boldsymbol{\Sigma}$ for $t = \eta$, $t_0 \vee \frac{19\kappa^2}{n}\log\frac{n}{2\delta} < \eta$, $n \ge 405\kappa^2 \vee 67\kappa^2\log\frac{12\kappa^2}{\delta}$.*
Then it holds, with probability at least $1-\delta$:

$$\|\boldsymbol{\Sigma}_\eta^{1/2}(\mathbf{I} - \mathbf{P}_M)\| \le 3\eta \implies \|\mathbf{H}^{1/2}(f^\star)(\mathbf{I} - \mathbf{P}_M)\|^2 \le \lambda\frac{\sqrt{2}}{480}.$$

Proof. The proof is a direct application of the lemmas in Sec. 4.I .1. Indeed, note that since $\boldsymbol{\Sigma} = \mathbb{E}\left[K_X \otimes K_X\right]$, then the results can be applied with $Z \leftarrow X$ and $v_z \leftarrow K_x$. Indeed, from Assumption 4.2, it holds:

$$\sup_{x \in \mathrm{supp}(X)} \|K_x\|^2 \le \kappa^2.$$

□

We can now combine proposition 4.11 and proposition 4.10 to obtain the following statistical bounds for the optimizer of the projected Nyström problem $\beta_{M,\lambda}$.

Theorem 4.6. *Suppose that Assumptions 4.1 to 4.3 are satisfied. Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \le \mathsf{B}_2^\star \wedge 720\sqrt{2}(\mathsf{b}_2^\star \vee 1)\|\boldsymbol{\Sigma}\|$. Assume*

$$n \ge \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda}\log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}, \qquad \mathsf{C}_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n}\ \log\frac{2}{\delta}} \le \frac{\lambda^{1/2}}{R}, \qquad \mathsf{C}_1\mathsf{b}_\lambda \le \frac{\lambda^{1/2}}{R},$$

Let $\eta = \frac{\lambda\sqrt{2}}{1440(\mathsf{b}_2^\star \vee 1)}$. Assume the samples $\{\tilde{x}_1, ..., \tilde{x}_M\}$ are obtained with one of the following.
1. *$n \ge M \ge \left(10 + 160\mathcal{N}_\infty^{\boldsymbol{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ using uniform sampling;*
2. *$M \ge \left(6 + 486q^2\mathcal{N}^{\boldsymbol{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ using q-approximate leverage scores with respect to $\boldsymbol{\Sigma}$ for $t = \eta$, $t_0 \vee \frac{19\kappa^2}{n}\log\frac{n}{2\delta} < \eta$, $n \ge 405\kappa^2 \vee 67\kappa^2\log\frac{12\kappa^2}{\delta}$.*
The following holds, with probability at least $1 - 3\delta$.

$$L(f_{\beta_{M,\lambda}}) - L(f^\star) \le \mathsf{K}_1\ \mathsf{b}_\lambda^2 + \mathsf{K}_2\ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n}\ \log\frac{2}{\delta}, \qquad R\|\beta_{M,\lambda}\| \le R\|f^\star\| + 10,$$

where $\mathsf{K}_1 \le$ 6.0e4, $\mathsf{K}_2 \le$ 6.0e6 and $\mathsf{K}_3 \le 810$, C_1 is defined in Lemma 4.19, and the other constants are defined in Theorem 4.8.

Proof. This is simply a reformulation of proposition 4.10, noting that $\widehat{\nu}_{M,\lambda}(\beta_{M,\lambda}) = 0$ and that proposition 4.11 implies the condition on the Hessian at the optimum. □

Provided source condition holds with $r > 0$, the conditions of this theorem are not void.

4.D .5 Performing the globalization scheme to approximate $\beta_{M,\lambda}$

In order to apply proposition 4.10, one needs to control $\widehat{\nu}_{M,\lambda}(\beta)$.

We will apply our general scheme to $\widehat{L}_{M,\lambda}$ in order to obtain such a control.

Performing approximate Newton steps

The key element in the globalization scheme is to be able to compute $\frac{1}{7}$-approximate Newton steps.

Note that at a given point β and for a given $\mu > 0$ the Hessian is of the form:

$$\widehat{\mathbf{H}}_{M,\mu}(\beta) = \frac{1}{n}\mathbf{T}^{-\top}\mathbf{K}_{Mn}\mathbf{D}_n(\beta)\mathbf{K}_{nM}\mathbf{T}^{-1} + \mu\mathbf{I}_M,$$

where $\mathbf{D}_n(\beta) = \operatorname{diag}((d_i(\beta))_{1\leq i\leq n})$ is a diagonal matrix whose elements are given by $d_i(\beta) = \ell''_{z_i}(e_i^\top \mathbf{K}_{nM}\mathbf{T}^{-1}\beta)$.

Note that we can always write

$$\widehat{\mathbf{H}}_{M,\mu}(\beta) = \frac{1}{n}\sum_{i=1}^{n} u_i(\beta)u_i(\beta)^\top + \mu\mathbf{I}, \qquad u_i(\beta) = \sqrt{d_i(\beta)}\mathbf{T}^{-\top}\mathbf{K}_{Mn}e_i$$

The gradient can be put in the following form:

$$\nabla\widehat{L}_{M,\mu}(\beta) = \frac{1}{n}\mathbf{T}^{-\top}\mathbf{K}_{Mn}v + \mu\beta, \qquad v = (\ell'_{z_i}(e_i^\top \mathbf{K}_{nM}\mathbf{T}^{-1}\beta))_{1\leq i\leq n}.$$

Computing the gradient at one point therefore costs $O(nM + M^2)$, this being the cost of computing $\mathbf{K}_{nM}$ times a vector costs $O(nM)$ and computing $\mathbf{T}^{-1}$ times a vector takes $O(M^2)$ since $\mathbf{T}$ is triangular. Moreover, the cost in memory is $O(M^2 + n)$, M^2 being needed for the saving of $\mathbf{T}$ and n for the saving of the gradient; $\mathbf{K}_{nM}$ times a vector can also be done in $O(n)$ memory, provided we compute it by blocks.

On the other hand, computing the full Hessian matrix would cost nM^2 operations, which is un-tractable. However, computing a Hessian vector product can be done in $O(nM + M^2)$ time, as for the gradient, which suggest using an iterative solver with preconditioning.

Computing $x \in \operatorname{LinApprox}(\mathbf{A}, b, \rho)$ through pre-conditioned conjugate gradient descent. Assume we wish to solve the problem $\mathbf{A}x = b$ where $\mathbf{A} \in \mathbb{R}^{M\times M}$ is a positive definite matrix and b is a vector of $\mathbb{R}^M$. If one uses the conjugate gradient method starting from zero, then if x_k denotes the k-the iterate of the conjugate gradient algorithm, Theorem 6.6 in Saad (2003) shows that

$$x_k \in \operatorname{LinApprox}(\mathbf{A}, b, \rho), \qquad \rho = 2\left(\frac{\sqrt{\operatorname{Cond}(\mathbf{A})} - 1}{\sqrt{\operatorname{Cond}(\mathbf{A})} + 1}\right)^k.$$

where $\operatorname{Cond}(\mathbf{A})$ is the condition number of the matrix $\mathbf{A}$, namely the ratio $\frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$. If $\operatorname{Cond}(\mathbf{A})$ is large, this convergence can be very slow. The idea of preconditioning is to compute an

approximation matrix $\widetilde{\mathbf{A}}$ such that

$$\frac{1}{2}\widetilde{\mathbf{A}} \preceq \mathbf{A} \preceq \frac{3}{2}\widetilde{\mathbf{A}}. \tag{4.34}$$

We then compute $\mathbf{B}$ a triangular matrix such that $\mathbf{B}^\top\mathbf{B} = \widetilde{\mathbf{A}}$ using a cholesky decomposition, which can be done in $O(M^3)$, and note that $\mathbf{B}^{-\top}\mathbf{A}\mathbf{B}^{-1}$ is very well conditioned; indeed, its condition number is bounded by 3.

Perform a conjugate gradient method to solve the pre-conditioned problem $\mathbf{B}^{-\top}\mathbf{A}\mathbf{B}^{-1}z = \mathbf{B}^{-\top}b$, and denote with z_τ the τ-th iteration of this method. Then using the bound on the condition number, we find

$$z_\tau \in \mathrm{LinApprox}(\mathbf{B}^{-\top}\mathbf{A}\mathbf{B}^{-1}, \mathbf{B}^{-\top}b, \rho), \qquad \rho = 2\left(\frac{\sqrt{3}-1}{\sqrt{3}+1}\right)^\tau,$$

which in turn implies that by setting $x_\tau := \mathbf{B}^{-1}z_\tau$,

$$x_\tau \in \mathrm{LinApprox}(\mathbf{A}, b, \rho),\ \rho = 2\left(\frac{\sqrt{3}-1}{\sqrt{3}+1}\right)^\tau.$$

This shows that after at most $\tau = 3$ iterations, provided $\widetilde{\mathbf{A}}$ satisfies Eq. (4.34), $x_\tau \in \mathrm{LinApprox}(\mathbf{A}, b, \frac{1}{7})$. The cost of this method is therefore $O(M^3 + nM)$ in time, and $O(n + M^2)$ due to the computing of the preconditioner and computing matrix vector products by block. This does not include the cost of finding a suitable $\widetilde{\mathbf{A}}$.

Computing a suitable approximation of $\widehat{\mathbf{H}}_{M,\mu}(\beta)$ To compute a good pre-conditioner, we will subsample Q points $i_1, ..., i_Q$ points from $\{1, ..., n\}$, and sketch the Hessian using these Q points.

Proposition 4.12 (Computing approximate newton steps)**.** *Let $\delta > 0$. Let $\beta \in \mathbb{R}^M$ and $\mu \geq \lambda$, and assume $\frac{19\mathsf{b}_2(f_\beta)\kappa^2}{n}\log\frac{n}{2\delta} < \lambda$ and $n \geq 405\mathsf{b}_2(f_\beta)\kappa^2 \vee 67\mathsf{b}_2(f_\beta)\kappa^2\log\frac{12\mathsf{b}_2(f_\beta)\kappa^2}{\delta}$. Let $\tilde{\mu} = \min(\mu, \|\mathbf{H}(f_\beta)\|)$. Assume one of the following properties is satisfied*

1. *$Q \geq \left(10 + 160\mathcal{N}_\infty^{\mathbf{H}(f_\beta)}(\tilde{\mu})\right)\log\frac{8\mathsf{b}_2(f_\beta)\kappa^2}{\tilde{\mu}\delta}$ with uniform sampling of the $\{i_1, ..., i_Q\}$. We set $\mathbf{D}_Q = \mathrm{diag}(\ell''_{z_{i_j}}(f_\beta(x_{i_j})))_{1\leq j\leq Q}$*

2. *$Q \geq \left(6 + 486q^2\mathcal{N}^{\mathbf{H}(f_\beta)}(\tilde{\mu})\right)\log\frac{8\mathsf{b}_2(f_\beta)\kappa^2}{\tilde{\mu}\delta}$ using q-approximate leverage scores associated to $\mathbf{H}(f_\beta)$ for $t = \tilde{\mu}$. We set $\mathbf{D}_Q = \mathrm{diag}\left(\frac{\ell''_{z_{i_j}}(f_\beta(x_{i_j}))}{p_{i_j}}\right)$, where the p_{i_j} are the probabilities computed from the leverage scores.*

Assume we use a pre-conditioner $\mathbf{B}$ such that

$$\mathbf{B}^\top\mathbf{B} = \frac{1}{Q}\mathbf{T}^{-\top}\mathbf{K}_{MQ}\mathbf{D}_Q\mathbf{K}_{QM}\mathbf{T}^{-1} + \mu\mathbf{I}_M, \qquad \mathbf{K}_{QM} = (K(x_{i_j}, \tilde{x}_k))_{\substack{1\leq j\leq Q\\ 1\leq k\leq M}}.$$

If we perform $\tau = \log(\rho/2)/\log((\sqrt{3}+1)/\sqrt{3}-1)$ iterations of the conjugate gradient descent on the pre-conditioned Newton system using $\mathbf{B}$ as a preconditioner, then with probability at least

$1-\delta$, this procedure is returns $\widetilde{\Delta} \in \mathrm{LinApprox}(\widehat{\mathbf{H}}_{M,\lambda}(\beta), \nabla \widehat{L}_{M,\lambda}(\beta), \rho)$, and the computational time is of order $O(\tau(Mn + M^2Q + M^3 + \mathsf{c}_{samp}))$, and the memory requirements can be reduced to $O(M^2 + n)$. Here c_{samp} stands for the complexity of computing Nystrom leverage scores, and using Remark 11 or Rudi, Calandriello, Carratino, and Rosasco (2018), $\mathsf{c}_{samp} = O(1)$ if uniform sampling is used, and $\mathsf{c}_{samp} = O(\mathcal{N}^{\mathbf{H}(f_\beta)}(\tilde{\mu})^2/\lambda)$ if Nystrom sub-sampling is used. Note that for $\tau = 3$, $\rho = \frac{1}{7}$.

Proof. Start by defining the following operators:

- $K_n : f \in \mathcal{H} \to (f(x_i))_{1 \le i \le n} \in \mathbb{R}^n$;
- $K_M : f \in \mathcal{H} \to (f(\tilde{x}_j))_{1 \le j \le M} \in \mathbb{R}^M$;
- $V = K_M^* \mathbf{T}^{-1}$, where $\mathbf{T}$ is an upper triangular matrix such that $\mathbf{T}^\top \mathbf{T} = \mathbf{K}_{MM} = K_M K_M^*$.

Note that $K_n V = \mathbf{K}_{nM} \mathbf{T}^{-1}$.

Now note that

$$\forall f \in \mathcal{H},\ \mathbf{H}(f) = \mathbb{E}\left[v_z \otimes v_z\right], \qquad \widehat{\mathbf{H}}(f) = \frac{1}{n}\sum_{i=1}^{n} v_{z_i} \otimes v_{z_i}, \qquad v_z = \sqrt{\ell_z''(f(x))} K_x.$$

Since for any $f \in \mathcal{H}$, $\widehat{\mathbf{H}}(f) = \frac{1}{n} K_n^* \mathbf{D}_n(f) K_n$, where $\mathbf{D}_n(f) = \mathrm{diag}(\ell_{z_i}''(f(x_i)))$, we see that

$$\widehat{\mathbf{H}}_{M,\mu}(\beta) = V^* \widehat{\mathbf{H}}(f_\beta) V + \mu \mathbf{I}_M.$$

Thus, the last lemma of Sec. 4.I .1 can be applied, using the fact that $\|v_z\|^2 \le \mathsf{b}_2(f)\kappa^2$, to get that in both cases of the proposition, under the corresponding assumptions:

$$\frac{1}{2}\left(\frac{1}{Q}\mathbf{T}^{-\top}\mathbf{K}_{MQ}\mathbf{D}_Q\mathbf{K}_{QM}\mathbf{T}^{-1} + \mu\mathbf{I}_M\right) \preceq \widehat{\mathbf{H}}_{M,\mu}(\beta) \preceq \frac{3}{2}\left(\frac{1}{Q}\mathbf{T}^{-\top}\mathbf{K}_{MQ}\mathbf{D}_Q\mathbf{K}_{QM}\mathbf{T}^{-1} + \mu\mathbf{I}_M\right).$$

The rest of the proposition follows from the previous discussion.

□

Applying the globalization scheme to control $\widehat{\nu}_{M,\lambda}(\beta)$

In order to apply proposition 4.12 to each point β in our method, we need to have a globalized version of the condition of this proposition.

First, we start by localizing the different values of β we will visit throughout the algorithm.

Definition 4.6 (path of regularized solutions)**.** *Let $\lambda > 0$, $\varepsilon > 0$. Define the path of regularized solutions*

$$\widehat{\Gamma}_\lambda^M := \{\beta_{M,\mu} \ :\ \mu \ge \lambda\}. \tag{4.35}$$

And the ε approximation of this path:

$$\widehat{\Gamma}_{\lambda,\varepsilon}^M := \left\{\beta \in \mathbb{R}^M \ :\ d(\beta, \widehat{\Gamma}_\lambda^M) \le \varepsilon\right\}. \tag{4.36}$$

Note that we always have $\widehat{\Gamma}^M_\lambda \subset \mathcal{B}_{\mathbb{R}^M}(\|\beta_{M,\lambda}\|)$. We now state a lemma proving that all the values visited during the algorithm will lie in an approximation of this path.

Lemma 4.15. *Define Let $\beta \in \mathbb{R}^M$ such that $\widehat{\nu}_{M,\mu}(\beta) \leq \frac{\mu^{1/2}}{7R}$ for some $\mu \geq \lambda$. Then the following holds:*

$$\beta \in \widehat{\Gamma}^M_{\lambda,\frac{1}{6R}}.$$

Proof. Bound

$$R\|\beta - \beta_{M,\mu}\| \leq \frac{R}{\mu^{1/2}}\|\beta - \beta_{M,\mu}\|_{\widehat{\mathbf{H}}_{M,\mu}(\beta)} \leq \frac{1}{\underline{\phi}(\mathsf{t}_M(\beta - \beta_{M,\mu}))}\frac{R\widehat{\nu}_{M,\mu}(\beta)}{\mu^{1/2}}.$$

Just apply Eq. (4.18) to obtain $R\|\beta - \beta_{M,\mu}\| \leq \frac{1}{6}$. □

We now introduce the following quantities which will allow to control the number of sub-samples throughout the whole algorithm.

Definition 4.7. *Define*

- $\overline{\mathsf{b}}_2 := \sup_{\beta \in \widehat{\Gamma}^M_{\lambda,1/6R}} \mathsf{b}_2(f_\beta)$.
- $\overline{\mathcal{N}}^{\mathbf{H}}(\lambda) = \sup_{\beta \in \widehat{\Gamma}^M_{\lambda,1/6R}} \mathcal{N}^{\mathbf{H}(f_\beta)}(\lambda)$.
- $\overline{\mathcal{N}}^{\mathbf{H}}_\infty(\lambda) = \sup_{\beta \in \widehat{\Gamma}^M_{\lambda,1/6R}} \mathcal{N}^{\mathbf{H}(f_\beta)}_\infty(\lambda)$.
- $\overline{\|\mathbf{H}\|} = \min_{\beta \in \widehat{\Gamma}^M_{\lambda,1/6R}} \|\mathbf{H}(f_\beta)\|$.

Proposition 4.13 (Performance of the globalization scheme)**.** *Let $\varepsilon > 0$, $\delta > 0$, $\tilde{\lambda} = \min(\lambda, \overline{\|\mathbf{H}\|})$. Assume $\frac{19\overline{\mathsf{b}}_2\kappa^2}{n}\log\frac{n}{2\delta} < \tilde{\lambda}$ and $n \geq 405\overline{\mathsf{b}}_2\kappa^2 \vee 67\overline{\mathsf{b}}_2\kappa^2\log\frac{12\overline{\mathsf{b}}_2\kappa^2}{\delta}$.*

Assume we perform the globalization scheme with the parameters in Theorem 4.1, where in order to compute any ρ approximation of a regularized Newton step, we use a conjugate gradient descent on the pre-conditioned system, where the pre-conditioner is computed as in proposition 4.12 using

1. $Q \geq \left(10 + 160\overline{\mathcal{N}}^{\mathbf{H}}_\infty(\tilde{\lambda})\right)\log\frac{8\overline{\mathsf{b}}_2\kappa^2}{\tilde{\lambda}\delta}$ *if using uniform sampling*
2. $Q \geq \left(6 + 486q^2\overline{\mathcal{N}}^{\mathbf{H}}(\tilde{\lambda})\right)\log\frac{8\overline{\mathsf{b}}_2\kappa^2}{\tilde{\lambda}\delta}$ *if using Nyström leverage scores*

Recall that t denotes the number of approximate Newton steps performed at for each μ in Phase I and T denotes the number of approximate Newton steps performed in Phase II, and that using Theorem 4.1, $t = 2$ and $T = \lceil\log_2\sqrt{1 \vee (\lambda\varepsilon^{-1}/R^2)}\rceil$. Moreover, recall that K denotes the number of steps performed in Phase I. Define

$$N_{ns} := 2\left\lfloor(3 + 11R\|\beta_{M,\lambda}\|)\log_2(7R\|\nabla\widehat{L}_M(0)\|/\lambda)\right\rfloor + \lceil\log_2\sqrt{1 \vee (\lambda\varepsilon^{-1}/R^2)}\rceil.$$

Then with probability at least $(1-\delta)^{N_{ns}}$:

- *The method presented in proposition 4.12 returns a 1/7- approximate Newton step at each time it is called in the algorithm.*
- *If β denotes the result of the method, $\widehat{\nu}_{M,\lambda}(\beta) \leq \sqrt{\varepsilon}$.*

- *The number of approximate Newton steps computed during the algorithm is bounded by N_{ns}; the complexity of the method is therefore of order $O(N_{ns}(M^2\max(M,Q)+nM+\mathsf{c}_{samp}(\lambda)))$ in time and $O(MQ+M^2+n)$ in memory, where $\mathsf{c}_{samp}(\lambda)$ is a bound on the complexity associated to the computing of leverage scores (see Rudi, Calandriello, Carratino, and Rosasco (2018) for details).*

The algorithm is detailed in Sec. 4.E , in algorithm 1. Note however that the notations are those of the main paper, which are slightly different from the ones used here.

Proof. If we take the globalization scheme, using the parameters of Theorem 4.1. Assume that all previous approximate Newton steps have been computed in a good way. Then the β at which we are belongs to $\widehat{\Gamma}^M_{\lambda,1/6R}$. Thus, the hypotheses of this proposition imply that the hypothesis of proposition 4.12 are satisfied; and hence, up to a $(1-\delta)$ probability factor, we can assume that the next approximate Newton step is performed correctly, continuing the globalization scheme in the right way. Thus, the globalization scheme converges as in Theorem 4.1. □

4.D .6 Statistical properties of the algorithm

The following theorem describes the computational and statistical behavior of our algorithm.

Proposition 4.14 (Behavior of an approximation to the projected problem)**.** *Suppose that Assumptions 4.1 to 4.3 are satisfied.*
Let $n\in\mathbb{N}$, $\varepsilon>0$, $\delta\in(0,1/2]$, $0<\lambda\leq \mathsf{B}_2^\star$.
Define $\tilde{\lambda}=\min(\lambda,\overline{\|\mathbf{H}\|})$ and assume $\frac{19\bar{\mathsf{b}}_2\kappa^2}{n}\log\frac{n}{2\delta}<\tilde{\lambda}$, $n\geq 405\bar{\mathsf{b}}_2\kappa^2\vee 67\bar{\mathsf{b}}_2\kappa^2\log\frac{12\bar{\mathsf{b}}_2\kappa^2}{\delta}$, and $n\geq\triangle_1\frac{\mathsf{B}_2^\star}{\lambda}\log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}$. Assume

$$\mathsf{C}_1\sqrt{\frac{\mathsf{df}_\lambda\vee(\mathsf{Q}^\star)^2}{n}\log\frac{2}{\delta}}\leq\frac{\lambda^{1/2}}{R},\qquad \mathsf{C}_1\mathsf{b}_\lambda\leq\frac{\lambda^{1/2}}{R},\qquad 126\sqrt{\varepsilon}\leq\frac{\lambda^{1/2}}{R}.$$

Assume that the M points $\tilde{x}_1,\dots,\tilde{x}_M$ are obtained through Nyström sub-sampling using $\eta=\|\mathbf{\Sigma}\|\wedge\frac{\lambda\sqrt{2}}{1440(\mathsf{b}_2^\star\vee 1)}$, with either
1. *$M\geq\left(10+160\mathcal{N}_\infty^{\mathbf{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ if using uniform sampling;*
2. *$M\geq\left(6+486q^2\mathcal{N}^{\mathbf{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ if using q-approximate leverage scores for η, associated to the co-variance operator $\mathbf{\Sigma}$.*

Assume we perform the globalization scheme as in proposition 4.13, i.e. with the parameters in Theorem 4.1, where in order to compute any ρ approximation of a regularized Newton step, we use a conjugate gradient descent on the pre-conditioned system, where the pre-conditioner is computed as in proposition 4.12 using
1. *$Q\geq\left(10+160\overline{\mathcal{N}}_\infty^{\mathbf{H}}(\tilde{\lambda})\right)\log\frac{8\bar{\mathsf{b}}_2\kappa^2}{\tilde{\lambda}\delta}$ if using uniform sampling*
2. *$Q\geq\left(6+486q^2\overline{\mathcal{N}}^{\mathbf{H}}(\tilde{\lambda})\right)\log\frac{8\bar{\mathsf{b}}_2\kappa^2}{\tilde{\lambda}\delta}$ if using Nyström leverage scores*
Let N_{ns} be defined as in proposition 4.13. Recall N_{ns} is an upper bound for the number of approximate Newton steps performed in the algorithm. One can bound

$$N_{ns}\leq 2\left\lfloor(113+11R\|f^\star\|)\log_2\frac{7R\|\nabla\widehat{L}_M(0)\|}{\lambda}\right\rfloor+\left\lceil\log_2\frac{\lambda^{1/2}}{R\varepsilon}\right\rceil.$$

Moreover, with probability at least $1-(N_{ns}+2)\delta$, the following holds:

$$L(f_\beta)-L(f^\star)\leq\mathsf{K}_1\,\mathsf{b}_\lambda^2+\mathsf{K}_2\,\frac{\mathsf{df}_\lambda\vee(\mathsf{Q}^\star)^2}{n}\log\frac{2}{\delta}+\mathsf{K}_3\,\varepsilon.$$

where $\mathsf{K}_1 \leq 6.0\text{e}4$, $\mathsf{K}_2 \leq 6.0\text{e}6$ *and* $\mathsf{K}_3 \leq 810$, C_1 *is defined in Lemma 4.19, and the other constants are defined in Theorem 4.8.*

Proof. This is a simple combination between propositions 4.10, 4.11 and 4.13. To bound the number of Newton steps N_{ns}, one simply uses the fact that under the conditions of the theorem, $R\|\beta_{M,\lambda}\| \leq 10 + R\|f^\star\|$. □

Remark 12 (Complexity)**.** *Let* $L = \bar{\mathsf{b}}_2\kappa^2$. *The complexity of the previous method using leverage scores computed for* $\boldsymbol{\Sigma}$ *for the Nystrom projections and for* $\mathbf{H}(f_\beta)$ *for choosing the* Q *points at the different stages is the following. The total complexity in time will be of order:*

$$O\left(N_{ns}\left(n\mathcal{N}^{\mathbf{H}}(\lambda)\log(L\lambda^{-1}\delta^{-1}) + \bar{\mathsf{b}}_2^3\mathcal{N}^{\boldsymbol{\Sigma}}(\lambda)^3\log^3(L\lambda^{-1}\delta^{-1}) + L/\lambda\ \bar{\mathsf{b}}_2^2\mathcal{N}^{\boldsymbol{\Sigma}}(\lambda)^2\right)\right).$$

The memory complexity can be bounded by

$$O(\bar{\mathsf{b}}_2^2\mathcal{N}^{\boldsymbol{\Sigma}}(\lambda)^2\log^2(L\lambda^{-1}\delta^{-1}) + n).$$

Here, we use the fact that $\mathbf{H} \leq \bar{\mathsf{b}}_2\boldsymbol{\Sigma}$.

We can now write down the previous proposition by classifying problems using Assumptions 4.4 and 4.5 and in order to get optimal rates.

Theorem 4.7 (Performance of the scheme using pre-conditioning)**.** *Let* $\delta > 0$. *Assume Assumptions 4.1 to 4.5 are satisfied. Let* $n \geq \tilde{N}$, *where* $\tilde{N}$ *is characterized in the proof,* $\lambda = \left(\left(\frac{\mathsf{Q}}{\mathsf{L}}\right)^2 \frac{1}{n}\right)^{\frac{\alpha}{\alpha(1+2r)+1}}$.
Assume that the M *points* $\tilde{x}_1, ..., \tilde{x}_M$ *are obtained through Nyström sub-sampling using* $\eta = \frac{\lambda\sqrt{2}}{1440(\mathsf{b}_2^\star \vee 1)}$, *with either*

1. $M \geq \left(10 + 160\mathcal{N}_\infty^{\boldsymbol{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ *if using uniform sampling;*

2. $M \geq \left(6 + 486q^2\mathcal{N}^{\boldsymbol{\Sigma}}(\eta)\right)\log\frac{8\kappa^2}{\eta\delta}$ *if using* q*-approximate leverage scores for* η, *associated to the co-variance operator* $\boldsymbol{\Sigma}$.

Assume we perform the globalization scheme as in proposition 4.13, i.e. with the parameters in Theorem 4.1, where in order to compute any ρ *approximation of a regularized Newton step, we use a conjugate gradient descent on the pre-conditioned system, where the pre-conditioner is computed as in proposition 4.12 using*

1. $Q \geq \left(10 + 160\overline{\mathcal{N}}_\infty^{\mathbf{H}}(\lambda)\right)\log\frac{8\bar{\mathsf{b}}_2\kappa^2}{\lambda\delta}$ *if using uniform sampling*

2. $Q \geq \left(6 + 486q^2\overline{\mathcal{N}}^{\mathbf{H}}(\lambda)\right)\log\frac{8\bar{\mathsf{b}}_2\kappa^2}{\lambda\delta}$ *if using Nyström leverage scores*

Let N_{ns} *be defined as in proposition 4.13. Recall* N_{ns} *is an upper bound for the number of approximate Newton steps performed in the algorithm. One can bound*

$$N_{ns} \leq (227 + 22R\|f^\star\|)\left(\left\lceil\log_2\left(7R\|\nabla\widehat{L}_M(0)\|\right)\right\rceil + \left\lceil\log_2\frac{n\mathsf{L}^2}{\mathsf{Q}^2}\right\rceil + \left\lceil\log_2\frac{1}{R\mathsf{L}}\right\rceil\right).$$

Moreover, with probability at least $1 - (N_{ns} + 2)\delta$, *the following holds:*

- *all of the approximate Newton methods yield* $\frac{1}{7}$*-approximate Newton steps*
- *The scheme finishes, and the number of approximate Newton steps is bounded by* N_{ns}. *The total complexity of the method is therefore*

$$O((nM + M^3 + M^2Q + \mathsf{c}_{samp})N_{ns}) \text{ in time }, \qquad O(n + M^2) \text{ in memory.}$$

- *The returned β is statistically optimal:*

$$L(f_\beta) - L(f^\star) \leq \mathsf{K}\left(\mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma}\right)^2\ \frac{1}{n^\gamma}\log\frac{2}{\delta},$$

where K is defined in Theorem 4.5.

Proof. The proof consists mainly of combining propositions 4.11 and 4.13 and Theorem 4.5.

Recall that we set $\lambda = \left(\frac{\mathsf{Q}^2}{\mathsf{L}^2}\frac{1}{n}\right)^{\frac{\alpha}{\alpha(1+2r)+1}}$.

1. Start by defining $\tilde{N}$ such that:

- $\tilde{N} \geq N$ where N is defined in Theorem 4.5;
- $\forall n \geq \tilde{N},\ \lambda \leq \overline{\|\mathbf{H}\|}$. This is possible as $\frac{\alpha}{\alpha(1+2r)+1}$ is a strictly positive exponent.
- $\forall n \geq \tilde{N},\ \frac{19\overline{\mathsf{b}}_2 \vee 1\ \kappa^2}{n}\log\frac{n}{2\delta} < \lambda$; this is possible as soon as $\frac{\alpha}{\alpha(1+2r)+1} < 1$, i.e. this is satisfied since $r > 0$;
- $\tilde{N} \geq 405\overline{\mathsf{b}}_2 \vee 1\ \kappa^2 \vee 67\overline{\mathsf{b}}_2 \vee 1\ \kappa^2 \log\frac{12\overline{\mathsf{b}}_2 \vee 1\ \kappa^2}{\delta}$;
- $\forall n \geq \tilde{N},\ \frac{\lambda\sqrt{2}}{1440(\mathsf{b}_2^\star \vee 1)} \leq \|\mathbf{\Sigma}\|$.

We see that such a $\tilde{N}$ can be defined explicitly.

2. Combining the assumptions on $\tilde{N}$ with the ones on M, we see that all the assumptions of proposition 4.11 are satisfied and thus that with probability at least $1-\delta$, all the hypotheses for Theorem 4.5 are satisfied except the bound on $\widehat{\nu}_{M,\lambda}(\beta)$.

3. Applying proposition 4.13, taking $\sqrt{\varepsilon} = \mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma}n^{-\gamma/2}$ and $\lambda = \left(\frac{\mathsf{Q}^2}{\mathsf{L}^2}\frac{1}{n}\right)^{\frac{\alpha}{\alpha(1+2r)+1}}$, we see that under these hypotheses,

$$\begin{aligned} N_{ns} :=& 2\left\lfloor (3+11R\|\beta_{M,\lambda}\|)\log_2\left(7R\|\nabla\widehat{L}_M(0)\|\left(\frac{n\mathsf{L}^2}{\mathsf{Q}^2}\right)^{\frac{\alpha}{\alpha(1+2r)+1}}\right)\right\rfloor \\ &+ \left\lceil \log_2\left(\frac{1}{R\mathsf{L}}\left(\frac{n\mathsf{L}^2}{\mathsf{Q}^2}\right)^{\frac{r\alpha}{\alpha(1+2r)+1}}\right)\right\rceil. \end{aligned}$$

Now we can bound this harshly:

$$N_{ns} \leq (7+22R\|\beta_{M,\lambda}\|)\left(\left\lceil\log_2\left(7R\|\nabla\widehat{L}_M(0)\|\right)\right\rceil + \left\lceil\log_2\frac{n\mathsf{L}^2}{\mathsf{Q}^2}\right\rceil + \left\lceil\log_2\frac{1}{R\mathsf{L}}\right\rceil\right).$$

Now bounding $R\|\beta_{M,\lambda}\| \leq 10 + R\|f^\star\|$, we get

$$N_{ns} \leq (227+22R\|f^\star\|)\left(\left\lceil\log_2\left(7R\|\nabla\widehat{L}_M(0)\|\right)\right\rceil + \left\lceil\log_2\frac{n\mathsf{L}^2}{\mathsf{Q}^2}\right\rceil + \left\lceil\log_2\frac{1}{R\mathsf{L}}\right\rceil\right).$$

4. Finally, we use a union bound to conclude. □

4.E Algorithm

Let $N, M \in \mathbb{N}$ with $M \leq N$. In Alg. 1, `leverage-scores-sampling`$((z_i)_{i=1}^N, M, k, \lambda)$ returns a subset of $(z_i)_{i=1}^N$ of cardinality M sampled by using (approximate) leverage scores at scale $\lambda > 0$ and computed using the kernel k. An explicit example of an algorithm computing `leverage-scores-sampling` is in Rudi, Calandriello, Carratino, and Rosasco (2018). Moreover `kernel-matrix`$((x_i)_{i=1}^N, (x'_i)_{i=1}^M, k)$ computes the kernel matrix $K \in \mathbb{R}^{N \times M}$ where $K_{ij} = k(x_i, x'_j)$, with $N, M \in \mathbb{N}$.

Algorithm 1 Algorithm efficient non-parametric learning for generalized self-concordant losses with optimal statistical guarantees discussed in Sec. 4.4 of the main paper.

Input: $(x_i, y_i)_{i=1}^n$, $n \in \mathbb{N}$, ℓ loss function, k kernel function and $\lambda > 0$.
Return: estimated function $\widehat{g} : \mathcal{X} \to \mathbb{R}$
Parameters: $Q, M, T \in \mathbb{N}$, $\mu_0 > 0$, $(q_k)_{k\in\mathbb{N}}$.
Fixed parameters: $t = 2$ from Theorem 4.1, $\tau = 3$ from proposition 4.12 in Sec. 4.D .5.
$(\bar{x}_j)_{j=1}^M \leftarrow$ `leverage-scores-sampling`$((x_i)_{i=1}^n, M, \lambda, k)$
$\mathbf{K} \leftarrow$ `kernel-matrix`$((\bar{x}_j)_{j=1}^M, (\bar{x}_j)_{j=1}^M)$
$\mathbf{T} \leftarrow$ `cholesky-upper-triangular`$(\mathbf{K})$
define the function $v(\cdot) = (k(\bar{x}_1, \cdot), \ldots, k(\bar{x}_M, \cdot)) \in \mathbb{R}^M$

define `compute-preconditioner`:
Input: $\alpha \in \mathbb{R}^M, \lambda > 0$
$c_i \leftarrow \sqrt{\ell^{(2)}(v(x_i)^\top \mathbf{T}^{-1}\alpha, y_i)}$ for all $i = 1, \ldots, n$
define the function $k'(\circ, \bullet)$ as $k'(\circ, \bullet) := c_\circ \times c_\bullet \times k(x_\circ, x_\bullet)$ for $\circ, \bullet \in \{1, \ldots, n\}$
$(h_s)_{s=1}^Q \leftarrow$ `leverage-scores-sampling`$((i)_{i=1}^n, Q, \lambda, k')$
$\mathbf{G} \leftarrow$ `kernel-matrix`$((\bar{x}_j)_{i=1}^M, (x_{h_s})_{s=1}^Q, k)$
$\mathbf{H} \leftarrow \mathbf{T}^{-\top} \times \mathbf{G} \times \mathrm{diag}((c_{l_h}^2)_{h=1}^Q) \times \mathbf{G}^\top \times \mathbf{T}^{-1}$
$\mathbf{B} \leftarrow$ `cholesky-upper-triangular`$(\frac{1}{Q}\mathbf{H} + \lambda I)$
return $\mathbf{B}$

define `preconditioned-conj-grad`:
Input: $\alpha \in \mathbb{R}^M, \mu > 0, r \in \mathbb{R}^M, \tau \in \mathbb{N}, \mathbf{B} \in \mathbb{R}^{M\times M}$
$p \leftarrow r, s_0 \leftarrow \|r\|^2, \beta \leftarrow 0$
For $i = 1, \ldots, \tau$
$z \leftarrow \mu \mathbf{B}^{-\top}\mathbf{B}^{-1}p + \frac{1}{n}\sum_{i=1}^n \ell^{(2)}(v(x_i)^\top \mathbf{T}^{-1}\alpha, y_i)\ (v(x_i)^\top \mathbf{T}^{-1}\mathbf{B}^{-1}p)\ \mathbf{B}^{-\top}\mathbf{T}^{-\top}v(x_i)$
$a \leftarrow s_0/(p^\top z)$
$\beta \leftarrow \beta + ap$
$r \leftarrow r - az,\ s_1 \leftarrow \|r\|^2$
$p \leftarrow r + (s_1/s_0)p$
$s_0 \leftarrow s_1$
return β

define `appr-linear-solver`:
Input: $\alpha \in \mathbb{R}^M, \mu > 0, g \in \mathbb{R}^M$
$\mathbf{B} \leftarrow$ `compute-preconditioner`(α, μ)
$u \leftarrow$ `preconditioned-conjugate-gradient`$(\alpha, \mu, \mathbf{B}^{-\top}g, \tau = 3, \mathbf{B})$
return $\mathbf{B}^{-1}u$

define `approximate-Newton`:
Input: $\alpha_0 \in \mathbb{R}^M, \mu > 0, t \in \mathbb{N}$
For $j = 1, \ldots, t$
$g \leftarrow \mu\alpha_{j-1} + \frac{1}{n}\sum_{i=1}^n \ell^{(1)}(v(x_i)^\top \mathbf{T}^{-1}\alpha_{j-1}, y_i)\ \mathbf{T}^{-\top}v(x_i)$
$\alpha_j \leftarrow \alpha_{j-1} -$ `appr-linear-solver`(α_{j-1}, μ, g)
return α_t

$\alpha_0 \leftarrow 0$
For $k \in \mathbb{N}$
$\alpha_{k+1} \leftarrow$ `approximate-Newton`$(\alpha_k, \mu_k, t = 2)$
$\mu_{k+1} \leftarrow q_{k+1}\mu_k$
Stop when $\mu_{k+1} < \lambda$ and set $\alpha_{last} \leftarrow \alpha_k$
$\widehat{\alpha} \leftarrow$ `approximate-Newton`$(\alpha_{last}, \lambda, T)$
return $\widehat{g}(\cdot) := v(\cdot)^\top \mathbf{T}^{-1}\widehat{\alpha}$

4.F Experiments

We present our algorithm's performance for logistic regression on two large scale data sets: Higgs and Susy. We have implemented our method using pytorch, and performed computations on one node of a Tesla P100-PCIE-16GB GPU. Recall that in the case of logistic regression, $\ell_{(x,y)}(t) = \log(1 + e^{-yt})$.

In what follows, denote with n the cardinality of the data set and d the number of features of this data set. The error is measured in terms of classification error for both data sets. In both cases, we pre-process the data by substracting the mean and dividing by the standard deviation for each feature. The data sets are the following.

Susy ($n = 5 \times 10^6$, $d = 18$, binary classification). We always use a Gaussian Kernel with $\sigma = 5$ for logistic loss (obtained through a grid search; note that in Rudi, Carratino, and Rosasco (2017), $\sigma = 4$ is used for the square loss), and will always use 10^4 Nystrom points.

Higgs ($n = 1.1 \times 10^7$, $d = 28$, binary classification). We then apply a Gaussian Kernel with $\sigma = 5$, as in Rudi, Carratino, and Rosasco (2017) (we have also performed a grid search).

For these data sets, we do not have a fixed test set, and thus set apart 20% of the data set at random to be the test set, and use the rest of the 80% to train the classifier.

In practice, we perform our globally convergent scheme with the following parameters.

- We use $Q = M$ uniform random features to compute the pre-conditioner for each approximate Newton step;
- In the first phase, we decrease μ in a very fast way to λ by starting at $\mu = 1$ and dividing μ by 1000 after performing only a single approximate Newton step (using 2 iterations of conjugate gradient descent);
- In the second phase, we perform 10 approximate Newton steps (each ANS is computed using 8 iterations of conjugate gradient descent).

Selection of λ In the introduction, we claim that in many a learning problem, the parameter λ obtained through cross validation is often much smaller than the ones obtained in statistical bounds which are usually of order $\frac{1}{\sqrt{n}}$. This leads to very ill conditioned problems.

For both data sets, we select λ (and σ, but we omit the double tables from this paper) by computing the test loss and classification errors for different values of λ, and report the evolution of these losses as a function of the parameter λ in Fig. 4.2 for the Higgs data set, and Fig. 4.3 for the Susy data set. We see that the optimal λ yield strongly ill-conditioned problems.

Comparison with accelerated methods Given the M Nystrom points, our aims to minimize $\widehat{L}_{M,\lambda}$. From an optimization point of view, i.e. from a point of view where the aim is to minimize $\widehat{L}_{M,\lambda}$, we compare our method with a large mini-batch version of Katyusha accelerated SVRG (see Allen-Zhu (2017)).
Indeed, we perform this method using batch sizes of size M; the theoretical bounds provided in Allen-Zhu (2017) show that the algorithm has linear convergence, with a time complexity of

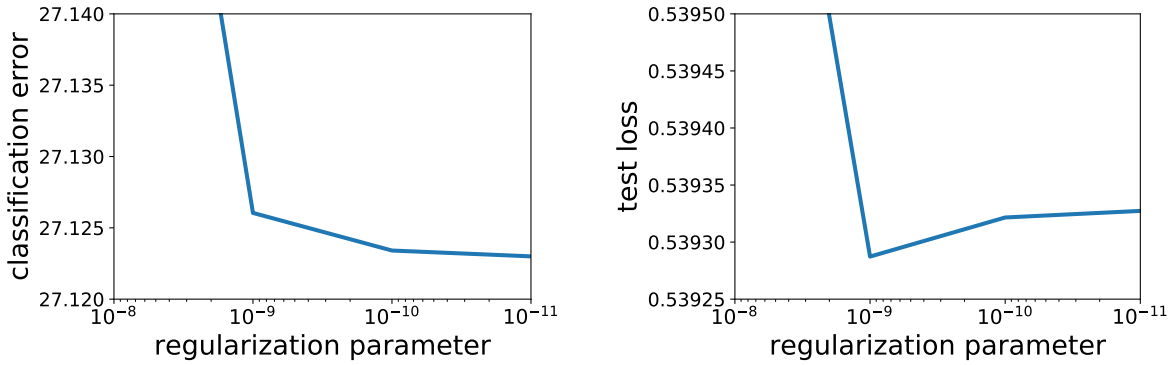

Figure 4.2: **(Left)** Classification error as a function of the regularization parameter and **(Right)** test loss as a function of the regularization parameter, when performing a logistic regression with $M = 2 \times 10^4$ Nyström features on the entire Higgs data set; we select $\lambda = 10^{-9}$.

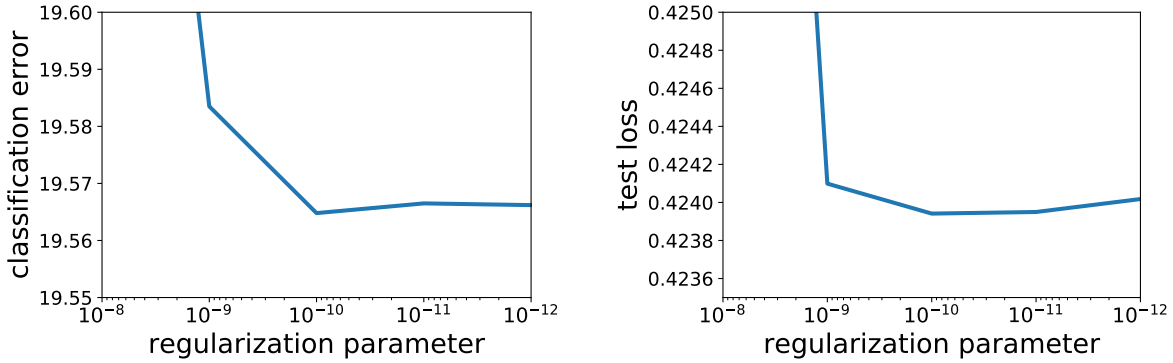

Figure 4.3: **(Left)** Classification error as a function of the regularization parameter and **(Right)** test loss as a function of the regularization parameter, when performing a logistic regression with $M = 10^4$ Nyström features on the entire Susy data set; we select $\lambda = 10^{-10}$.

order $O(nM + M^3 + M^2\sqrt{\frac{L}{\lambda}}) \log \frac{1}{\varepsilon}$ to reach precision ε. In the following plots, we compare both methods in terms of passes and time.

By pass, we mean the following.

- In the case of our second-order scheme, we define a pass on the data to be one step of the conjugate gradient descent used to compute approximate newton steps.
- In the case of Katyusha SVRG, we define a pass on the data to be either a full gradient computation or n/M computations of the type $K_{\tau M} T^{-1} \beta$ where T is an upper triangular matrix, and $K_{\tau M}$ is a $M \times M$ kernel matrix, associated to one batch gradient.

We use this notion to measure the speed of our method as they both correspond to natural $O(nM)$ operations, and incorporate the essential of the computing time. However, the second point is often much slower to compute than the first, due to the solving of the triangular system. Thus, the notion of passes is to take with precaution, as a pass for the accelerated SVRG algorithm takes much longer to run that a pass for our method. This is confirmed by the time plots (see Fig. 4.5 for in instance).

Comparison between the two methods - Due to the running time of K-SVRG, we compare both methods for $M = 10000$ Nyström points for both data sets. We compare the performance of these two algorithm with respect to the distance to the optimum in function values as well as classification error Fig. 4.4 for the Higgs data set, and in Fig. 4.5 for the Susy data set.

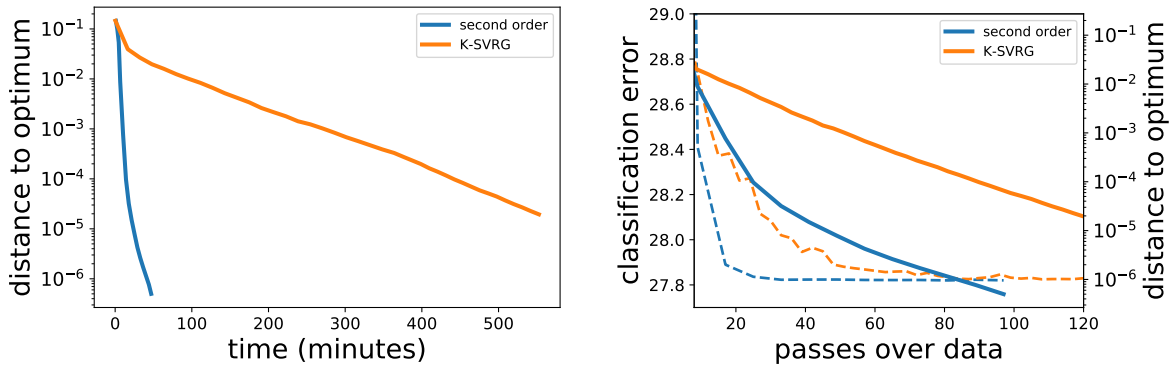

Figure 4.4: **(Left)** Distance to optimum as a function of time and **(Right)** distance to optimum and classification error as a function of the number of passes on the data when performing our second order scheme and K-SVRG to minimize the train loss on Higgs, with 1.0×10^4 Nyström points and $\lambda = 10^{-9}$.

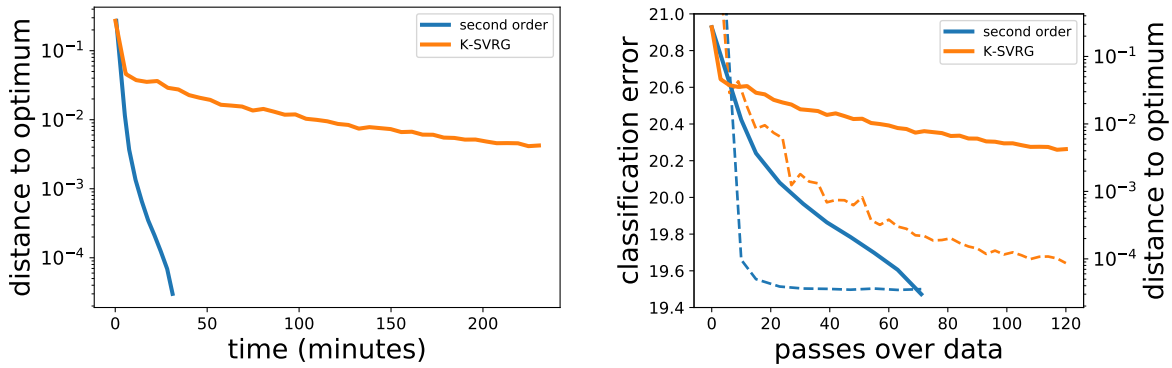

Figure 4.5: **(Left)** Distance to optimum as a function of time and **(Right)** distance to optimum and classification error as a function of the number of passes on the data when performing our second order scheme and K-SVRG to minimize the train loss on Susy, with 1.0×10^4 Nyström points and $\lambda = 10^{-10}$.

Note on the need for precise optimization - As noted in the introduction, we see in both Fig. 4.5 and Fig. 4.4 that precise optimization of the objective function is needed in order to get a good classification error. This justifies a posteriori the use of a second order method. In particular, in Fig. 4.5, one notes the difference in behavior between the two methods : the second order method converges linearly in a fast way while the first order method slows down because of the condition number.

Note on ill-conditioning - First note that in order to optimize test error, one gets very poorly conditioned problems. As predicted by the rates, we observe that K-SVRG is more sensible to ill-conditioning than our second order scheme. Indeed, in Fig. 4.6, we have plotted the results for Susy for a smaller condition number with $\lambda = 10^{-8}$, compared to $\lambda = 10^{-10}$ to get optimal

test error in Fig. 4.5. We see that the difference in number of passes needed to reach a certain precision is much lower when $\lambda = 10^{-8}$ in Fig. 4.6, confirming that K-SVRG behaves better when the condition number is smaller.

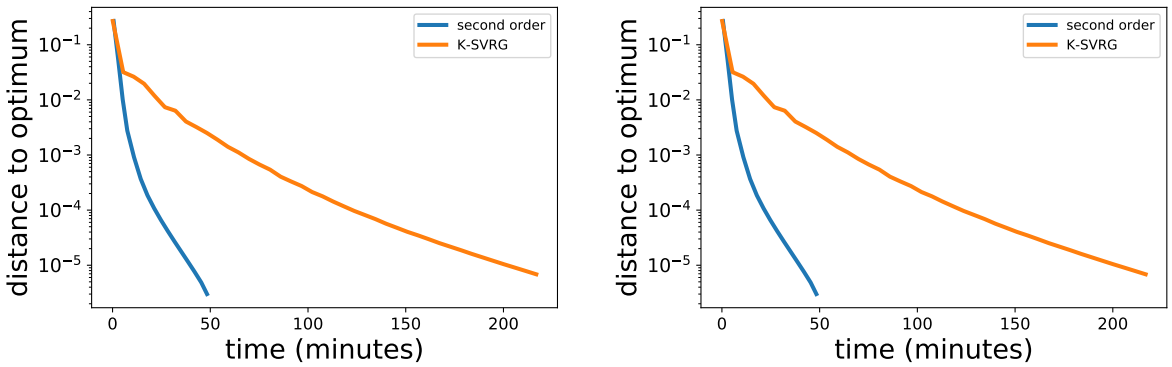

Figure 4.6: **(Left)** Distance to optimum as a function of time and **(Right)** distance to optimum and classification error as a function of the number of passes on the data when performing our second order scheme and K-SVRG to minimize the train loss on Susy, with 1.0×10^4 Nyström points and $\lambda = 10^{-8}$.

Performance of our method. In Table 4.1, we record the performance of the following methods, taking the λ values we have obtained previously for the different data sets.

For FALKON (see Rudi, Carratino, and Rosasco (2017)), we take the parameters suggested in the paper (except for the number of Nyström points needed for Higgs, as our computational capacity is limited).

Method	Susy			
	c-error	M	time(m)	c-error
Logistic regression with K-SVRG	19.64%	10^4	230	27.82 %
Logistic regression with our scheme	19.5%	10^4	15	26.9 %
Ridge Regression with FALKON (Rudi, Carratino, and Rosasco (2017))	19.7%	10^4	5	27.16 %

Table 4.1: Classification error of different methods

4.G Solving a projected problem to reduce dimension

4.G .1 Introduction and notations

In this section, we give ourselves a generalized self-concordant function f whose associated subset we denote with $\mathcal{G}$. Once again, we will always omit the subscript f in the notations associated to f.

The aim of this section is the following. Given f and $\lambda > 0$, computing an approximate solution to

$$x^\star_\lambda = \arg\min_{x\in\mathcal{H}} f_\lambda(x),$$

is often too costly. Instead, we look for a solution in a small subset of $\mathcal{H}$ which we see as the image of a certain orthogonal projector $\mathbf{P}$ and which we denote $\mathcal{H}_\mathbf{P}$. Usually, this subset will be finite dimensional and admit an easy parametrization. Thus we will compare an approximation of $x^\star_\lambda$ to an approximation of

$$x^*_{\mathbf{P},\lambda} = \arg\min_{x\in\mathcal{H}_\mathbf{P}} f_\lambda(x) = \arg\min_{x\in\mathcal{H}} f(\mathbf{P}x) + \frac{\lambda}{2}\|x\|^2.$$

Denote with $f_\mathbf{P}$ the mapping $x \in \mathcal{H} \mapsto f(\mathbf{P}x)$. It is easy to see that, as f is a generalized self-concordant function with $\mathcal{G}$, $f_\mathbf{P}$ is naturally a generalized self-concordant with $\mathcal{G}_\mathbf{P} := \mathbf{P}\mathcal{G} = \{\mathbf{P}g \ : \ g \in \mathcal{G}\}$. Moreover, $x^*_{\mathbf{P},\lambda} = x^\star_{f_\mathbf{P},\lambda}$.

We will adopt the following notations for the quantities related to the generalized self-concordant function $f_\mathbf{P}$. Essentially, we always replace $f_\mathbf{P}$ simply by $\mathbf{P}$ from our definitions in appendix.

- For the regularized function :

$$\forall x \in \mathcal{H},\ \forall\lambda > 0,\ f_{\mathbf{P},\lambda}(x) = f_\mathbf{P}(x) + \frac{\lambda}{2}\|x\|^2.$$

- For the Hessians

$$\forall x \in \mathcal{H},\ \lambda > 0,\ \mathbf{H}_{\mathbf{P},\lambda}(x) = \mathbf{H}_{f_\mathbf{P},\lambda}(x) = \mathbf{P}\mathbf{H}(\mathbf{P}x)\mathbf{P} + \lambda\mathbf{I}.$$

- $\forall h \in \mathcal{H},\ \mathsf{t}_\mathbf{P}(h) := \mathsf{t}_{f_\mathbf{P}}(h) = \mathsf{t}(\mathbf{P}h)$.
- For the Newton decrement:

$$\forall x \in \mathcal{H},\ \lambda > 0,\ \nu_{\mathbf{P},\lambda}(x) = \nu_{f_\mathbf{P},\lambda}(x) = \|\nabla f_{\mathbf{P},\lambda}\|_{\mathbf{H}^{-1}_{\mathbf{P},\lambda}(x)} = \|\mathbf{P}\nabla f(\mathbf{P}x) + \lambda x\|_{\mathbf{H}^{-1}_{\mathbf{P},\lambda}(x)}.$$

- For the Dikin ellipsoid radius:

$$\forall\lambda > 0,\ \forall x \in \mathcal{H},\ \mathsf{r}_{\mathbf{P},\lambda}(x) := \mathsf{r}_{f_\mathbf{P},\lambda}(x) = \frac{1}{\sup_{g\in\mathcal{G}} \|\mathbf{P}g\|_{\mathbf{H}^{-1}_{\lambda,\mathbf{P}}(x)}};$$

- For the Dikin ellipsoid:

$$\forall\lambda > 0,\ \forall\mathsf{c} \geq 0,\ \mathsf{D}_{\mathbf{P},\lambda}(\mathsf{c}) := \mathsf{D}_{f_\mathbf{P},\lambda}(\mathsf{c}).$$

Note that for any $x \in \mathcal{H}_{\mathbf{P}}$, $\mathsf{r}_{\mathbf{P},\lambda}(x) \geq \mathsf{r}_\lambda(x)$.

We will now introduce the key quantities in order to compare an approximation of $x^*_{\mathbf{P},\lambda}$ to an approximation of $x^\star_\lambda$.

Definition 4.8 (key quantities)**.** *Define the following quantities*

- *For any $\lambda > 0$, the source term $\mathsf{s}_\lambda := \lambda \|x^\star_\lambda\|_{\mathbf{H}^{-1}_\lambda(x^\star_\lambda)} = \|\nabla f(x^\star_\lambda)\|_{\mathbf{H}^{-1}_\lambda(x^\star_\lambda)}$;*
- *Given an orthogonal projector $\mathbf{P}$, $\lambda > 0$, and $x \in \mathcal{H}$, the capacity of the projector $\mathsf{C}_{\mathbf{P}}(x,\lambda) := \frac{\|\mathbf{H}(x)^{1/2}(\mathbf{I}-\mathbf{P})\|^2}{\lambda}$.*

4.G .2 Relating the projected to the original problem

Given $x \in \mathcal{H}_{\mathbf{P}}$, our aim is to bound $\nu_\lambda(x)$ given $\nu_{\lambda,\mathbf{P}}(x)$ and s_λ.

Proposition 4.15. *Let $x \in \mathcal{H}_{\mathbf{P}}$. If*

$$\frac{\mathsf{s}_\lambda}{\mathsf{r}_\lambda(x^\star_\lambda)} \leq \frac{1}{4},\ \mathsf{C}_{\mathbf{P}}(x^\star_\lambda,\lambda) \leq \frac{1}{120},\ \nu_{\mathbf{P},\lambda}(x) \leq \frac{\mathsf{r}_{\mathbf{P},\lambda}(x)}{2},$$

Then it holds:

$$\nu_\lambda(x) \leq 3(\nu_{\mathbf{P},\lambda}(x) + \mathsf{s}_\lambda).$$

Moreover, under these conditions,

- $\|x - x^\star_\lambda\| \leq 7\lambda^{-1/2}(\nu_{\mathbf{P},\lambda}(x) + \mathsf{s}_\lambda)$;
- $\lambda\|x\|_{\mathbf{H}^{-1}_{\mathbf{P},\lambda}(x)} \leq 7\nu_{\mathbf{P},\lambda}(x) + 9\mathsf{s}_\lambda$.

Proof. In this proof, introduce the following auxiliary quantity:

$$\gamma_\lambda := \frac{\mathsf{s}_\lambda}{\mathsf{r}_\lambda(x^\star_\lambda)}.$$

1) Start by bounding $\mathsf{t}(\mathbf{P}x^\star_\lambda - x^\star_\lambda)$. It holds:

$$\begin{aligned}
\mathsf{t}(\mathbf{P}x - x^\star_\lambda) &= \sup_{g\in\mathcal{G}} |g\cdot(\mathbf{I}-\mathbf{P})x^\star_\lambda| \\
&\leq \frac{1}{\mathsf{r}_\lambda(x^\star_\lambda)}\ \|(\mathbf{I}-\mathbf{P})x^\star_\lambda\|_{\mathbf{H}_\lambda(x^\star_\lambda)} \\
&\leq \frac{1}{\mathsf{r}_\lambda(x^\star_\lambda)}\ \|\mathbf{H}_\lambda(x^\star_\lambda)^{1/2}(\mathbf{I}-\mathbf{P})\mathbf{H}_\lambda(x^\star_\lambda)^{1/2}\|\ \|\mathbf{H}^{-1/2}_\lambda(x^\star_\lambda)x^\star_\lambda\| \\
&= (1+\mathsf{C}_{\mathbf{P}}(x^\star_\lambda,\lambda))\ \frac{\lambda\|\mathbf{H}^{-1/2}_\lambda(x^\star_\lambda)x^\star_\lambda\|}{\mathsf{r}_\lambda(x^\star_\lambda)} \\
&= (1+\mathsf{C}_{\mathbf{P}}(x^\star_\lambda,\lambda))\ \gamma_\lambda.
\end{aligned}$$

2) Then bound $\mathsf{t}(x^*_{\mathbf{P},\lambda} - \mathbf{P}x^\star_\lambda)$ First, bound $\nu_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda)$:

$$\begin{aligned}\nu_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda) &= \|\mathbf{P}\nabla f_\lambda(\mathbf{P}x^\star_\lambda)\|_{\mathbf{H}_{\lambda,\mathbf{P}}(\mathbf{P}x^\star_\lambda)^{-1}} \\ &\leq \|\nabla f_\lambda(\mathbf{P}x^\star_\lambda)\|_{\mathbf{H}_\lambda(\mathbf{P}x^\star_\lambda)^{-1}}.\end{aligned}$$

Using Eq. (4.17), we get $\|\nabla f_\lambda(\mathbf{P}x^\star_\lambda)\|_{\mathbf{H}_\lambda(\mathbf{P}x^\star_\lambda)^{-1}} \leq e^{\mathsf{t}((\mathbf{I}-\mathbf{P})x^\star_\lambda)/2}\nu_\lambda(\mathbf{P}x^\star_\lambda)$. Using Eq. (4.20), we can bound

$$\nu_\lambda(\mathbf{P}x^\star_\lambda) \leq \overline{\phi}(\mathsf{t}((\mathbf{I}-\mathbf{P})x^\star_\lambda))\ \|(\mathbf{I}-\mathbf{P})x^\star_\lambda\|_{\mathbf{H}_\lambda(x^\star_\lambda)} \leq \overline{\phi}(\mathsf{t}((\mathbf{I}-\mathbf{P})x^\star_\lambda))\ (1+\mathsf{C}_\mathbf{P}(x^\star_\lambda,\lambda))\mathsf{s}_\lambda.$$

Putting things together,

$$\nu_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda) \leq e^{\mathsf{t}((\mathbf{I}-\mathbf{P})x^\star_\lambda)/2}\overline{\phi}(\mathsf{t}((\mathbf{I}-\mathbf{P})x^\star_\lambda))\ (1+\mathsf{C}_\mathbf{P}(x^\star_\lambda,\lambda))\mathsf{s}_\lambda.$$

Now

$$\frac{1}{\mathsf{r}_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda)} \leq \frac{1}{\mathsf{r}_\lambda(\mathbf{P}x^\star_\lambda)} \leq e^{\mathsf{t}((\mathbf{I}-\mathbf{P})x^\star_\lambda)/2}\frac{1}{\mathsf{r}_\lambda(x^\star_\lambda)}.$$

Hence,

$$\frac{\nu_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda)}{\mathsf{r}_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda)} \leq e^{\tilde{t}_\lambda}\overline{\phi}(\tilde{t}_\lambda)\ \tilde{t}_\lambda, \qquad \tilde{t}_\lambda = (1+\mathsf{C}_\mathbf{P}(x^\star_\lambda,\lambda))\gamma_\lambda.$$

Since $t \mapsto e^t\overline{\phi}(t)\ t$ is an increasing function whose value in 0 is 0, we find numerically that for $t = \frac{3}{10}$, $e^t\overline{\phi}(t)\ t \leq \frac{1}{2}$. Hence, if $(1+\mathsf{C}_\mathbf{P}(x^\star_\lambda,\lambda))\gamma_\lambda \leq \frac{3}{10}$, then $\frac{\nu_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda)}{\mathsf{r}_{\mathbf{P},\lambda}(\mathbf{P}x^\star_\lambda)} \leq \frac{1}{2}$. Using Lemma 4.5, this shows that

$$\mathsf{t}_\mathbf{P}(\mathbf{P}x^\star_\lambda - x^*_{\mathbf{P},\lambda}) = \mathsf{t}(\mathbf{P}x^\star_\lambda - x^*_{\mathbf{P},\lambda}) \leq \log 2.$$

3) Getting a bound for $\mathsf{t}(x - x^\star_\lambda)$. To do so, combine the two previous bounds with the fact that if $\nu_{\mathbf{P},\lambda}(x) \leq \frac{\mathsf{r}_{\mathbf{P},\lambda}(x)}{2}$, then using Lemma 4.5 with $f_\mathbf{P}$, $\mathsf{t}_\mathbf{P}(x - x^*_{\mathbf{P},\lambda}) = \mathsf{t}(x - x^*_{\mathbf{P},\lambda}) \leq \log 2$. Thus, if

$$(1+\mathsf{C}_\mathbf{P}(x^\star_\lambda,\lambda))\gamma_\lambda \leq \frac{3}{10},\ \nu_{\mathbf{P},\lambda}(x) \leq \frac{\mathsf{r}_{\mathbf{P},\lambda}(x)}{2},$$

then it holds

$$\mathsf{t}(x - x^\star_\lambda) \leq \frac{3}{10} + 2\log 2.$$

4) A technical result to bound $\|\mathbf{H}_\lambda(x)^{-1/2}\mathbf{H}_{\mathbf{P},\lambda}(x)^{1/2}\|$. Using the fact that $\mathbf{P}x = x$, and Lemma 4.23, applied to $\mathbf{A} = \mathbf{H}(x)$, we get

$$\|\mathbf{H}_\lambda(x)^{-1/2}\mathbf{H}_{\mathbf{P},\lambda}(x)^{1/2}\| \leq 1 + \sqrt{\mathsf{C}_\mathbf{P}(x,\lambda)}.$$

Then, one can easily bound $\mathsf{C}_\mathbf{P}(x,\lambda) \leq e^{\mathsf{t}(x-x^\star_\lambda)}\mathsf{C}_\mathbf{P}(x^\star_\lambda,\lambda)$.

5) Let us now bound $\nu_\lambda(x)$. First, decompose the term

$$\nu_\lambda(x) = \|\nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \|\mathbf{P}\nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)} + \|(\mathbf{I}-\mathbf{P})\nabla f(x)\|_{\mathbf{H}_\lambda^{-1}(x)}.$$

Since $x \in \mathcal{H}_\mathbf{P}$, $\|\mathbf{P}\nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)} = \|\nabla f_{\mathbf{P},\lambda}(x)\|_{\mathbf{H}_\lambda^{-1}(x)}$, and using the previous point, we get

$$\|\mathbf{P}\nabla f_\lambda(x)\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \left(1 + e^{\mathsf{t}(x-x_\lambda^\star)/2}\sqrt{\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\right)\nu_{\mathbf{P},\lambda}(x).$$

Let us now bound the second term. We divide it into two terms:

$$\|(\mathbf{I}-\mathbf{P})\nabla f(x)\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \|(\mathbf{I}-\mathbf{P})\left(\nabla f(x) - \nabla f(x_\lambda^\star)\right)\|_{\mathbf{H}_\lambda^{-1}(x)} + \|(\mathbf{I}-\mathbf{P})\nabla f(x_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(x)}.$$

The second term can be bounded in the following way:

$$\|(\mathbf{I}-\mathbf{P})\nabla f(x_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \frac{1}{\sqrt{\lambda}}\|(\mathbf{I}-\mathbf{P})\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\|\ \|\nabla f(x_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(x_\lambda^\star)} \leq \sqrt{1+\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\ \mathsf{s}_\lambda.$$

For the first term, we proceed in the following way.

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{P})\left(\nabla f(x) - \nabla f(x_\lambda^\star)\right)\|_{\mathbf{H}_\lambda^{-1}(x)} &= \|\int_0^1 \mathbf{H}_\lambda^{-1/2}(x)(\mathbf{I}-\mathbf{P})\mathbf{H}(x_t)(x-x_\lambda^\star)\ dt\| \\
&\leq \frac{1}{\sqrt{\lambda}}\int_0^1 \|(\mathbf{I}-\mathbf{P})\mathbf{H}^{1/2}(x_t)\|\ \|\mathbf{H}^{1/2}(x_t)(x-x_\lambda^\star)\|\ dt \\
&\leq \sqrt{\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\ \overline{\phi}(\mathsf{t}(x-x_\lambda^\star))\ \|x-x_\lambda^\star\|_{\mathbf{H}(x_\lambda^\star)} \\
&\leq \sqrt{\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\ e^{\mathsf{t}(x-x_\lambda^\star)}\nu_\lambda(x).
\end{aligned}$$

Hence the final bound:

$$\left(1 - \sqrt{\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\ e^{\mathsf{t}(x-x_\lambda^\star)}\right)\nu_\lambda(x) \leq \left(1 + e^{\mathsf{t}(x-x_\lambda^\star)/2}\sqrt{\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\right)\nu_{\mathbf{P},\lambda}(x) + \sqrt{1+\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\ \mathsf{s}_\lambda.$$

Now if $\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda) \leq \frac{1}{120}$, we see that $\sqrt{\mathsf{C}_\mathbf{P}(x_\lambda^\star,\lambda)}\ e^{\mathsf{t}(x-x_\lambda^\star)} \leq \frac{1}{2}$, and hence, using the bound on $\mathsf{t}(x-x_\lambda^\star)$,

$$\nu_\lambda(x) \leq 3(\nu_{\mathbf{P},\lambda}(x) + \mathsf{s}_\lambda).$$

6) Showing the last two points . We leverage the fact that $\nu_\lambda(x) \leq 3(\nu_{\mathbf{P},\lambda}(x) + \mathsf{s}_\lambda)$ and $\mathsf{t}(x-x_\lambda^\star) \leq \frac{3}{10} + 2\log 2$.
To show the first bound, we plug in the previous results in the following equation:

$$\|x-x_\lambda^\star\| \leq \lambda^{-1/2}\|x-x_\lambda^\star\|_{\mathbf{H}_\lambda(x)} \leq \frac{1}{\underline{\phi}(\mathsf{t}(x-x_\lambda^\star))}\ \lambda^{-1/2}\nu_\lambda(x).$$

The last inequality is obtained using Eq. (4.18).

To show the second point, we use the fact that $x \in \mathcal{H}_\mathbf{P}$ to show that

$$\lambda\|x\|_{\mathbf{H}_{\mathbf{P},\lambda}^{-1}(x)} \leq \lambda\|x\|_{\mathbf{H}_\lambda^{-1}(x)} \leq \lambda\|x-x_\lambda^\star\|_{\mathbf{H}_\lambda(x)} + \lambda\|x_\lambda^\star\|_{\mathbf{H}_\lambda^{-1}(x)}.$$

Then applying Eq. (4.17) and Eq. (4.18):

$$\lambda\|x\|_{\mathbf{H}^{-1}_{\mathbf{P},\lambda}(x)} \leq \frac{1}{\underline{\phi}(\mathsf{t}(x-x^\star_\lambda))}\ \nu_\lambda(x) + e^{\mathsf{t}(x-x^\star_\lambda)/2}\mathsf{s}_\lambda.$$

We then use the previous results to conclude. □

4.G .3 Finding a good projector

Lemma 4.16. *If for a certain $\eta \leq \lambda$ and for a certain constant C, $\|\mathbf{H}^{1/2}_\eta(x)(\mathbf{I}-\mathbf{P})\|^2 \leq C\eta$, then*

$$\mathsf{C}_{\mathbf{P}}(x,\lambda) \leq \frac{C\eta}{\lambda}.$$

Proof. This is completely direct, using the fact that $\mathbf{H}^{1/2}(x) \preceq \mathbf{H}^{1/2}_\eta(x)$. □

4.H Relations between statistical problems and empirical problem.

In this section, we recall and reformulate the framework from Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019).

4.H .1 Statistical problem and ERM estimator

Let $\mathcal{Z}$ be a Polish space and Z be a random variable on $\mathcal{Z}$ with distribution ρ. Let $\mathcal{H}$ be a separable Hilbert space, with norm $\|\cdot\|$, and let $(f_z)_{z\in\mathcal{Z}}$ be a family of functions on $\mathcal{H}$. Our goal is to minimize the *expected risk* with respect to $x \in \mathcal{H}$:

$$\inf_{x\in\mathcal{H}} \; f(x) := \mathbb{E}\left[f_Z(x)\right].$$

Given $(z_i)_{i=1}^n \in \mathcal{Z}^n$, we define the *empirical risk*:

$$\widehat{f}(x) := \frac{1}{n}\sum_{i=1}^n f_{z_i}(x),$$

and consider the following estimator based on regularized empirical risk minimization given $\lambda > 0$ (note that the minimizer is unique in this case):

$$\widehat{x}_\lambda^\star = \operatorname*{arg\,min}_{x\in\mathcal{H}} \widehat{f}_\lambda(x) := \widehat{f}(x) + \frac{\lambda}{2}\|x\|^2,$$

where we assume the following.

Assumption 4.6 (i.i.d. data)**.** *The samples $(z_i)_{1\le i\le n}$ are independently and identically distributed according to ρ.*

We make the following assumption on the family (f_z) (this is a reformulation of Assumption 8 in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019))

Assumption 4.7 (Generalized self-concordance)**.** *For any $z \in \mathcal{Z}$, there exists an associated subset $\mathcal{G}_z \subset \mathcal{H}$ such that $(f_z, \mathcal{G}_z)$ is generalized self-concordant in the sense of definition 4.3.*

Moreover we require the following technical assumption to guarantee that f and and its derivatives are well defined for any $x \in \mathcal{H}$ (this is a reformulation of Assumptions 3 and 4 in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), and the necessary conditions to obtain proposition 4.3).

Assumption 4.8 (Technical assumptions)**.** *The mapping $(z, x) \in \mathcal{Z}\times\mathcal{H} \mapsto f_z(x)$ is measurable. Moreover,*

- *the random variables $\|f_Z(0)\|, \|\nabla f_Z(0)\|, \mathrm{Tr}(\nabla^2 f_Z(0))$ are are bounded;*
- *$\mathcal{G} := \bigcup_{z\in\mathrm{supp}(Z)} \mathcal{G}_z$ is a bounded subset of $\mathcal{H}$.*

The assumptions above are usually easy to check in practice. In particular, if the support of ρ is bounded, the mappings $z \mapsto \ell_z(0), \nabla\ell_z(0), \mathrm{Tr}(\nabla^2\ell_z(0))$ are continuous, and $z \mapsto \mathcal{G}_z$ is uniformly bounded on bounded sets, then they hold.

Proposition 4.16. *Under Assumptions 4.7 and 4.8, the function $(f, \mathcal{G})$ (or simply f) is generalized self-concordant.*

Moreover, under Assumption 4.6, define

$$\widehat{\mathcal{G}} := \bigcup_{i=1}^{n} \mathcal{G}_{z_i}.$$

Then $(\widehat{f}, \widehat{\mathcal{G}})$ (or simply $\widehat{f}$) is generalized self-concordant. Moreover, note that $\widehat{\mathcal{G}} \subset \mathcal{G}$.

The main regularity assumption we make on our statistical problems follows (see Assumption 5 in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019)).

Assumption 4.9 (Existence of a minimizer)**.** *There exists $x^\star \in \mathcal{H}$ such that $f(x^\star) = \inf_{x \in \mathcal{H}} f(x)$.*

Notations We adopt all the notations from Sec. 4.A for f and $\widehat{f}$, which are generalized self-concordant functions with associated subsets given in proposition 4.16 with the following conventions:

- For all quantities relating to f, we omit the subscript f as usual;
- For all quantities relating to $\widehat{f}$, we omit the subscript $\widehat{f}$ and instead put a hat over all these quantities. For instance:

$$\widehat{\mathbf{H}}(x) := \mathbf{H}_{\widehat{f}}(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_{z_i}(x),\ \widehat{\mathsf{r}}_\lambda(x) := \mathsf{r}_{\widehat{f},\lambda}(x) = \frac{1}{\sup_{g \in \widehat{\mathcal{G}}} \|g\|_{\widehat{\mathbf{H}}_\lambda^{-1}(x)}},\ \text{etc...}$$

Recall the two main quantities introduced in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) to establish the quality of our estimator $\widehat{x}_\lambda^\star$ (in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), this is a mix between Proposition 2 and Definition 3).

Proposition 4.17 (Bias, degrees of freedom)**.** *Suppose Assumptions 4.7 to 4.9 are satisfied. The following key quantities are well defined:*

- *the* bias $\mathsf{b}_\lambda = \|\mathbf{H}_\lambda(x^\star)^{-1/2}\nabla f_\lambda(x^\star)\|$*;*
- *the* effective dimension $\mathsf{df}_\lambda = \mathbb{E}\left[\|\mathbf{H}_\lambda(x^\star)^{-1/2}\nabla f_Z(x^\star)\|^2\right]$.

Moreover, we also introduce the following quantities:

$$\mathsf{B}_1^\star := \sup_{z \in \mathrm{supp}(Z)} \|\nabla f_z(x^\star)\|, \qquad \mathsf{B}_2^\star := \sup_{z \in \mathrm{supp}(Z)} \mathrm{Tr}(\nabla^2 f_z(x^\star)), \qquad \mathsf{Q}^\star = \frac{\mathsf{B}_1^\star}{\sqrt{\mathsf{B}_2^\star}}.$$

We can now recall the main theorem of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) (Theorem 4), which quantifies the behavior of the ERM estimator:

Theorem 4.8 (Bound for the ERM estimator)**.** *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Whenever*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \qquad \sqrt{\triangle_2 \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n} \log\frac{2}{\delta}} \leq \mathsf{r}_\lambda(x^\star), \qquad 2\mathsf{b}_\lambda \leq \mathsf{r}_\lambda(x^\star),$$

then with probability at least $1 - 2\delta$, it holds

$$f(\widehat{x}_\lambda^\star) - f(x^\star) \leq \mathsf{C}_{\text{bias}}\ \mathsf{b}_\lambda^2 + \mathsf{C}_{\text{var}}\ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n}\ \log\frac{2}{\delta}, \tag{4.37}$$

where $\mathsf{C}_{\text{bias}}, \mathsf{C}_{\text{var}}, \square_1 \leq 414,\ \triangle_1,\ \triangle_2 \leq 5184$.

4.H .2 Link between a good approximation of $\widehat{x}_\lambda^\star$ and $x^\star$

In this paper, we provide an algorithm which can effectively compute a good approximation of $\widehat{x}_\lambda^\star$ (as it is a finite sum problem which can be solved). This algorithm will return a certain $x \in \mathcal{H}$, whose precision with respect to the empirical problem will be characterized by $\widehat{\nu}_\lambda(x)$. The aim of the following lemma is to see how this approximation x behaves with respect to the statistical problem.

Lemma 4.17. *Suppose the conditions for Theorem 4.8 are satisfied, i.e. let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$ and suppose*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \qquad \sqrt{\triangle_2 \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n} \log \frac{2}{\delta}} \leq \mathsf{r}_\lambda(x^\star), \qquad 2\mathsf{b}_\lambda \leq \mathsf{r}_\lambda(x^\star).$$

Let x be an approximation of $\widehat{x}_\lambda^\star$ characterized by its Newton decrement $\widehat{\nu}_\lambda(x)$. If

$$\widehat{\nu}_\lambda(x) \leq \frac{\widehat{\mathsf{r}}_\lambda(x)}{2}, \ \widehat{\nu}_\lambda(x) \leq \frac{\mathsf{r}_\lambda(x^\star)}{2},$$

then with probability at least $1 - 2\delta$, it holds

$$f(x) - f(x^\star) \leq 14(f(\widehat{x}_\lambda^\star) - f(x^\star)) + 30\widehat{\nu}_\lambda(x)^2.$$

Proof. Using Eq. (4.16),

$$\begin{aligned} f(x) - f(\widehat{x}_\lambda^\star) &\leq \langle \nabla f(\widehat{x}_\lambda^\star), x - \widehat{x}_\lambda^\star \rangle_{\mathcal{H}} + \psi(\mathsf{t}(x - \widehat{x}_\lambda^\star)) \|x - \widehat{x}_\lambda^\star\|^2_{\mathbf{H}_\lambda(\widehat{x}_\lambda^\star)} \\ &\leq \frac{1}{2} \|\nabla f(\widehat{x}_\lambda^\star)\|^2_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} + \left(\psi(\mathsf{t}(x - \widehat{x}_\lambda^\star)) + \frac{1}{2}\right) \|x - \widehat{x}_\lambda^\star\|^2_{\mathbf{H}_\lambda(\widehat{x}_\lambda^\star)}. \end{aligned}$$

1. Let us bound $\|\nabla f(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)}$

$$\begin{aligned} \|\nabla f(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} &\leq \int_0^1 \|\mathbf{H}_\lambda^{-1/2}(\widehat{x}_\lambda^\star) \mathbf{H}(x_t)(\widehat{x}_\lambda^\star - x^\star)\| \, dt, \qquad x_t = (1-t)\widehat{x}_\lambda^\star + tx^\star \\ &\leq \int_0^1 \|\mathbf{H}_\lambda^{-1/2}(\widehat{x}_\lambda^\star) \mathbf{H}^{1/2}(x_t)\| \, \|\mathbf{H}^{1/2}(x_t)(\widehat{x}_\lambda^\star - x^\star)\| \, dt. \end{aligned}$$

Now using equation Eq. (4.14)

$$\mathbf{H}(x_t) \preceq e^{t\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)} \mathbf{H}(\widehat{x}_\lambda^\star), \qquad \mathbf{H}(x_t) \preceq e^{(1-t)\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)}.$$

Thus:

$$\|\nabla f(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} \leq e^{\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)/2} \, \|\widehat{x}_\lambda^\star - x^\star\|_{\mathbf{H}(x^\star)}.$$

Finally, using equation Eq. (4.16)

$$\|\nabla f(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} \leq \frac{e^{\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)/2}}{\psi(-\mathsf{t}(\widehat{x}_\lambda^\star - x^\star))^{1/2}} \left(f(\widehat{x}_\lambda^\star) - f(x^\star)\right)^{1/2}.$$

2. Let us bound the terms involving $\|x - \widehat{x}_\lambda^\star\|_{\mathbf{H}_\lambda(\widehat{x}_\lambda^\star)}$ Note that using Eq. (4.18) and Eq. (4.17) applied to $\widehat{f}$,

$$\|x - \widehat{x}_\lambda^\star\|_{\mathbf{H}_\lambda(\widehat{x}_\lambda^\star)} \leq \|\mathbf{H}_\lambda^{1/2}(\widehat{x}_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\widehat{x}_\lambda^\star)\| \; \frac{e^{\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star)/2}}{\underline{\phi}(\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star))}\widehat{\nu}_\lambda(x).$$

This also leads to:

$$\begin{aligned}\mathsf{t}(x - \widehat{x}_\lambda^\star) &\leq \frac{1}{\mathsf{r}_\lambda(\widehat{x}_\lambda^\star)} \; \|\mathbf{H}_\lambda^{1/2}(\widehat{x}_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\widehat{x}_\lambda^\star)\| \; \|x - \widehat{x}_\lambda^\star\|_{\widehat{\mathbf{H}}_\lambda(\widehat{x}_\lambda^\star)} \\ &\leq \frac{1}{\mathsf{r}_\lambda(\widehat{x}_\lambda^\star)} \; \|\mathbf{H}_\lambda^{1/2}(\widehat{x}_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(\widehat{x}_\lambda^\star)\| \; \frac{e^{\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star)/2}}{\underline{\phi}(\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star))}\widehat{\nu}_\lambda(x).\end{aligned}$$

3. Putting things together In the end, we get

$$\begin{aligned}f(x) - f(x^\star) \leq &\left(1 + \frac{e^{\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)}}{\psi(-\mathsf{t}(\widehat{x}_\lambda^\star - x^\star))}\right)(f(\widehat{x}_\lambda^\star) - f(x^\star)) \\ &+ \left(\psi(\mathsf{t}(x - \widehat{x}_\lambda^\star)) + \frac{1}{2}\right)\left(e^{\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|\frac{e^{\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star)/2}}{\underline{\phi}(\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star))}\right) \; \widehat{\nu}_\lambda(x)^2.\end{aligned}$$

Moreover, we bound

$$\mathsf{t}(x - \widehat{x}_\lambda^\star) \leq e^{(\mathsf{t}(x^\star - \widehat{x}_\lambda^\star) + \mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star))/2} \; \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\| \; \frac{e^{\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star)/2}}{\underline{\phi}(\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star))}\frac{\widehat{\nu}_\lambda(x)}{\mathsf{r}_\lambda(x^\star)}.$$

4. Plugging in previous results Under the assumptions of this lemma, which include the assumptions of Theorem 4. in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), we get the following bounds.

- In Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019),the assumptions of Theorem 4 imply that we can use Lemma 9, which uses Lemma 8 in which we show that with probability at least $1 - \delta$,
$$\|\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\mathbf{H}_\lambda(x_\lambda^\star)^{1/2}\|^2 \leq 2.$$
- Still using the assumptions of Theorem 4, we see in the proof of this theorem that the assumptions of Theorem 7 of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) are satisfied in the case where $\mathsf{b}_\lambda \leq \frac{\mathsf{r}_\lambda(x^\star)}{2}$, and thus that
$$\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star) \leq \log 2, \;\; \mathsf{t}(x_\lambda^\star - x^\star) \leq \log 2.$$

Plugging in all these bounds, we get

$$\left(1 + \frac{e^{\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)}}{\psi(-\mathsf{t}(\widehat{x}_\lambda^\star - x^\star))}\right) \leq 14, \;\; \mathsf{t}(x - \widehat{x}_\lambda^\star) \leq 6,$$

$$\left(\psi(\mathsf{t}(x-\widehat{x}_\lambda^\star))+\frac{1}{2}\right)\left(e^{\mathsf{t}(\widehat{x}_\lambda^\star-x_\lambda^\star)/2}\|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|\frac{e^{\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star)/2}}{\underline{\phi}(\widehat{\mathsf{t}}(x-\widehat{x}_\lambda^\star))}\right)\leq 30.$$

□

4.H .3 Bounds when we solve a projected empirical problem

In this section, we place ourselves in the setting of Sec. 4.G . In this section, we had argued that for computational purposes, it was less costly to compute an approximate solution to a projected problem.

In this section, we assume that we are going to project the regularized empirical problem, that is solve approximately

$$x\approx\operatorname*{arg\,min}_{x\in\mathcal{H}}\widehat{f}_{\mathbf{P},\lambda}(x)=\widehat{f}(\mathbf{P}x)+\frac{\lambda}{2}\|x\|^2.$$

for a given orthogonal projection $\mathbf{P}$. Recall from Sec. 4.G that there is a natural way of seeing $\widehat{f}_{\mathbf{P}}$ as a generalized self-concordant function. We import all the notations from this section, keeping a $\widehat{\cdot}$ over all notations to mark the fact that we are projecting $\widehat{f}$ and not f.

To quantify the quality of the approximation x, we will use the Newton decrement for the empirical projected problem $\widehat{\nu}_{\mathbf{P},\lambda}(x):=\nu_{\widehat{f}_{\mathbf{P}},\lambda}(x)$.

As we see in proposition 4.15, under certain conditions, bounding $\widehat{\nu}_\lambda(x)$ amounts to bounding two terms:

- The empirical source $\widehat{s}_\lambda:=\lambda\|\widehat{x}_\lambda^\star\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\widehat{x}_\lambda^\star)}$,
- The projected empirical Newton decrement $\widehat{\nu}_{\mathbf{P},\lambda}(x)$.

1. Bounding the empirical source term $\widehat{s}_\lambda$ Start by bounding the source empirical source term using quantities we know.

Lemma 4.18 (Empirical source)**.** *Let $n\in\mathbb{N}$, $\delta\in(0,1/2]$, $0<\lambda\leq\mathsf{B}_2^\star$. Whenever*

$$n\geq\triangle_1\frac{\mathsf{B}_2^\star}{\lambda}\log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta},\qquad\sqrt{\triangle_2\frac{\mathsf{df}_\lambda\vee(\mathsf{Q}^\star)^2}{n}\log\frac{2}{\delta}}\leq\mathsf{r}_\lambda(x^\star),\qquad 2\mathsf{b}_\lambda\leq\mathsf{r}_\lambda(x^\star).$$

The following holds, with probability at least $1-2\delta$.

$$\widehat{s}_\lambda\leq 8\ \mathsf{b}_\lambda+80\sqrt{\frac{\mathsf{df}_\lambda\vee(\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}.$$

Moreover, we also have the following bound :

$$\|\widehat{x}_\lambda^\star-x^\star\|\leq 3\ \lambda^{-1/2}\ \mathsf{b}_\lambda+8\ \lambda^{-1/2}\sqrt{\frac{\mathsf{df}_\lambda\vee(\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}.$$

Proof. We first decompose the source term into two terms, and then apply different bounds from Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) to effectively bound it. We will use the following quantity:

$$\widehat{\mathsf{Var}}_\lambda := \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|^2 \, \|\nabla \widehat{f}_\lambda(x_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(x_\lambda^\star)}.$$

It is also defined in equation (23) in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019).

1. Dividing $\widehat{s}_\lambda$ into two controllable terms . Decompose

$$\begin{aligned}\widehat{s}_\lambda = \|\lambda \widehat{x}_\lambda^\star\|_{\widehat{\mathbf{H}}_\lambda^{-1}(\widehat{x}_\lambda^\star)} &\leq \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\widehat{x}_\lambda^\star)\mathbf{H}_\lambda^{1/2}(\widehat{x}_\lambda^\star)\| \, \|\lambda \widehat{x}_\lambda^\star\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} \\ &\leq \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\widehat{x}_\lambda^\star)\mathbf{H}_\lambda^{1/2}(\widehat{x}_\lambda^\star)\| \, \left(\|\nabla f_\lambda(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} + \|\nabla f(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)}\right).\end{aligned}$$

On the one hand, from the previous proof, we get

$$\begin{aligned}\|\nabla f(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} &\leq e^{\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)/2} \, \|\widehat{x}_\lambda^\star - x^\star\|_{\mathbf{H}(x^\star)} \\ &\leq e^{\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)/2} \, \left(e^{\mathsf{t}(x_\lambda^\star - x^\star)}\|\widehat{x}_\lambda^\star - x_\lambda^\star\|_{\mathbf{H}_\lambda(x_\lambda^\star)} + \|x_\lambda^\star - x^\star\|_{\mathbf{H}_\lambda(x^\star)}\right) \\ &\leq e^{\mathsf{t}(\widehat{x}_\lambda^\star - x^\star)/2} \, \left(\frac{e^{\mathsf{t}(x_\lambda^\star - x^\star)}}{\underline{\phi}(\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star))}\widehat{\mathsf{Var}}_\lambda + \frac{1}{\underline{\phi}(\mathsf{t}(x_\lambda^\star - x^\star))}\mathsf{b}_\lambda\right).\end{aligned}$$

In the last line, we use the fact that $\|\widehat{x}_\lambda^\star - x_\lambda^\star\|_{\mathbf{H}_\lambda(x_\lambda^\star)} \leq \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\| \, \|\widehat{x}_\lambda^\star - x_\lambda^\star\|_{\widehat{\mathbf{H}}_\lambda(x_\lambda^\star)}$ and then bound it using Eq. (4.18) applied to $\widehat{f}$ to get

$$\begin{aligned}\|\widehat{x}_\lambda^\star - x_\lambda^\star\|_{\widehat{\mathbf{H}}_\lambda(x_\lambda^\star)} &\leq \frac{1}{\underline{\phi}(\widehat{\mathsf{t}}(x_\lambda^\star - \widehat{x}_\lambda^\star))}\|\nabla \widehat{f}_\lambda(x_\lambda^\star)\|_{\widehat{\mathbf{H}}_\lambda^{-1}(x_\lambda^\star)} \\ &\leq \frac{1}{\underline{\phi}(\mathsf{t}(x_\lambda^\star - \widehat{x}_\lambda^\star))}\|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\| \, \|\nabla \widehat{f}_\lambda(x_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(x_\lambda^\star)}.\end{aligned}$$

On the other hand, apply successively Eq. (4.18) to f and $\widehat{f}$ using the fact that $\widehat{\mathsf{t}} \leq \mathsf{t}$ to get

$$\begin{aligned}\|\nabla f_\lambda(\widehat{x}_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} &= \|\nabla f_\lambda(\widehat{x}_\lambda^\star) - \nabla f_\lambda(x_\lambda^\star)\|_{\mathbf{H}_\lambda^{-1}(\widehat{x}_\lambda^\star)} \\ &\leq e^{\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\overline{\phi}(\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)) \, \|\widehat{x}_\lambda^\star - x_\lambda^\star\|_{\mathbf{H}_\lambda(x_\lambda^\star)} \\ &\leq e^{\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\overline{\phi}(\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)) \, \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\| \, \|\widehat{x}_\lambda^\star - x_\lambda^\star\|_{\widehat{\mathbf{H}}_\lambda(x_\lambda^\star)} \\ &\leq \frac{e^{\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\overline{\phi}(\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star))}{\underline{\phi}(\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star))} \, \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|^2 \, \|\nabla \widehat{f}_\lambda(x_\lambda^\star)\|_{\mathbf{H}_\lambda(x_\lambda^\star)} \\ &= e^{3\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\widehat{\mathsf{Var}}_\lambda.\end{aligned}$$

Putting things together:

$$\widehat{s}_\lambda \leq \|\widehat{\mathbf{H}}_\lambda^{-1/2}(\widehat{x}_\lambda^\star)\mathbf{H}_\lambda^{1/2}(\widehat{x}_\lambda^\star)\| \, \left(e^{3\mathsf{t}(x_\lambda^\star - \widehat{x}_\lambda^\star)/2}\left(1 + \frac{1}{\underline{\phi}(\mathsf{t}(x_\lambda^\star - \widehat{x}_\lambda^\star))}\right)\widehat{\mathsf{Var}}_\lambda + \frac{e^{\mathsf{t}(x_\lambda^\star - \widehat{x}_\lambda^\star)/2}}{\underline{\phi}(\mathsf{t}(x_\lambda^\star - x^\star))}\mathsf{b}_\lambda\right).$$

2. We now import the results from Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) . The following hypotheses imply those of Thms 4 and 7 in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019):

Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Whenever

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta},\ n \geq \triangle_2\ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{\mathsf{r}_\lambda(x^\star)^2}\ \log \frac{2}{\delta}, \mathsf{b}_\lambda \leq \frac{\mathsf{r}_\lambda(x^\star)}{2}.$$

In particular, they imply that with probability at least $1 - 2\delta$:

- $\widehat{\mathsf{Var}}_\lambda \leq \frac{1}{2}\mathsf{b}_\lambda + 4\square_1 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log \frac{2}{\delta}}{n}}$;
- $\|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\| \leq \sqrt{2}$;
- $\mathsf{t}(x^\star - x_\lambda^\star) \leq \log 2$;
- $\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star) \leq \log 2$.

Hence, plugging these bounds in the previous equation, we get

$$\widehat{s}_\lambda \leq 8\mathsf{b}_\lambda + 80\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log \frac{2}{\delta}}{n}}.$$

3. Note that in what has been done previously, we can bound:

$$\|\widehat{x}_\lambda^\star - x_\lambda^\star\|_{\mathbf{H}_\lambda(x_\lambda^\star)} \leq \frac{1}{\underline{\phi}(\mathsf{t}(x_\lambda^\star - \widehat{x}_\lambda^\star))}\widehat{\mathsf{Var}}_\lambda \leq \mathsf{b}_\lambda + 8\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log \frac{2}{\delta}}{n}}.$$

Moreover,

$$\|x_\lambda^\star - x^\star\|_{\mathbf{H}_\lambda(x^\star)} \leq \frac{1}{\underline{\phi}(\mathsf{t}(x_\lambda^\star - x^\star))}\|\nabla f_\lambda(x^\star)\|_{\mathbf{H}_\lambda^{-1}(x^\star)} \leq 2\mathsf{b}_\lambda.$$

Hence:

$$\|\widehat{x}_\lambda^\star - x^\star\| \leq 3\ \lambda^{-1/2}\ \mathsf{b}_\lambda + 8\ \lambda^{-1/2}\ \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log \frac{2}{\delta}}{n}}.$$

□

2. Final bound for the projected ERM approximation In this paragraph, denote with $\mathsf{C}_\mathbf{P}(x, \lambda)$ the quantity $\frac{\|\mathbf{H}^{1/2}(x)(\mathbf{I}-\mathbf{P})\|^2}{\lambda}$ and $\widehat{\mathsf{C}}_\mathbf{P}(x, \lambda)$ the quantity $\frac{\|\widehat{\mathbf{H}}^{1/2}(x)(\mathbf{I}-\mathbf{P})\|^2}{\lambda}$

Lemma 4.19. *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Whenever*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8\square_1^2 \mathsf{B}_2^\star}{\lambda\delta}, \qquad \mathsf{C}_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log \frac{2}{\delta}}{n}} \leq \mathsf{r}_\lambda(x^\star), \qquad \mathsf{C}_1\mathsf{b}_\lambda \leq \mathsf{r}_\lambda(x^\star),$$

if

$$\mathsf{C}_\mathbf{P}(x^\star, \lambda) \leq \frac{\sqrt{2}}{480},\ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \frac{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)}{2}\ \wedge\ \frac{\mathsf{r}_\lambda(x^\star)}{126},$$

the following holds, with probability at least $1 - 2\delta$.

$$\widehat{\nu}_\lambda(x) \leq \frac{\widehat{\mathsf{r}}_\lambda(x)}{2},\ \widehat{\nu}_\lambda(x) \leq \frac{\mathsf{r}_\lambda(x^\star)}{2}.$$

Here, $\mathsf{C}_1 = 1008$.

Proof. Proceed in the following way.

1. It is easy to see that the conditions of this lemma imply the conditions of Theorem 4.8. Hence, as in the previous proofs, the following hold:

- $\|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\| \leq \sqrt{2}$;
- $\mathsf{t}(x^\star - x_\lambda^\star) \leq \log 2$;
- $\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star) \leq \log 2$.

2. Let us now apply proposition 4.15 to $\widehat{f}$. If

$$\frac{\widehat{s}_\lambda}{\widehat{\mathsf{r}}_\lambda(\widehat{x}_\lambda^\star)} \leq \frac{1}{4},\ \widehat{\mathsf{C}}_{\mathbf{P}}(\widehat{x}_\lambda^\star, \lambda) \leq \frac{1}{120},\ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \frac{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)}{2},$$

Then it holds:

$$\widehat{\nu}_\lambda(x) \leq 3(\widehat{\nu}_{\mathbf{P},\lambda}(x) + \widehat{s}_\lambda), \qquad \widehat{\mathsf{t}}(x - \widehat{x}_\lambda^\star) \leq \frac{3}{10} + 2\log 2. \tag{4.38}$$

where the second bound is obtained in the proof of this proposition. Now since

$$\begin{aligned}
\frac{1}{\widehat{\mathsf{r}}_\lambda(\widehat{x}_\lambda^\star)} &\leq e^{\widehat{\mathsf{t}}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2} \frac{1}{\widehat{\mathsf{r}}_\lambda(x_\lambda^\star)} && \text{Eq. (4.17)} \\
&\leq e^{\widehat{\mathsf{t}}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\ \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|\ \sup_{g \in \widehat{\mathcal{G}}} \|g\|_{\mathbf{H}_\lambda^{-1}(x_\lambda^\star)} && \text{Def} \\
&\leq e^{\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\ \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|\ \sup_{g \in \mathcal{G}} \|g\|_{\mathbf{H}_\lambda^{-1}(x_\lambda^\star)} && \widehat{\mathcal{G}} \subset \mathcal{G} \\
&= e^{\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star)/2}\ \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|\ \frac{1}{\mathsf{r}_\lambda(x_\lambda^\star)} && \text{Def} \\
&\leq e^{(\mathsf{t}(\widehat{x}_\lambda^\star - x_\lambda^\star) + \mathsf{t}(x_\lambda^\star - x^\star))/2}\ \|\mathbf{H}_\lambda^{1/2}(x_\lambda^\star)\widehat{\mathbf{H}}_\lambda^{-1/2}(x_\lambda^\star)\|\ \frac{1}{\mathsf{r}_\lambda(x^\star)} && \text{Eq. (4.17)} \\
&\leq \frac{2\sqrt{2}}{\mathsf{r}_\lambda(x^\star)}. && \text{previous bounds}
\end{aligned}$$

In a similar way, we get $\widehat{\mathsf{C}}_{\mathbf{P}}(\widehat{x}_\lambda^\star, \lambda) \leq 2\sqrt{2}\mathsf{C}_{\mathbf{P}}(x^\star, \lambda)$. Thus, the conditions above are satisfied if the following conditions are satisfied:

$$\frac{\widehat{s}_\lambda}{\mathsf{r}_\lambda(x^\star)} \leq \frac{\sqrt{2}}{16},\ \mathsf{C}_{\mathbf{P}}(x^\star, \lambda) \leq \frac{\sqrt{2}}{480},\ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \frac{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)}{2}.$$

Finally, note that under these conditions,

$$\frac{1}{\widehat{\mathsf{r}}_\lambda(x)} \leq \frac{e^{\widehat{\mathsf{t}}(x - \widehat{x}_\lambda^\star)/2}}{\widehat{\mathsf{r}}_\lambda(x)} \leq \frac{7}{\mathsf{r}_\lambda(x^\star)}. \tag{4.39}$$

using the previous bound and the bound on $\widehat{\mathsf{t}}(x - \widehat{x}_\lambda^\star)$.

3. Let us assume

$$\frac{\widehat{s}_\lambda}{\mathsf{r}_\lambda(x^\star)} \leq \frac{\sqrt{2}}{16},\ \mathsf{C}_{\mathbf{P}}(x^\star,\lambda) \leq \frac{\sqrt{2}}{480},\ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \frac{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)}{2}.$$

According to Eq. (4.39), and to Eq. (4.38), if

$$\widehat{\nu}_{\mathbf{P},\lambda}(x) + \widehat{s}_\lambda \leq \frac{\mathsf{r}_\lambda(x^\star)}{42},$$

then it holds

$$\widehat{\nu}_\lambda(x) \leq \frac{\widehat{\mathsf{r}}_\lambda(x)}{2},\ \widehat{\nu}_\lambda(x) \leq \frac{\mathsf{r}_\lambda(x^\star)}{2}.$$

We simplify this condition as:

$$\widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \frac{\mathsf{r}_\lambda(x^\star)}{126}, \qquad \widehat{s}_\lambda \leq \frac{2\mathsf{r}_\lambda(x^\star)}{126}.$$

4. Now using the fact that under the conditions of this lemma, those of Lemma 4.18 are satisfied:

$$\widehat{s}_\lambda \leq 8\ \mathsf{b}_\lambda + 80\ \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}.$$

Thus, $\widehat{s}_\lambda \leq \frac{2\mathsf{r}_\lambda(x^\star)}{126}$ holds, provided

$$\mathsf{b}_\lambda \leq \frac{\mathsf{r}_\lambda(x^\star)}{\mathsf{C}_1},\ n \geq \mathsf{C}_1^2 \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{\mathsf{r}_\lambda(x^\star)^2},$$

where $\mathsf{C}_1 = 1008$. □

Proposition 4.18 (Behavior of an approximation to the projected problem)**.** *Let $n \in \mathbb{N}$, $\delta \in (0,1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Let $x \in \mathcal{H}_{\mathbf{P}}$. Whenever*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}, \qquad \mathsf{C}_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}} \leq \mathsf{r}_\lambda(x^\star), \qquad \mathsf{C}_1\mathsf{b}_\lambda \leq \mathsf{r}_\lambda(x^\star),$$

if

$$\mathsf{C}_{\mathbf{P}}(x^\star,\lambda) \leq \frac{\sqrt{2}}{480},\ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \frac{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)}{2}\ \wedge\ \frac{\mathsf{r}_\lambda(x^\star)}{126}.$$

The following holds, with probability at least $1-2\delta$.

$$f(x) - f(x^\star) \leq \mathsf{K}_1\ \mathsf{b}_\lambda^2 + \mathsf{K}_2\ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n}\ \log\frac{2}{\delta} + \mathsf{K}_3\ \widehat{\nu}_{\mathbf{P},\lambda}^2(x),$$

where $\mathsf{K}_1 \leq$ 6.0e4, $\mathsf{K}_2 \leq$ 6.0e6 and $\mathsf{K}_3 \leq 810$, C_1 are defined in Lemma 4.19, and the other constants are defined in Theorem 4.8.

Remark 13 (Constants)**.** *In this result, absolutely huge constants are obtained. They are (of course) totally sub-optimal. Indeed, this analysis has been simplified by dividing the bound into blocks: error of the empirical risk minimization with regularization, error of the projection compared to this empirical risk minimizer. Going back and forth from empirical to statistical, from projected to non projected induces exponential explosion of the constants. There is a way of doing the analysis directly by projecting the statistical problem. However, in order to relate to our previous work Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) and avoid re-doing all of our work we discarded this. If we were to perform this more direct analysis, we could keep the constants to a reasonable level, of order 10^2.*

Proof. We apply Lemma 4.17, using the previous lemma to guarantee the conditions.

1. Under the conditions of this proposition, applying Lemma 4.19, the conditions of Lemma 4.17 are satisfied. Thus,

$$f(x) - f(x^\star) \leq 14(f(\widehat{x}^\star_\lambda) - f(x^\star)) + 30\widehat{\nu}_\lambda(x)^2.$$

Moreover, from the previous proof,

$$\widehat{\nu}_\lambda(x) \leq 3(\widehat{\nu}_{\mathbf{P},\lambda}(x) + \widehat{s}_\lambda),$$

and seeing as Lemma 4.18 is satisfied,

$$\widehat{s}_\lambda \leq 8\ \mathsf{b}_\lambda + 80\ \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}.$$

This therefore yields:

$$\widehat{\nu}_\lambda(x)^2 \leq 27\widehat{\nu}_{\mathbf{P},\lambda}(x)^2 + 1726\mathsf{b}_\lambda^2 + 172600\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}.$$

2. Moreover, from Theorem 4.8, it holds:

$$f(\widehat{x}^\star_\lambda) - f(x^\star) \leq 414\ \mathsf{b}_\lambda^2 + 414\ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n}\ \log\frac{2}{\delta}.$$

3. Putting things together:

$$f(x) - f(x^\star) \leq \mathsf{K}_1\ \mathsf{b}_\lambda^2 + \mathsf{K}_2\ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n}\ \log\frac{2}{\delta} + \mathsf{K}_3\ \widehat{\nu}^2_{\mathbf{P},\lambda}(x).$$

We bound the constants in the theorem.

□

Lemma 4.20. *Under the conditions of the previous theorem, the following hold:*

- $\frac{1}{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)} \leq \frac{8}{\mathsf{r}_\lambda(x^\star)}$;
- $\lambda^{1/2}\|x - x^\star\| \leq 7\widehat{\nu}_{\mathbf{P},\lambda}(x) + 59\mathsf{b}_\lambda + 568\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}$;
- $\lambda\|x\|_{\widehat{\mathbf{H}}^{-1}_{\mathbf{P},\lambda}(x)} \leq 7\widehat{\nu}_{\mathbf{P},\lambda}(x) + 72\mathsf{b}_\lambda + 720\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2\ \log\frac{2}{\delta}}{n}}$.

In particular, $\frac{\lambda\|x\|_{\widehat{\mathbf{H}}^{-1}_{\mathbf{P},\lambda}(x)}}{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)} \leq 11.$

Proof. Let us prove the three statements.

1. Write $\frac{1}{\widehat{\mathsf{r}}_{\mathbf{P},\lambda}(x)} = \sup_{g\in\widehat{\mathcal{G}}} \|\mathbf{P}g\|_{\widehat{\mathbf{H}}^{-1}_{\mathbf{P},\lambda}(x)}$. Now

$$\sup_{g\in\widehat{\mathcal{G}}} \|\mathbf{P}g\|_{\widehat{\mathbf{H}}^{-1}_{\mathbf{P},\lambda}(x)} \leq \sup_{g\in\widehat{\mathcal{G}}} \|g\|_{\widehat{\mathbf{H}}^{-1}_{\lambda}(x)} \leq e^{\widehat{\mathsf{t}}(x-\widehat{x}^\star_\lambda)/2} \sup_{g\in\widehat{\mathcal{G}}} \|g\|_{\widehat{\mathbf{H}}^{-1}_{\lambda}(\widehat{x}^\star_\lambda)}.$$

Now bound

$$\sup_{g\in\widehat{\mathcal{G}}} \|g\|_{\widehat{\mathbf{H}}^{-1}_{\lambda}(\widehat{x}^\star_\lambda)} \leq e^{\widehat{\mathsf{t}}(x^\star_\lambda-\widehat{x}^\star_\lambda)/2} \sup_{g\in\widehat{\mathcal{G}}} \|g\|_{\widehat{\mathbf{H}}^{-1}_{\lambda}(x^\star_\lambda)} \leq e^{\widehat{\mathsf{t}}(x^\star_\lambda-\widehat{x}^\star_\lambda)/2} \ \|\mathbf{H}^{1/2}_\lambda(x^\star_\lambda)\widehat{\mathbf{H}}^{-1/2}_\lambda(x^\star_\lambda)\| \ \sup_{g\in\widehat{\mathcal{G}}} \|g\|_{\mathbf{H}^{-1}_{\lambda}(x^\star_\lambda)}.$$

Finally bound

$$\sup_{g\in\widehat{\mathcal{G}}} \|g\|_{\mathbf{H}^{-1}_{\lambda}(x^\star_\lambda)} \leq e^{\mathsf{t}(x^\star-x^\star_\lambda)/2} \frac{1}{\mathsf{r}_\lambda(x^\star)}.$$

Now using the fact that under the previous assumptions $\mathsf{t}(x^\star - x^\star_\lambda), \mathsf{t}(x^\star_\lambda - \widehat{x}^\star_\lambda) \leq \log 2$, $\widehat{\mathsf{t}}(x-\widehat{x}^\star_\lambda) \leq \frac{3}{10} + 2\log 2$ and $\|\mathbf{H}^{1/2}_\lambda(x^\star_\lambda)\widehat{\mathbf{H}}^{-1/2}_\lambda(x^\star_\lambda)\| \leq \sqrt{2}$, we get the first equation.

2. In order to bound $\lambda^{1/2}\|x - x^\star\|$, decompose

$$\lambda^{1/2}\|x - x^\star\| \leq \lambda^{1/2}\|x - \widehat{x}^\star_\lambda\| + \lambda^{1/2}\|\widehat{x}^\star_\lambda - x^\star\|.$$

Now use proposition 4.15 to bound $\lambda^{1/2}\|x - \widehat{x}^\star_\lambda\| \leq 7(\widehat{\nu}_{\mathbf{P},\lambda}(x) + \widehat{s}_\lambda)$. Using Lemma 4.18, under the conditions above,

$$\widehat{s}_\lambda \leq 8 \ \mathsf{b}_\lambda + 80 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

Hence

$$\lambda^{1/2}\|x - \widehat{x}^\star_\lambda\| \leq 7\widehat{\nu}_{\mathbf{P},\lambda}(x) + 56\mathsf{b}_\lambda + 560\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

Moreover, using again Lemma 4.18

$$\lambda^{1/2}\|\widehat{x}^\star_\lambda - x^\star\| \leq 3 \ \mathsf{b}_\lambda + 8 \sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

Combining these two inequalities, we get:

$$\lambda^{1/2}\|x - x^\star\| \leq 7\widehat{\nu}_{\mathbf{P},\lambda}(x) + 59\mathsf{b}_\lambda + 568\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

3. In order to bound $\lambda\|x\|_{\widehat{\mathbf{H}}_{\mathbf{P},\lambda}^{-1}(x)}$, use proposition 4.15 to get $\lambda\|x\|_{\widehat{\mathbf{H}}_{\mathbf{P},\lambda}^{-1}(x)} \leq 7\widehat{\nu}_{\mathbf{P},\lambda}(x) + 9\widehat{s}_\lambda$.

Now using Lemma 4.18, the following bound holds:

$$\lambda\|x\|_{\widehat{\mathbf{H}}_{\mathbf{P},\lambda}^{-1}(x)} \leq 7\widehat{\nu}_{\mathbf{P},\lambda}(x) + 72\mathsf{b}_\lambda + 720\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}}.$$

□

Proposition 4.19 (Simplification). *Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$. Let $x \in \mathcal{H}_\mathbf{P}$. Whenever*

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}, \qquad \mathsf{C}_1\sqrt{\frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \ \log\frac{2}{\delta}}{n}} \leq \frac{\sqrt{\lambda}}{R}, \qquad \mathsf{C}_1\mathsf{b}_\lambda \leq \frac{\sqrt{\lambda}}{R},$$

if

$$\mathsf{C}_\mathbf{P}(x^\star, \lambda) \leq \frac{\sqrt{2}}{480}, \ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \frac{\sqrt{\lambda}}{126R},$$

then the following holds, with probability at least $1 - 2\delta$.

$$f(x) - f(x^\star) \leq \mathsf{K}_1 \ \mathsf{b}_\lambda^2 + \mathsf{K}_2 \ \frac{\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2}{n} \ \log\frac{2}{\delta} + \mathsf{K}_3 \ \widehat{\nu}_{\mathbf{P},\lambda}^2(x),$$

where $\mathsf{K}_1 \leq$ 6.0e4, $\mathsf{K}_2 \leq$ 6.0e6 and $\mathsf{K}_3 \leq 810$, C_1 are defined in Lemma 4.19, and the other constants are defined in Theorem 4.8.

Moreover, in that case, $R\|x - x^\star\| \leq 10$.

4.H .4 Optimal choice of λ, specific source conditions

In this part, we continue to assume Assumptions 4.6 to 4.9. We present a classification of distributions ρ and show that we can achieve better rates than the classical slow rates, as presented in Appendix F of Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019).

Classification of distributions and statistical bounds for the ERM

We use the following classification for distributions.

Definition 4.9 (class of distributions). *Let $\alpha \in [1, +\infty]$ and $r \in [0, 1/2]$.*
We denote with $\mathcal{P}_{\alpha,r}$ the set of probability distributions ρ such that there exists $\mathsf{L}, \mathsf{Q} \geq 0$,

- $\mathsf{b}_\lambda \leq \mathsf{L} \ \lambda^{\frac{1+2r}{2}}$;
- $\mathsf{df}_\lambda \leq \mathsf{Q}^2 \ \lambda^{-1/\alpha}$;

where this holds for any $0 < \lambda \leq 1$. For simplicity, if $\alpha = +\infty$, we assume that $\mathsf{Q} \geq \mathsf{Q}^\star$.

Note that given our assumptions, we always have

$$\rho \in \mathcal{P}_{1,0}, \ \mathsf{L} = \|x^\star\|, \ \mathsf{Q} = \mathsf{B}_1^\star. \tag{4.40}$$

We also define

$$\lambda_1 = \left(\frac{\mathsf{Q}}{\mathsf{Q}^\star}\right)^{2\alpha} \wedge 1, \tag{4.41}$$

such that

$$\forall \lambda \leq \lambda_1,\ \mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \leq \frac{\mathsf{Q}^2}{\lambda^{1/\alpha}}.$$

Interpretation of the classes

- The bias term b_λ characterizes the regularity of the objective $x^\star$. In a sense, if r is big, then this means $x^\star$ is very regular and will be easier to estimate. The following results reformulates this intuition.

  Remark 14 (source condition)**.** *Assume there exists $0 \leq r \leq 1/2$ and $v \in \mathcal{H}$ such that*

  $$\mathbf{P}_{\mathbf{H}(x^\star)} x^\star = \mathbf{H}(x^\star)^r v.$$

  Then it holds:

  $$\forall \lambda > 0,\ \mathsf{b}_\lambda \leq \mathsf{L}\ \lambda^{\frac{1+2r}{2}},\ \mathsf{L} = \|\mathbf{H}(x^\star)^{-r} x^\star\|.$$

- The effective dimension df_λ characterizes the size of the space $\mathcal{H}$ with respect to the problem. The higher α, the smaller the space. If $\mathcal{H}$ is finite dimensional for instance, $\alpha = +\infty$.

In this section, for any given pair (α, r) characterizing the regularity and size of the problem, we associate

$$\beta = \frac{1}{1 + 2r + 1/\alpha},\ \gamma = \frac{\alpha(1+2r)}{\alpha(1+2r)+1}.$$

In Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) (see corollary 3), explicit bounds are given for the performance of the regularized expected risk minimizer $\widehat{x}^\star_\lambda$ depending on which class ρ belongs to, i.e., as a function of α, r.

Corollary 4.4. *Let $\delta \in (0, 1/2]$. Under Assumptions 4.6 to 4.9, if $\rho \in \mathcal{P}_{\alpha,r}$ with $r > 0$, when $n \geq N$ and $\lambda = (C_0/n)^\beta$, then with probability at least $1 - 2\delta$,*

$$f(\widehat{x}^\star_\lambda) - f(x^\star) \leq C_1 n^{-\gamma} \log \frac{2}{\delta},$$

with $C_0 = 256(\mathsf{Q}/\mathsf{L})^2$, $C_1 = 8(256)^\gamma\ (\mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma})^2$ and N defined in Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019), and satisfying $N = O(\mathrm{poly}(\mathsf{B}_1^\star, \mathsf{B}_2^\star, \mathsf{L}, \mathsf{Q}, R, \log(1/\delta)))$.

Quantitative bounds for the projected problem

In this part, the aim is to show that if we approximately solve the projected problem up to a certain precision, then this approximation has the same statistical rates as the regularized ERM with the good choice of λ. For the sake of simplicity, we will assume that $r > 0$.

In what follows, we define

$$N = \frac{\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta} \vee \left(2.1e4 \frac{1}{1-\beta} A \log\left(1.4e6 \frac{1}{1-\beta} A^2 \frac{1}{\delta}\right)\right)^{1/(1-\beta)}, \tag{4.42}$$

where $A = \frac{\mathsf{B}_2^\star \mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$, $\lambda_0 = (\mathsf{C}_1 \mathsf{L} R \log\frac{2}{\delta})^{-1/r} \wedge 1$ and $\lambda_1 = \frac{\mathsf{Q}^{2\alpha}}{(\mathsf{Q}^\star)^{2\alpha}}$.

Theorem 4.9 (Quantitative result with source $r > 0$)**.** *Let $\rho \in \mathcal{P}_{\alpha,r}$ and assume $r > 0$. Let $\delta \in (0, \frac{1}{2}]$.*
Let $\mathbf{P}$ be an orthogonal projection, $x \in \mathcal{H}$. If

$$n \geq N,\ \lambda = \left(\left(\frac{\mathsf{Q}}{\mathsf{L}}\right)^2 \frac{1}{n}\right)^{\beta},\ \mathsf{C}_{\mathbf{P}}(x^\star, \lambda) \leq \frac{\sqrt{2}}{480},\ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma} n^{-\gamma/2}$$

then with probability at least $1 - 2\delta$,

$$f(x) - f(x^\star) \leq \mathsf{K}\left(\mathsf{Q}^\gamma\ \mathsf{L}^{1-\gamma}\right)^2\ \frac{1}{n^\gamma} \log \frac{2}{\delta},$$

where N is defined in Eq. (4.42) and $\mathsf{K} \leq 7.0e6$. Moreover, $R\|x - x^\star\| \leq 10$.

Proof. Using the definition of λ_1, as soon as $\lambda \leq \lambda_1$,it holds: $\mathsf{df}_\lambda \vee (\mathsf{Q}^\star)^2 \leq \mathsf{Q}^2 \lambda^{-1/\alpha}$.

Let us formulate proposition 4.19 using the fact that $\rho \in \mathcal{P}_{\alpha,r}$.

Let $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, $0 < \lambda \leq \mathsf{B}_2^\star$, $x \in \mathcal{H}_{\mathbf{P}}$. Whenever

$$n \geq \triangle_1 \frac{\mathsf{B}_2^\star}{\lambda} \log \frac{8 \square_1^2 \mathsf{B}_2^\star}{\lambda \delta},\ \mathsf{C}_1 \sqrt{\frac{\mathsf{Q}^2}{\lambda^{1/\alpha} n}\ \log \frac{2}{\delta}} \leq \frac{\lambda^{1/2}}{R}, \mathsf{C}_1\ \mathsf{L} \lambda^{1/2+r} \leq \frac{\lambda^{1/2}}{R},$$

if

$$\mathsf{C}_{\mathbf{P}}(x^\star, \lambda) \leq \frac{\sqrt{2}}{480},\ \widehat{\nu}_{\mathbf{P},\lambda}(x) \leq \mathsf{L} \lambda^{1/2+r},$$

The following holds, with probability at least $1 - 2\delta$.

$$f(x) - f(x^\star) \leq (\mathsf{K}_1 + \mathsf{K}_3) \mathsf{L}^2 \lambda^{1+2r} + \mathsf{K}_2\ \frac{\mathsf{Q}^2}{\lambda^{1/\alpha} n}\ \log \frac{2}{\delta}, \qquad R\|x - x^\star\| \leq 10,$$

where all constants are defined in proposition 4.19.

Assume that $r > 0$. Define

$$\lambda_0 = (\mathsf{C}_1 \mathsf{L} R \log \frac{2}{\delta})^{-1/r} \wedge 1.$$

Then for any $\lambda \leq \lambda_0$:

$$\mathsf{L} \lambda^{1/2+r} \leq \frac{1}{\mathsf{C}_1} \frac{\sqrt{\lambda}}{R}.$$

1) First, we find a simple condition to guarantee

$$\mathsf{r}_\lambda(x^\star)^2 \lambda^{1/\alpha} \geq \mathsf{C}_2\ \mathsf{Q}^2 \frac{1}{n}\ \log \frac{2}{\delta}.$$

We see that if $\lambda \leq \lambda_0$, then $\mathsf{r}_\lambda \geq \mathsf{C}_1 \mathsf{L} \lambda^{1/2+r} \log \frac{2}{\delta}$. Hence, this condition is satisfied if

$$\lambda \leq \lambda_0,\ \mathsf{C}_1^2 \mathsf{L}^2 \lambda^{1+2r+1/\alpha} \geq \mathsf{C}_2\ \mathsf{Q}^2 \frac{1}{n}.$$

Using the fact that $\mathsf{C}_2 = \mathsf{C}_1^2$, we reformulate:

$$\lambda \leq \lambda_0,\ \mathsf{L}^2\lambda^{1+2r+1/\alpha} \geq \mathsf{Q}^2\frac{1}{n}.$$

2) Now fix

$$\lambda^{1+2r+1/\alpha} = \frac{\mathsf{Q}^2}{\mathsf{L}^2}\frac{1}{n} \Longleftrightarrow \lambda = \left(\frac{\mathsf{Q}^2}{\mathsf{L}^2}\frac{1}{n}\right)^{\beta}.$$

where $\beta = 1/(1+2r+1/\lambda) \in [1/2, 1)$.

Using our restatement of proposition 4.18, with probability at least $1-2\delta$,

$$L(x) - L(x^\star) \leq \left(\mathsf{K}_1 + \mathsf{K}_3 + \mathsf{K}_2 \log\frac{2}{\delta}\right)\ \mathsf{L}^2\lambda^{1+2r} \leq \mathsf{K}\ \log\frac{2}{\delta}\ \mathsf{L}^2\lambda^{1+2r},$$

where $\mathsf{K} = \mathsf{K}_1 + \mathsf{K}_3 + \mathsf{K}_2 \leq 7.0e6$ (see proposition 4.18).
This result holds provided

$$0 < \lambda \leq \mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1,\ n \geq \triangle_1\frac{\mathsf{B}_2^\star}{\lambda}\log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}. \tag{4.43}$$

Indeed, it is shown in the previous point that the other conditions are satisfied.

3) Let us now work to guarantee the conditions in Eq. (4.43).
First, to guarantee $n \geq \triangle_1\frac{\mathsf{B}_2^\star}{\lambda}\log\frac{8\square_1^2\mathsf{B}_2^\star}{\lambda\delta}$, bound

$$\frac{\mathsf{B}_2^\star}{\lambda} = \frac{\mathsf{B}_2^\star\mathsf{L}^{2\beta}n^\beta}{\mathsf{Q}^{2\beta}\ \log^\beta\frac{2}{\delta}} \leq 2\ \frac{\mathsf{B}_2^\star\mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}n^\beta.$$

Then apply lemma 15 from Marteau-Ferey, Ostrovskii, Bach, and Rudi (2019) with $a_1 = 2\triangle_1$, $a_2 = 16\square_1^2$, $A = \frac{\mathsf{B}_2^\star\mathsf{L}^{2\beta}}{\mathsf{Q}^{2\beta}}$. Since $\beta \geq 1/2$, using the bounds in Theorem 4.8, we find $a_1 \leq 10400$ and $a_2 \leq 64$, hence the following sufficient condition:

$$n \geq \left(2.1e4\frac{1}{1-\beta}A\log\left(1.4e6\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)}.$$

Then, to guarantee the condition

$$\lambda \leq \mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1,$$

we simply need

$$n \geq \frac{\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta}.$$

Hence, defining

$$N = \frac{\mathsf{Q}^2}{\mathsf{L}^2}\left(\mathsf{B}_2^\star \wedge \lambda_0 \wedge \lambda_1\right)^{-1/\beta} \vee \left(2.1e4\frac{1}{1-\beta}A\log\left(1.4e6\frac{1}{1-\beta}A^2\frac{1}{\delta}\right)\right)^{1/(1-\beta)},$$

we see that as soon as $n \geq N$, Eq. (4.43) holds.

□

4.I Multiplicative approximations for Hermitian operators

In this section, we put together useful tools for approximating linear operators and solving linear systems with regularization.

In this section, $\mathbf{A}$ and $\mathbf{B}$ will always denote positive semi-definite Hermitian operators on a Hilbert space $\mathcal{H}$, and $\mathbf{P}$ will denote an orthogonal projection operator. Moreover, given a positive semi-definite operator $\mathbf{A}$, and $\lambda > 0$, $\mathbf{A}_\lambda$ will stand for the regularized operator $\mathbf{A} + \lambda \mathbf{I}$.

Lemma 4.21 (Equivalence of Hermitian operators)**.** *Let $\mathbf{A}$ and $\mathbf{B}$ be two semi-definite Hermitian operators. Let $\lambda > 0$. Assume you have access to*

$$t := \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B} - \mathbf{A})\mathbf{A}_\lambda^{-1/2}\|.$$

It holds:

$$\|\mathbf{A}_\lambda^{-1/2}\mathbf{B}_\lambda^{1/2}\|^2 \leq 1 + t \Leftrightarrow \mathbf{B}_\lambda \preceq (1+t)\mathbf{A}_\lambda.$$

Moreover, if $t < 1$,

$$\|\mathbf{B}_\lambda^{-1/2}\mathbf{A}_\lambda^{1/2}\|^2 \leq \frac{1}{1-t} \Leftrightarrow (1-t)\mathbf{A}_\lambda \preceq \mathbf{B}_\lambda.$$

Proof. For the first point, simply note that:

$$\|\mathbf{A}_\lambda^{-1/2}\mathbf{B}_\lambda^{1/2}\|^2 = \|\mathbf{A}_\lambda^{-1/2}\mathbf{B}_\lambda\mathbf{A}_\lambda^{-1/2}\| = \|\mathbf{I} + \mathbf{A}_\lambda^{-1/2}\left(\mathbf{B} - \mathbf{A}\right)\mathbf{A}_\lambda^{-1/2}\| \leq 1 + t.$$

For the second point,

$$\|\mathbf{B}_\lambda^{-1/2}\mathbf{A}_\lambda^{1/2}\|^2 = \|\left(\mathbf{A}_\lambda^{-1/2}\mathbf{B}_\lambda\mathbf{A}_\lambda^{-1/2}\right)^{-1}\| = \|\left(\mathbf{I} + \mathbf{A}_\lambda^{-1/2}\left(\mathbf{B} - \mathbf{A}\right)\mathbf{A}_\lambda^{-1/2}\right)^{-1}\|.$$

Moreover, we know that if $\|\mathbf{H}\| < 1$ with $\mathbf{H}$ a Hermitian operator, then $\|(\mathbf{I} + \mathbf{H})^{-1}\| \leq \frac{1}{1-\|\mathbf{H}\|}$. The result follows. □

We will now state a technical lemma which describes how combining approximation behaves.

Lemma 4.22 (Combination of approximations)**.** *Let $N \geq 1$. Let $(\mathbf{A}_i)_{1 \leq i \leq N+1}$ be a sequence of positive semi-definite Hermitian operators. Define*

$$t_i := \|\mathbf{A}_{i,\lambda}^{-1/2}(\mathbf{A}_{i+1} - \mathbf{A}_i)\mathbf{A}_{i,\lambda}^{-1/2}\|.$$

For any $1 \leq i, j \leq N+1$, define

$$t_{i:j} := \|\mathbf{A}_{i,\lambda}^{-1/2}(\mathbf{A}_j - \mathbf{A}_i)\mathbf{A}_{i,\lambda}^{-1/2}\|.$$

In particular, $t_i = t_{i:i+1}$. Then the following holds:

$$\forall 1 \leq i \leq j \leq N,\ 1 + t_{i:j} \leq \prod_{k=i}^{j-1}(1 + t_k)$$

Moreover, if $t_i < 1$, then it holds:

$$\|\mathbf{A}_{i+1,\lambda}^{-1/2}(\mathbf{A}_i - \mathbf{A}_{i+1})\mathbf{A}_{i+1,\lambda}^{-1/2}\| \leq \frac{t_i}{1 - t_i}$$

Hence, in that case

$$\forall 1 \leq j \leq i \leq N,\ 1 + t_{j:i} \leq \prod_{k=i}^{j-1}\frac{1}{1 - t_k}$$

Proof. Let us prove everything for a sequence of three operators; the rest follows by induction. Let $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ be three positive semi-definite operators.

1. Bound

$$\begin{aligned} t_{1:3} &= \|\mathbf{A}_{1,\lambda}^{-1/2} (\mathbf{A}_1 - \mathbf{A}_3) \mathbf{A}_{1,\lambda}^{-1/2}\| \\ &\leq \|\mathbf{A}_{1,\lambda}^{-1/2} (\mathbf{A}_1 - \mathbf{A}_2) \mathbf{A}_{1,\lambda}^{-1/2}\| + \|\mathbf{A}_{1,\lambda}^{-1/2} (\mathbf{A}_2 - \mathbf{A}_3) \mathbf{A}_{1,\lambda}^{-1/2}\| \\ &\leq t_{1:2} + \|\mathbf{A}_{1,\lambda}^{-1/2} \mathbf{A}_{2,\lambda}^{1/2}\|^2 t_{2:3} \\ &\leq t_{1:2} + (1 + t_{1:2}) t_{2:3}. \end{aligned}$$

The last line comes from Lemma 4.21. Thus

$$1 + t_{1:3} \leq 1 + t_{1:2} + t_{2:3} + t_{1:2} t_{2:3} = (1 + t_{1:2})(1 + t_{2:3}).$$

2. Let us now bound $t_{2:1}$ knowing $t_{1:2}$. This will imply the rest of the lemma.

Indeed, simply note that

$$t_{2:1} = \|\mathbf{A}_{2,\lambda}^{-1/2} (\mathbf{A}_2 - \mathbf{A}_1) \mathbf{A}_{2,\lambda}^{-1/2}\| \leq \|\mathbf{A}_{2,\lambda}^{-1/2} \mathbf{A}_{1,\lambda}^{1/2}\|^2 \; t_{1:2}.$$

Using Lemma 4.21, if $t_{1:2} < 1$, $\|\mathbf{A}_{2,\lambda}^{-1/2} \mathbf{A}_{1,\lambda}^{1/2}\|^2 \leq \frac{1}{1 - t_{1:2}}$, hence

$$t_{2:1} \leq \frac{t_{1:2}}{1 - t_{1:2}}.$$

□

Lemma 4.23 (Projection of Hermitian operators)**.** *For any Hermitian operator* $\mathbf{A}$ *and orthogonal projection* $\mathbf{P}$*, the following holds:*

$$\|\mathbf{A}_\lambda^{-1/2} (\mathbf{A} - \mathbf{PAP}) \mathbf{A}_\lambda^{-1/2}\| \leq \left(1 + \frac{\|\mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P})\|}{\sqrt{\lambda}}\right)^2 - 1.$$

In particular,

$$\|\mathbf{A}_\lambda^{-1/2} (\mathbf{PAP} + \lambda \mathbf{I})^{1/2} \| \leq 1 + \frac{\|\mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P})\|}{\sqrt{\lambda}}.$$

Moreover, if

$$\frac{\|\mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P})\|}{\sqrt{\lambda}} < \sqrt{2} - 1,$$

then it holds

$$\|\mathbf{A}_\lambda^{1/2} (\mathbf{PAP} + \lambda \mathbf{I})^{-1/2} \|^2 \leq \frac{1}{2 - \left(1 + \frac{\|\mathbf{A}^{1/2}(\mathbf{I} - \mathbf{P})\|}{\sqrt{\lambda}}\right)^2}.$$

We also always have:

$$\| (\mathbf{PAP} + \lambda \mathbf{I})^{-1/2} \mathbf{PA}_\lambda^{1/2}\|^2 \leq 1.$$

Proof. For any Hermitian operator $\mathbf{A}$, the following computation holds:

$$\begin{aligned}\|\mathbf{A}_\lambda^{-1/2}(\mathbf{A}-\mathbf{PAP})\mathbf{A}_\lambda^{-1/2}\| &= \|\mathbf{A}_\lambda^{-1/2}(\mathbf{A}-(\mathbf{I}-(\mathbf{I}-\mathbf{P}))\mathbf{A}(\mathbf{I}-(\mathbf{I}-\mathbf{P}))\mathbf{A}_\lambda^{-1/2}\| \\ &\leq 2\|\mathbf{A}_\lambda^{-1/2}(\mathbf{I}-\mathbf{P})\mathbf{A}\mathbf{A}_\lambda^{-1/2}\| + \|\mathbf{A}_\lambda^{-1/2}(\mathbf{I}-\mathbf{P})\mathbf{A}(\mathbf{I}-\mathbf{P})\mathbf{A}_\lambda^{-1/2}\| \\ &\leq \frac{2\|\mathbf{A}^{1/2}(\mathbf{I}-\mathbf{P})\|}{\sqrt{\lambda}} + \frac{\|\mathbf{A}^{1/2}(\mathbf{I}-\mathbf{P})\|^2}{\lambda} \\ &= \left(1+\frac{\|\mathbf{A}^{1/2}(\mathbf{I}-\mathbf{P})\|}{\sqrt{\lambda}}\right)^2 - 1.\end{aligned}$$

□

Lemma 4.24 (Relationship between approximations)**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two positive semi-definite hermitian operators. Let* $\lambda > 0$*,* $b \in \mathcal{H}$ *and* $\rho > 0$*. If*

$$\|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\| \leq \frac{1}{2} \wedge \frac{\rho}{4}, \qquad \widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{B}_\lambda, b, \rho/4),$$

then it holds:

$$\widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{A}_\lambda, b, \rho).$$

Proof. Assume $\widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{B}_\lambda, b, \tilde{\rho}/4)$ for a certain $\tilde{\rho}$. Decompose

$$\begin{aligned}\|\mathbf{A}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{A}_\lambda} &\leq \|\mathbf{A}_\lambda^{-1}b - \mathbf{B}_\lambda^{-1}b\|_{\mathbf{A}_\lambda} + \|\mathbf{B}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{A}_\lambda} \\ &\leq \|\mathbf{A}_\lambda^{1/2}(\mathbf{A}_\lambda^{-1}-\mathbf{B}_\lambda^{-1})\mathbf{A}_\lambda^{1/2}\|\ \|b\|_{\mathbf{A}_\lambda^{-1}} + \|\mathbf{A}_\lambda^{1/2}\mathbf{B}_\lambda^{-1/2}\|\ \|\mathbf{B}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{B}_\lambda}.\end{aligned}$$

Now using the fact that $\mathbf{A}_\lambda^{-1} - \mathbf{B}_\lambda^{-1} = \mathbf{B}_\lambda^{-1}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1}$,

$$\begin{aligned}\|\mathbf{A}_\lambda^{1/2}(\mathbf{A}_\lambda^{-1}-\mathbf{B}_\lambda^{-1})\mathbf{A}_\lambda^{1/2}\| &\leq \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|\ \|\mathbf{A}_\lambda^{1/2}\mathbf{B}_\lambda^{-1}\mathbf{A}_\lambda^{1/2}\| \\ &= \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|\ \|\mathbf{A}_\lambda^{1/2}\mathbf{B}_\lambda^{-1/2}\|^2.\end{aligned}$$

Moreover,

$$\|\mathbf{B}_\lambda^{-1}b - \widetilde{\Delta}\|_{\mathbf{B}_\lambda} \leq \tilde{\rho}\|b\|_{\mathbf{B}_\lambda^{-1}} \leq \|\mathbf{A}^{1/2}\mathbf{B}^{-1/2}\|\ \|b\|_{\mathbf{A}_\lambda^{-1}}.$$

Putting things together, and noting that from Lemma 4.21, $\|\mathbf{A}^{1/2}\mathbf{B}^{-1/2}\|^2 \leq \frac{1}{1-\|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|}$ as soon as $\|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\| < 1$, it holds:

$$\widetilde{\Delta} \in \mathrm{LinApprox}(\mathbf{A}_\lambda, b, \rho),\ \rho = \frac{\tilde{\rho} + \|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|}{1-\|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|}.$$

Choosing the right values for $\tilde{\rho}$ and $\|\mathbf{A}_\lambda^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}_\lambda^{-1/2}\|$ yields the result. □

4.I .1 Results for Nystrom sub-sampling

Recall the notations from Sec. 4.D .

We write without proof the following lemmas, which are just restatements of lemmas 9 and 10 of Rudi, Carratino, and Rosasco (2017).

Lemma 4.25 (Uniform sampling)**.** *Let $\delta > 0$. If $\{\tilde{z}_1, ..., \tilde{z}_m\}$ are sampled uniformly, then if $0 < \lambda < \|\mathbf{A}\|$, $m \leq n$ and*

$$m \geq \left(10 + 160\mathcal{N}_\infty^{\mathbf{A}}(\lambda)\right) \log \frac{8\|v\|_{L^\infty(Z)}^2}{\lambda\delta}.$$

Then it holds, with probability at least $1 - \delta$:

$$\|\mathbf{A}_\lambda^{-1/2}(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{A}_\lambda^{-1/2}\| \leq \frac{1}{2}, \qquad \|\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}(\widehat{\mathbf{A}} - \widehat{\mathbf{A}}_m)\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}\| \leq \frac{1}{2}.$$

Lemma 4.26 (Nystrom sampling)**.** *Let $\delta > 0$. If $\{\tilde{z}_1, ..., \tilde{z}_m\}$ are sampled using q-approximate leverage scores for $t = \lambda$, then if $t_0 \vee \frac{19\|v\|_{L^\infty(Z)}^2}{n} \log \frac{n}{2\delta} < \lambda < \|\mathbf{A}\|$, and $n \geq 405\|v\|_{L^\infty(Z)}^2 \vee 67\|v\|_{L^\infty(Z)}^2 \log \frac{12\|v\|_{L^\infty(Z)}^2}{\delta}$, if*

$$m \geq \left(6 + 486q^2\mathcal{N}^{\mathbf{A}}(\lambda)\right) \log \frac{8\|v\|_{L^\infty(Z)}^2}{\lambda\delta}.$$

Then it holds, with probability at least $1 - \delta$:

$$\|\mathbf{A}_\lambda^{-1/2}(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{A}_\lambda^{-1/2}\| \leq \frac{1}{2}, \qquad \|\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}(\widehat{\mathbf{A}} - \widehat{\mathbf{A}}_m)\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}\| \leq \frac{1}{2}.$$

Lemma 4.27. *Let $\lambda > 0$. Assume:*

$$\|\mathbf{A}_\lambda^{-1/2}(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{A}_\lambda^{-1/2}\| \leq \frac{1}{2}, \qquad \|\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}(\widehat{\mathbf{A}} - \widehat{\mathbf{A}}_m)\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}\| \leq \frac{1}{2}.$$

Denote with P_m the projection on $\operatorname{span}(v_{\tilde{z}_j})_{1 \leq j \leq m}$. Then the following holds:

$$\|\mathbf{A}_\lambda^{1/2}(\mathbf{I} - \mathbf{P}_m)\|^2 \leq 3\lambda,$$

and for any partial isometry V,

$$\frac{1}{2}\left(V^*\widehat{\mathbf{A}}_m V + \lambda\mathbf{I}\right) \preceq V^*\widehat{\mathbf{A}}V + \lambda\mathbf{I} \preceq \frac{3}{2}\left(V^*\widehat{\mathbf{A}}_m V + \lambda\mathbf{I}\right).$$

Proof. For the first point, use the well known fact that

$$\mathbf{I} - \mathbf{P}_m \leq \lambda\widehat{\mathbf{A}}_{m,\lambda}^{-1},$$

since the range of $\mathbf{P}_m$ contains that of $\widehat{\mathbf{A}}_m$. Thus,

$$\|\mathbf{A}_\lambda^{1/2}(\mathbf{I} - \mathbf{P}_m)\|^2 \leq \lambda\|\mathbf{A}_\lambda^{1/2}\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}\|^2.$$

Now using Lemma 4.22,

$$\|\mathbf{A}_\lambda^{-1/2}(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{A}_\lambda^{-1/2}\| \leq \frac{1}{2} \implies \|\widehat{\mathbf{A}}_\lambda^{-1/2}(\widehat{\mathbf{A}} - \mathbf{A})\widehat{\mathbf{A}}_\lambda^{-1/2}\| \leq 1.$$

Hence, again using Lemma 4.22,

$$\|\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}(\widehat{\mathbf{A}}_m - \mathbf{A})\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}\| \leq 2,$$

and therefore, using Lemma 4.21,

$$\|\mathbf{A}_\lambda^{1/2}\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}\|^2 \leq 3.$$

For the second point, this is only a consequence of Lemma 4.21. □

Now state two results which show that

Lemma 4.28 (Uniform sampling yielding ρ-approximation)**.** *Let $0 < \rho \leq 1$ and $\delta > 0$. Let $b \in \mathcal{H}$. If $\{\tilde{z}_1, ..., \tilde{z}_m\}$ are sampled uniformly, $0 < \lambda < \|\mathbf{A}\|$, $m \leq n$ and*

$$m \geq \left(2 + \frac{48}{\rho} + \frac{5000}{\rho^2}\mathcal{N}_\infty^{\mathbf{A}}(\lambda)\right) \log \frac{8\|v\|_{L^\infty(Z)}^2}{\lambda\delta}.$$

Then it holds, with probability at least $1 - \delta$:

$$x \in \text{LinApprox}(\widehat{\mathbf{A}}_{m,\lambda}, b, \rho/4) \implies x \in \text{LinApprox}(\mathbf{A}_\lambda, b, \rho).$$

In particular, with probability $1 - \delta$,

$$\widehat{\mathbf{A}}_{m,\lambda}^{-1} b \in \text{LinApprox}(\mathbf{A}_\lambda, b, \rho).$$

Proof. Apply Lemma 9 from Rudi, Carratino, and Rosasco (2017) with $\eta = \frac{\rho}{12} < \frac{1}{2}$. We find that under the conditions above, with probability at least $1 - \delta$,

$$\|\mathbf{A}_\lambda^{-1/2}(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{A}_\lambda^{-1/2}\| \leq \eta, \qquad \|\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}(\widehat{\mathbf{A}} - \widehat{\mathbf{A}}_m)\widehat{\mathbf{A}}_{m,\lambda}^{-1/2}\| \leq \eta.$$

Now use Lemma 4.22 to see that

$$\|\mathbf{A}_\lambda^{-1/2}(\widehat{\mathbf{A}}_m - \mathbf{A})\mathbf{A}_\lambda^{-1/2}\| \leq (1 + \eta^2) - 1 \leq 3\eta \leq \rho/4.$$

Thus, we can apply Lemma 4.24 to get the desired result. □

Lemma 4.29 (Leverage scores Nystrom sampling yielding ρ-approximation)**.** *Let $\delta > 0$. If $\{\tilde{z}_1, ..., \tilde{z}_m\}$ are sampled using q-approximate leverage scores for $t = \lambda$, then if $t_0 \vee \frac{19\|v\|_{L^\infty(Z)}^2}{n} \log \frac{n}{2\delta} < \lambda < \|\mathbf{A}\|$, and $n \geq 405\|v\|_{L^\infty(Z)}^2 \vee 67\|v\|_{L^\infty(Z)}^2 \log \frac{12\|v\|_{L^\infty(Z)}^2}{\delta}$, if*

$$m \geq \left(2 + \frac{24}{\rho} + \frac{13000q^2}{\rho^2}\mathcal{N}^{\mathbf{A}}(\lambda)\right) \log \frac{8\|v\|_{L^\infty(Z)}^2}{\lambda\delta}.$$

Then it holds, with probability at least $1 - \delta$:

$$x \in \text{LinApprox}(\widehat{\mathbf{A}}_{m,\lambda}, b, \rho/4) \implies x \in \text{LinApprox}(\mathbf{A}_\lambda, b, \rho).$$

In particular, with probability $1 - \delta$,

$$\widehat{\mathbf{A}}_{m,\lambda}^{-1} b \in \text{LinApprox}(\mathbf{A}_\lambda, b, \rho).$$

Proof. The proof is exactly the same as that of the previous lemma, using Lemma 10 instead of Lemma 9 in Rudi, Carratino, and Rosasco (2017). □

Part II

Positive semidefinite models : theory and applications

Table of Contents

Introduction : representing non-negative functions in a flexible way 223

5 PSD models 225

6 Sampling from arbitrary functions via PSD models 265

Introduction : representing non-negative functions in a flexible way

This part is based on the two articles by Marteau-Ferey, Bach, and Rudi (2020, 2022a). In this part, we develop a model for non-negative functions based on reproducing kernel Hilbert spaces, and apply this model to sample from probability distributions given their un-normalized density function, breaking the curse of dimensionality with regularity. This introduction will be less detailed than the one in the two other parts of thes thesis for two reasons. The first is that, as we introduce a new model for non-negative functions, there is less backround on this topic. The second is that these articles are short and easily readable without context. We therefore do not go too deep into the related works and results of this part to avoid repetitions, and instead present a high level overview. In particular, we freely include elements from the introductions of chapters 5 and 6 and Sec. 1.3.2

In chapter 5, we describe the work by Marteau-Ferey, Bach, and Rudi (2020). We consider a class of models with non-negative outputs, as well as outputs in an affine convex cone, which exhibit the same properties as linear models and kernel methods. This model is to consider functions parametrized by a PSD operator on a Hilbert space $\mathcal{H}$ on which the input space $\mathcal{X}$ is mapped through $\phi : \mathcal{X} \to \mathcal{H}$ with linear models notations and $x \in \mathcal{X} \mapsto k_x \in \mathcal{H}$ in RKHS notation (see Eq. (5.4) in Sec. 5.3):

$$\{f_A \ : \ A \succeq 0\} \qquad f_A(x) = \phi(x)^\top A \phi(x) \text{ or } f_A(x) = \langle k_x, A k_x \rangle_{\mathcal{H}}, \ x \in \mathcal{X}. \tag{4.44}$$

We call these models *PSD models*. As this model is itself linear, it can directly be used in the same way to solve e.r.m. problems Eqs. (1.47) and (1.50), although different spectral regularizations may be used (see chapter 5). Moreover we derive a representer theorem similar to Theorem 1.2 for our models in the context of empirical risk minimization in Theorem 5.1, and provide a convex finite-dimensional dual formulation of the learning problem, depending only on the training examples in Theorem 5.2.

In terms of statistics, we prove that the proposed model is a universal approximator and is strictly richer than commonly used generalized linear models. Moreover, we show that its Rademacher complexity is comparable with the one of kernel methods (for more details, see Sec. 5.4).

In terms of algorithms, we show the effectiveness of the method through the problems of density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression. We

derive the corresponding learning algorithms for convex dual formulation, and compare it with standard techniques used for the specific problems on a few reference simulations. In this case, we use proximal optimization techniques, and in particular FISTA, presented in Sec. 1.1.3.

In chapter 6, we describe the work by Marteau-Ferey, Bach, and Rudi (2022a). This work applies the modeling framework above to the problem of sampling n i.i.d. samples from a distribution whose density is known up to a constant through function evaluations. Contrary to most of the existing methods in the literature such as rejection sampling or Monte-Carlo Markov chain methods (Gelman, Carlin, Stern, and Rubin, 2004; Liu, 2008; Lelièvre, Rousset, and Stoltz, 2010; Robert and Casella, 2013), we propose a two-step procedure by first modelling this density using a positive semidefinite model, and then sampling from this PSD model using an adapted algorithm. In particular, we use PSD models with the Gaussian kernel defined in Sec. 1.2.2 (we will call these models Gaussian PSD models), that is we approximate the target density with a function of the form

$$f_A(x) = \sum_{i,j=1}^{n} A_{ij} k_\sigma(x_i, x) k_\sigma(x_j, x), \qquad A \succeq 0, \;\; k_\sigma(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2)). \tag{4.45}$$

Modeling probability distributions with Gaussian PSD models has been developped by Rudi and Ciliberto (2021), showing that *a)* they are stable under key operations for probabilistic inference, such as marginalization, integration (also called "sum-rule"), and product, which can be done efficiently in practice, and *b)* they concisely approximate a large class of probability distributions. The contributions of this chapter are the following.

We derive an algorithm that is easy to implement and which can generate an arbitrary number of i.i.d. samples from a given Gaussian PSD model, with any given precision (see Sec. 6.3). This answers one of the open questions outlined by Rudi and Ciliberto (2021) and shows that one can indeed efficiently sample from a Gaussian PSD model.

We then show that we can sample an arbitrary number of i.i.d. samples from a target probability distribution that is regular enough, with any given precision. The algorithm consists in (a) approximating the un-normalized density p via a PSD model, using evaluations of p, and (b) extracting i.i.d. samples from the PSD model. We show that for sufficiently regular densities the resulting PSD model is concise and avoids the curse of dimensionality : to achieve error ε (we measure this error in Wasserstein, total variation and Hellinger distance), the Gaussian PSD model requires a number of parameters and a number of evaluations of p that are in the order $\varepsilon^{-2-d/\beta}$, where d is the dimension of the space and β is the order of differentiability of the density. For regular probabilities, i.e., when $\beta \geq d$, the rate does not depend exponentially on d and is bounded by $O(\varepsilon^{-3})$ (the constant term instead may depend exponentially on d). Note that the main technical difficulty lies in showing that the approximation phase can be done in a good way statistically (see Sec. 6.4), although this is very close in terms of tools to the setting of part I and especially to the least squares setting.

Chapter 5

Non-parametric models for non-negative functions

This chapter is a verbatim of the work :

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12816–12826. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/968b15768f3d19770471e9436d97913c-Paper.pdf.

Contents

5.1 Introduction **225**
5.2 Background **226**
5.3 Proposed Model for Non-negative Functions **228**
5.4 Approximation Properties of the Model **231**
5.5 Extensions: Integral Constraints and Output in Convex Cones . . . **232**
5.6 Numerical Simulations **234**
5.A Notation and basic definitions **237**
5.B Proofs and additional discussions **237**
5.C Additional proofs **258**
5.D Additional details on the other models **258**
5.E Additional details on the experiments **261**
5.F Relation to similar work **263**

5.1 Introduction

The richness and flexibility of linear models, with the aid of possibly infinite-dimensional feature maps, allowed to achieve great effectiveness from a theoretical, algorithmic, and practical viewpoint in many supervised and unsupervised learning problems, becoming one of the workhorses of statistical machine learning in the past decades (Hastie, Tibshirani, and Friedman, 2001; Scholkopf and Smola, 2001). Indeed linear models preserve convexity of the optimization problems where they are used. Moreover they can be evaluated, differentiated and also integrated very easily.

Linear models are adapted to represent functions with unconstrained real-valued or vector-valued outputs. However, in some applications, it is crucial to learn functions with constrained outputs, such as functions which are non-negative or whose outputs are in a convex set, possibly with additional constraints like an integral equal to one, such as in density estimation, regression of multiple quantiles (Bondell, Reich, and Wang, 2010), and isotonic regression (Barlow and Brunk, 1972). Note that the convex pointwise constraints on the outputs of the learned function must hold everywhere and not only on the training points. In this context, other models have been considered, such as generalized linear models (McCullagh and Nelder, 1989), at the expense of losing some important properties that hold for linear ones.

In this paper, we make the following contributions:

- We consider a class of models with non-negative outputs, as well as outputs in a chosen convex cone, which exhibit the same key properties of linear models. They can be used within empirical risk minimization with convex risks, preserving convexity. They are defined in terms of an arbitrary feature map and they can be evaluated, differentiated and integrated exactly.
- We derive a representer theorem for our models and provide a convex finite-dimensional dual formulation of the learning problem, depending only on the training examples. Interestingly, in the proposed formulation, the convex pointwise constraints on the outputs of the learned function are naturally converted to convex constraints on the coefficients of the model.
- We prove that the proposed model is a universal approximator and is strictly richer than commonly used generalized linear models. Moreover, we show that its Rademacher complexity is comparable with the one of linear models based on kernels.
- To show the effectiveness of the method in terms of formulation, algorithmic derivation and practical results, we express naturally the problems of density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression. We derive the corresponding learning algorithms for convex dual formulation, and compare it with standard techniques used for the specific problems on a few reference simulations.

5.2 Background

In a variety of fields ranging from *supervised learning*, to *Gaussian processes* (Williams and Rasmussen, 2006), *inverse problems* (Engl, Hanke, and Neubauer, 1996), *scattered data approximation techniques* (Wendland, 2004), and *quadrature methods* to compute multivariate integrals (Bach, 2017b), prototypical problems can be cast as

$$f^* \in \arg\min_{f \in \mathcal{F}} \; L(f(x_1), \ldots, f(x_n)) + \Omega(f). \tag{5.1}$$

Here $L : \mathbb{R}^n \to \mathbb{R}$ is a (often convex) functional, $\mathcal{F}$ a class of real-valued functions, $x_1, \ldots, x_n$ a given set of points in $\mathcal{X}$, and Ω a suitable regularizer (Scholkopf and Smola, 2001).

Linear models for the class of functions $\mathcal{F}$ are particularly suitable to solve such problems. They are classically defined in terms of a feature map $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{X}$ is the input space and $\mathcal{H}$ is a separable Hilbert space. Typically, $\mathcal{H} = \mathbb{R}^D$ with $D \in \mathbb{N}$, but $\mathcal{H}$ can also be infinite-dimensional. A linear model is determined by a parameter vector $w \in \mathcal{H}$ as

$$f_w(x) = \phi(x)^\top w, \tag{5.2}$$

leading to the space $\mathcal{F} = \{f_w \mid w \in \mathcal{H}\}$. These models are particularly effective for problems in the form Eq. (5.1) because they satisfy the following key properties.

P1. They preserve convexity of the loss function. Indeed, given $x_1, \dots, x_n \in \mathcal{X}$, if $L : \mathbb{R}^n \to \mathbb{R}$ is convex, then $L(f_w(x_1), \dots, f_w(x_n))$ is convex in w.

P2. They are universal approximators. Under mild conditions on ϕ and $\mathcal{H}$ (universality of the associated kernel function (Micchelli, Xu, and Zhang, 2006)) linear models can approximate any continuous function on $\mathcal{X}$. Moreover they can represent many classes of functions of interest, such as the class of polynomials, analytic functions, smooth functions on subsets of $\mathbb{R}^d$ or on manifolds, or Sobolev spaces (Scholkopf and Smola, 2001).

P3. They admit a finite-dimensional representation. Indeed, there is a so-called *representer theorem* (Cucker and Smale, 2002). Let L be a possibly non-convex functional, $\mathcal{F} = \{f_w \mid w \in \mathcal{H}\}$, and assume Ω is an increasing function of $w^\top w$ (see the work by Schölkopf, Herbrich, and Smola (2001) for more generality and details). Then, the optimal solution f^* of (5.1) corresponds to $f^* = f_{w^*}$, with $w^* = \sum_{i=1}^n \alpha_i \phi(x_i)$, and $\alpha_1, \dots \alpha_n \in \mathbb{R}$. Denoting by k the *kernel function* $k(x, x') := \phi(x)^\top \phi(x')$ for $x, x' \in \mathcal{X}$ (see, e.g., the work by Scholkopf and Smola (2001)), f^* can be rewritten as

$$f^*(x) = \sum_{i=1}^n \alpha_i k(x, x_i). \tag{5.3}$$

P4. They are differentiable/integrable in closed form. Assume that the kernel $k(x, x')$ is differentiable in the first variable. Then $\nabla_x f_{w^*}(x) = \sum_{i=1}^n \alpha_i \nabla_x k(x, x_i)$. Also the integral of f_{w^*} can be computed in closed form if we know how to integrate k. Indeed, for $p : \mathcal{X} \to \mathbb{R}$ integrable, we have $\int f_{w^*}(x)p(x)dx = \sum_{i=1}^n \alpha_i \int k(x, x_i)p(x)dx$.

Vector-valued models. By juxtaposing scalar-valued linear models, we obtain a vector valued linear model, i.e. $f_{w_1 \cdots w_p} : \mathcal{X} \to \mathbb{R}^p$ defined as $f_{w_1 \cdots w_p}(x) = (f_{w_1}(x), \dots, f_{w_p}(x)) \in \mathbb{R}^p$.

5.2 .1 Models for non-negative functions or functions with constrained outputs

While linear models provide a powerful formalization for functions from $\mathcal{X}$ to $\mathbb{R}$ or $\mathbb{R}^p$, in some important applications arising in the context of unsupervised learning, non-parametric Bayesian methods, or graphical models, additional conditions on the model are required. In particular, we will focus on *pointwise output constraints*. That is, given $\mathcal{Y} \subsetneq \mathbb{R}^p$, we want to obtain functions satisfying $f(x) \in \mathcal{Y}$ *for all* $x \in \mathcal{X}$. A prototypical example is the problem of density estimation.

Example 5.1 (density estimation problem)**.** *The goal is to estimate the density of a probability ρ on $\mathcal{X}$, given some i.i.d. samples $x_1, \dots, x_n$. It can be formalized in terms of Eq. (5.1) (e.g., through maximum likelihood), with the constraint that f is a density, i.e., $f(x) \geq 0$, $\forall x \in \mathcal{X}$, and $\int_{\mathcal{X}} f(x)dx = 1$.*

Despite the similarity with Eq. (5.1), linear models cannot be applied because of the constraint $f(x) \geq 0$. Existing approaches to deal with the problem above are reported below, but lack some of the crucial properties **P1-4** that make linear models so effective for problems of the form Eq. (5.1).

Generalized linear models (GLM). Given a suitable map $\psi : \mathbb{R}^p \to \mathcal{Y}$, these models are of the form $f(x) = \psi(w^\top \phi(x))$. In the case of non-negative functions, common choices are $\psi(z) = e^z$, leading to the *exponential family*, or the positive part function $\psi(z) = \max(0, z)$. GLM have an expressive power comparable to linear models, being able to represent a wide class of functions, and admit a finite-dimensional representation (Cheney and Light, 2009) (they thus satisfy **P2** and **P3**). However, in general they do not preserve convexity of the functionals where they are used (except for specific cases, such as $L = -\sum_{i=1}^n \log z_i$ and $\psi(z) = e^z$ (McCullagh and Nelder, 1989)). Moreover they cannot be integrated in closed form, except for specific ϕ, requiring some Monte Carlo approximations (Robert and Casella, 2013) (thus missing **P1** and **P4**). An elegant way to obtain a GLM-like non-negative model is via *non-parametric mirror descent* (Yang, Wang, Kiyavash, and He, 2019) (see, e.g., their Example 4). A favorable feature of this approach is that the map ψ is built implicitly according to the geometry of $\mathcal{Y}$. However, still the resulting model does not always satisfy **P3**, does not satisfy **P1** and **P4** , and is only efficient in small-dimensional input spaces.

Non-negative coefficients models (NCM). Leveraging the finite-dimensional representation of linear models in Eq. (5.3), the NCM models represent non-negative functions as $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$, with $\alpha_1, \dots \alpha_n \geq 0$, given a kernel $k(x, x') \geq 0$ for any $x, x' \in \mathcal{X}$, such as the Gaussian kernel $e^{-\|x-x'\|^2}$ or the Abel kernel $e^{-\|x-x'\|}$. By construction these models satisfy **P1**, **P3**, **P4**. However, they do not satisfy **P2**. Indeed the fact that $\alpha_1, \dots, \alpha_n \geq 0$ does not allow cancellation effects and thus strongly constrains the set of functions that can be represented, as illustrated below.

Example 5.2. *The NCM model cannot approximate arbitrarily well a function with a width strictly smaller that the width of the kernel. Take $k(x, x') = e^{-\|x-x'\|^2}$ and try to approximate the function $e^{-\|x\|^2/2}$ on $[-1, 1]$. Independently of the chosen n or the chosen locations of the points $(x_i)_{i=1}^n$, it will not be possible to achieve an error smaller than a fixed constant (Sec. 5.D for a simulation).*

Partially non-negative linear models (PNM). A partial solution to have a linear model that is pointwise non-negative is to require non-negativity only on the observed points $(x_i)_{i=1}^n$. That is, the model is of the form $w^\top \phi(x)$, with $w \in \{w \in \mathcal{H} \mid w^\top \phi(x_1) \geq 0, \dots, w^\top \phi(x_n) \geq 0\}$. While this model is easy to integrate in Eq. (5.1), this does not guarantee the non-negativity outside of a neighborhood of $(x_i)_{i=1}^n$. It is possible to enrich this construction with a set of points that cover the whole space $\mathcal{X}$ (i.e., a fine grid, if $\mathcal{X} = [-1, 1]^d$), but this usually leads to exponential costs in the dimension of $\mathcal{X}$ and is not feasible when $d \geq 4$.

5.3 Proposed Model for Non-negative Functions

In this section we consider a non-parametric model for non-negative functions and we show that it enjoys the same benefits of linear models. In particular, we prove that it satisfies at the same time all the properties **P1**, ..., **P4**. As linear models, the model we consider has a simple formulation in terms of a feature map $\phi : \mathcal{X} \to \mathcal{H}$.

Let $\mathcal{S}(\mathcal{H})$ be the set of bounded Hermitian linear operators from $\mathcal{H}$ to $\mathcal{H}$ (set of symmetric $D \times D$ matrices if $\mathcal{H} = \mathbb{R}^D$ with $D \in \mathbb{N}$) and denote by $A \succeq 0$ the fact that A is a positive semi-definite operator (a positive semi-definite matrix, when $\mathcal{H}$ is finite-dimensional) (Reed, 1980; Horn and Johnson, 2012). The model is defined for all $x \in \mathcal{X}$ as

$$f_A(x) = \phi(x)^\top A \phi(x), \qquad \text{where} \qquad A \in \mathcal{S}(\mathcal{H}),\ A \succeq 0. \tag{5.4}$$

The proposed model [1] is parametrized in terms of the operator (or matrix when $\mathcal{H}$ is finite dimensional) A, as in the work by Blondel, Fujino, and Ueda (2015), but with an additional positivity constraint. Note that, by construction, it is linear in A and at the same time non-negative for any $x \in \mathcal{X}$, due to the positiveness of the operator A, as reported below (the complete proof in Sec. 5.B .1).

Proposition 5.1 (Pointwise positivity and linearity in the parameters)**.** *Given $A, B \in \mathcal{S}(\mathcal{H})$ and $\alpha, \beta \in \mathbb{R}$, then $f_{\alpha A+\beta B}(x) = \alpha f_A(x) + \beta f_B(x)$. Moreover, $A \succeq 0 \;\Rightarrow\; f_A(x) \geq 0,\ \forall x \in \mathcal{X}$.*

An important consequence of linearity of f_A in the parameter is that, despite the pointwise non-negativity in x, it preserves **P1**, i.e., the convexity of the functional where it is used. First define the set $\mathcal{S}(\mathcal{H})_+$ as $\mathcal{S}(\mathcal{H})_+ = \{A \in \mathcal{S}(\mathcal{H}) \mid A \succeq 0\}$ and note that $\mathcal{S}(\mathcal{H})_+$ is convex (Boyd and Vandenberghe, 2004).

Proposition 5.2 (The model satisfies **P1**)**.** *Let $L : \mathbb{R}^n \to \mathbb{R}$ be a jointly convex function and $x_1, \ldots, x_n \in \mathcal{X}$. Then the function $A \mapsto L(f_A(x_1), \ldots, f_A(x_n))$ is convex on $\mathcal{S}(\mathcal{H})_+$.*

proposition 5.2 is proved in Sec. 5.B .2. The property above provides great freedom in choosing the functionals to be optimized with the proposed model. However, when $\mathcal{H}$ has very high dimensionality or it is infinite-dimensional, the resulting optimization problem may be quite expensive. In the next subsection we provide a representer theorem and finite-dimensional representation for our model, that makes the optimization independent from the dimensionality of $\mathcal{H}$.

5.3 .1 Finite-dimensional representations, representer theorem, dual formulation

Here we will provide a finite-dimensional representation for the solutions of the following problem,

$$\inf_{A \succeq 0} L(f_A(x_1), \ldots, f_A(x_n)) + \Omega(A), \tag{5.5}$$

given some points $x_1, \ldots, x_n \in \mathcal{H}$. However, the existence and uniqueness of solutions for the problem above depend crucially on the choice of the regularizer Ω as it happens for linear models when $\mathcal{H}$ is finite-dimensional (Engl, Hanke, and Neubauer, 1996). To derive a representer theorem for our model, we need to specify the class of regularizers we are considering. In the context of linear models a typical regularizer is Tikhonov regularization, i.e., $\Omega(w) = \lambda w^\top w$, for $w \in \mathcal{H}$. Since the proposed model is expressed in terms of a symmetric operator (matrix, if $\mathcal{H}$ is finite-dimensional), the equivalent of the Tikhonov regularizer is a functional that penalizes the squared Frobenius norm of A, i.e., $\Omega(A) = \lambda \operatorname{Tr}(A^\top A)$, for $A \in \mathcal{S}(\mathcal{H})$ also written as $\Omega(A) = \lambda \|A\|_F^2$ (Engl, Hanke, and Neubauer, 1996). However, since A is an operator, we can also consider different norms on its spectrum. From this viewpoint, an interesting regularizer corresponds to the *nuclear norm* $\|A\|_\star$, which induces sparsity on the spectrum of A, leading to low-rank solutions (Recht, Fazel, and Parrilo, 2010; Blondel, Fujino, and Ueda, 2015). In this paper, for the sake of simplicity we will present the results for the following regularizer, which is the matrix/operator equivalent of the *elastic-net* regularizer (Zou and Hastie, 2005):

$$\Omega(A) = \lambda_1 \|A\|_\star + \lambda_2 \|A\|_F^2, \quad \forall A \in \mathcal{S}(\mathcal{H}), \tag{5.6}$$

[1] Note that the model in Eq. (5.4) has already been considered by Bagnell and Farahmand (2015) with a similar goal as ours. However, this workshop publication has only be lightly peer-reviewed, the representer theorem they propose is incorrect, the optimization algorithm is based on an incorrect representation and inefficient at best. See Appendix 5.F for details.

with $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 > 0$. Note that Ω is strongly convex as soon as $\lambda_2 > 0$; we will therefore take $\lambda_2 > 0$ in practice in order to have easier optimization. Recall the definition of the kernel $k(x, x') := \phi(x)^\top \phi(x')$, $x, x' \in \mathcal{X}$ (Scholkopf and Smola, 2001). We have the following theorem.

Theorem 5.1 (Representer theorem, **P3**). *Let $L : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be lower semi-continuous and bounded below, and Ω as in Eq. (5.6). Then Eq. (5.5) has a solution A_* which can be written as*

$$\sum_{i,j=1}^{n} \mathbf{B}_{ij} \phi(x_i)\phi(x_j)^\top, \qquad \text{for some matrix } \mathbf{B} \in \mathbb{R}^{n \times n},\ \mathbf{B} \succeq 0. \tag{5.7}$$

A_* is unique if L is convex and $\lambda_2 > 0$. By Eq. (5.4), A_* corresponds to a function of the form

$$f_*(x) = \sum_{i,j=1}^{n} \mathbf{B}_{ij} k(x, x_i) k(x, x_j), \qquad \text{for some matrix } \mathbf{B} \in \mathbb{R}^{n \times n},\ \mathbf{B} \succeq 0.$$

The proof of the theorem above is in Sec. 5.B .3, where it is derived for the more general class of spectral regularizers (this thus extends a result from Abernethy, Bach, Evgeniou, and Vert (2009), from linear operators between potentially different spaces to positive self-adjoint operators). A direct consequence of Theorem 5.1 is the following finite-dimensional representation of the optimization problem in Eq. (5.5). Denote by $\mathbf{K} \in \mathbb{R}^{n \times n}$ the matrix $\mathbf{K}_{i,j} = k(x_i, x_j)$ and assume w.l.o.g. that it is full rank (always true when the n observations are distinct and k is a *universal kernel* such as the Gaussian kernel (Micchelli, Xu, and Zhang, 2006)). Let $\mathbf{V}$ be the Cholesky decomposition of $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{V}^\top \mathbf{V}$. Define the finite dimensional model

$$\tilde{f}_{\mathbf{A}}(x) = \Phi(x)^\top \mathbf{A} \Phi(x), \qquad \mathbf{A} \in \mathbb{R}^{n \times n},\ \mathbf{A} \succeq 0, \tag{5.8}$$

where $\Phi : \mathcal{X} \to \mathbb{R}^n$, defined as $\Phi(x) = \mathbf{V}^{-\top} v(x)$, with $v(x) = (k(x, x_i))_{i=1}^n \in \mathbb{R}^n$, is the classical *empirical feature map*. In particular, $\tilde{f}_{\mathbf{A}} = f_A$ where A is of the form Eq. (5.7) with $\mathbf{B} = \mathbf{V}^{-1} \mathbf{A} \mathbf{V}^{-\top}$. We will say that $\tilde{f}_{\mathbf{A}}$ is a solution of Eq. (5.5) if the corresponding A is a solution of Eq. (5.5).

Proposition 5.3 (Equivalent finite-dimensional formulation in the primal). *Under the assumptions of Theorem 5.1, the following problem has at least one solution, which is unique if $\lambda_2 > 0$ and L is convex :*

$$\min_{\mathbf{A} \succeq 0} L(\tilde{f}_{\mathbf{A}}(x_1), \ldots, \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}). \tag{5.9}$$

Moreover, for any given solution $\mathbf{A}^* \in \mathbb{R}^{n \times n}$ of Eq. (5.9), the function $\tilde{f}_{\mathbf{A}^*}$ is a minimizer of Eq. (5.5). Finally, note that problems Eq. (5.5) and Eq. (5.9) have the same condition number if it is exists.

The proposition above (proof in Sec. 5.B .4) characterizes the possibly infinite-dimensional optimization problem of Eq. (5.5) in terms of an optimization on $n \times n$ matrices. A crucial property is that the formulation in Eq. (5.9) *preserves convexity*, i.e., it is convex as soon as L is convex. To conclude, Sec. 5.B .4 provides a construction for $\mathbf{V}$ valid for possibly rank-deficient $\mathbf{K}$. We now provide a finer characterization in terms of a dual formulation on only n variables.

Convex dual formulation. We have seen above that the problem in Eq. (5.5) admits a finite-dimensional representation and can be cast in terms of an equivalent problem on $n \times n$ matrices. Here, when L is convex, we refine the analysis and provide a dual optimization problem on only n variables. The dual formulation is particularly suitable when L is a sum of functions as we will see later. In the following theorem $[\mathbf{A}]_-$ corresponds to the negative part[2] of $\mathbf{A} \in \mathcal{S}(\mathbb{R}^n)$.

Theorem 5.2 (Convex dual problem)**.** *Assume L is convex, lower semi-continuous and bounded below. Assume Ω is of the form Eq. (5.6) with $\lambda_2 > 0$. Assume that the problem has at least a strictly feasible point, i.e., there exists $A_0 \succeq 0$ such that L is continuous in $(f_{A_0}(x_1), ..., f_{A_0}(x_n)) \in \mathbb{R}^n$ (this condition is satisfied in simple cases; see examples in Sec. 5.B .5). Denoting with L^* the Fenchel conjugate of L (Boyd and Vandenberghe, 2004), problem Eq. (5.9) has the following dual formulation:*

$$\sup_{\alpha \in \mathbb{R}^n} -L^*(\alpha) - \tfrac{1}{2\lambda_2} \|[\mathbf{V} \operatorname{diag}(\alpha) \mathbf{V}^\top + \lambda_1 \mathbf{I}]_-\|_F^2, \tag{5.10}$$

and this supremum is atteined. Moreover, if $\alpha^* \in \mathbb{R}^n$ is a solution of (5.10), a solution of (5.5) is obtained via (5.7), with $\mathbf{B} \in \mathbb{R}^{n \times n}$ defined as

$$\mathbf{B} = \lambda_2^{-1} \mathbf{V}^{-1} [\mathbf{V} \operatorname{diag}(\alpha^*) \mathbf{V}^\top + \lambda_1 \mathbf{I}]_- \mathbf{V}^{-\top}. \tag{5.11}$$

The result above (proof in Sec. 5.B .5) is particularly interesting when L can be written in terms of a sum of functions, i.e., $L(z_1, \ldots, z_n) = \sum_{i=1}^n \ell_i(z_i)$ for some functions $\ell_i : \mathbb{R} \to \mathbb{R}$. Then the Fenchel dual is $L^*(\alpha) = \sum_{i=1}^n \ell_i^*(\alpha_i)$, where ℓ_i^* is the Fenchel dual of ℓ_i, and the optimization can be carried by using accelerated proximal splitting methods as FISTA (Beck and Teboulle, 2009), since $\|[\mathbf{V} \operatorname{diag}(\alpha) \mathbf{V}^\top + \lambda_1 \mathbf{I}]_-\|_F^2$ is differentiable in α. This corresponds to a complexity of $O(n^3)$ per iteration for FISTA, due to the computation of Eq. (5.11), and can be made comparable with fast algorithms for linear models based on kernels (Rudi, Carratino, and Rosasco, 2017), by using techniques from randomized linear algebra and Nyström approximation (Halko, Martinsson, and Tropp, 2011) (see more details in Sec. 5.B .5).

5.4 Approximation Properties of the Model

The goal of this section is to study the approximation properties of our model and to understand its "richness", i.e., which functions it can represent. In particular, we will prove that, under mild assumptions on ϕ, (a) the proposed model satisfies the property **P2**, i.e., it is a *universal approximator* for non-negative functions, and (b) that it is strictly richer than the family of exponential models with the same ϕ. First, define the set of functions belonging to our model

$$\mathcal{F}_\phi^\circ = \{f_A \mid A \in \mathcal{S}(\mathcal{H}),\ A \succeq 0, \|A\|_\circ < \infty\},$$

where $\|\cdot\|_\circ$ is a suitable norm for $\mathcal{S}(\mathcal{H})$. In particular, norms that we have seen to be relevant in the context of optimization are the nuclear norm $\|\cdot\|_\star$ and the Hilbert-Schmidt (Frobenius) norm $\|\cdot\|_F$. Given norms $\|\cdot\|_a, \|\cdot\|_b$, we denote the fact that $\|\cdot\|_a$ is stronger (or equivalent) than $\|\cdot\|_b$ with $\|\cdot\|_a \trianglerighteq \|\cdot\|_b$ (for example, $\|\cdot\|_\star \trianglerighteq \|\cdot\|_F$). In the next theorem we prove that when the feature map is universal (Micchelli, Xu, and Zhang, 2006), such as the one associated to the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2)$ or the Abel kernel $k(x, x') = \exp(-\|x - x'\|)$, then the proposed model is a universal approximator for non-negative functions over $\mathcal{X}$ (in particular, in the sense of *cc-universality* (Micchelli, Xu, and Zhang, 2006; Sriperumbudur, Fukumizu, and Lanckriet, 2011), see Sec. 5.B .6 for more details and the proof).

[2] Given the eigendecomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{U} \in \mathbb{R}^{n \times n}$ unitary and $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ diagonal, then $[\mathbf{A}]_- = \mathbf{U}\mathbf{\Lambda}_-\mathbf{U}^\top$, with $\mathbf{\Lambda}_-$ diagonal, defined as $(\mathbf{\Lambda}_-)_{i,i} = \min(0, \mathbf{\Lambda}_{i,i})$ for $i = 1, \ldots, n$.

Theorem 5.3 (Universality, **P2**). *Let $\mathcal{H}$ be a separable Hilbert space, $\phi : \mathcal{X} \to \mathcal{H}$ a universal map (Micchelli, Xu, and Zhang, 2006), and $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Then $\mathcal{F}_\phi^\circ$ is a universal approximator of non-negative functions over $\mathcal{X}$.*

The fact that the proposed model can approximate arbitrarily well any non-negative function on $\mathcal{X}$, when ϕ is universal, makes it a suitable candidate in the context of nonparametric approximation/interpolation or learning (Wendland, 2004; Tsybakov, 2008) of non-negative functions. In the following theorem, we give a more precise characterization of the functions contained in $\mathcal{F}_\phi^\circ$. Denote by $\mathcal{G}_\phi$ the set of linear models induced by ϕ, i.e., $\mathcal{G}_\phi = \{w^\top \phi(\cdot) \mid w \in \mathcal{H}\}$ and by $\mathcal{E}_\phi$ the set of *exponential models* induced by ϕ,

$$\mathcal{E}_\phi \;=\; \{\; e^f \mid f(\cdot) = w^\top \phi(\cdot),\; w \in \mathcal{H} \;\}.$$

Theorem 5.4 ($\mathcal{F}_\phi^\circ$ strictly richer than the exponential model). *Let $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Let $\mathcal{X} = [-R, R]^d$, with $R > 0$. Let ϕ such that $W_2^m(\mathcal{X}) = \mathcal{G}_\phi$, for some $m > 0$, where $W_2^m(\mathcal{X})$ is the* Sobolev space *of smoothness m (Adams and Fournier, 2003). Let $x_0 \in \mathcal{X}$. The following hold:*

(a) $\mathcal{E}_\phi \subsetneq \mathcal{F}_\phi^\circ$;

(b) the function $f_{x_0}(x) = e^{-\|x - x_0\|^{-2}} \in C^\infty(\mathcal{X})$ satisfies $f_{x_0} \in \mathcal{F}_\phi^\circ$ and $f_{x_0} \notin \mathcal{E}_\phi$.

Theorem 5.4 shows that if ϕ is rich enough, then the space of exponential models is strictly contained in the space of functions associated to the proposed model. In particular, the proposed model can represent functions that are exactly zero on some subset of $\mathcal{X}$ as showed by the example f_{x_0} in Theorem 5.4, while the exponential model can represent only strictly positive functions, by construction. Discussion on the condition $W_2^m(\mathcal{X}) = \mathcal{G}_\phi$, proof of Theorem 5.4 and its generalization to $\mathcal{X} \subseteq \mathbb{R}^d$ are in App. 5.B .7. Here we note only that the condition $W_2^m(\mathcal{X}) = \mathcal{G}_\phi$ is quite mild and satisfied by many kernels such as the Abel kernel $k(x, x') = \exp(-\|x - x'\|)$ (Wendland, 2004; Berlinet and Thomas-Agnan, 2011). We conclude with a bound on the *Rademacher complexity* (Boucheron, Bousquet, and Lugosi, 2005) of $\mathcal{F}_\phi^\circ$, which is a classical component for deriving generalization bounds (Shalev-Shwartz and Ben-David, 2014). Define $\mathcal{F}_{\phi,L}^\circ = \{f_A \mid A \succeq 0, \|A\|_\circ \leq L\}$, for $L > 0$. Theorem 5.5 shows that the Rademacher complexity of $\mathcal{F}_{\phi,L}^\circ$ depends on L and not on the dimensionality of $\mathcal{X}$, as for regular kernel methods (Boucheron, Bousquet, and Lugosi, 2005).

Theorem 5.5 (Rademacher complexity of $\mathcal{F}_\phi^\circ$). *Let $\|\cdot\|_\circ \trianglerighteq \|\cdot\|_F$ and $\sup_{x \in \mathcal{X}} \|\phi(x)\| \leq c < \infty$. Let $(x_i)_{i=1}^n$ be i.i.d. samples, $L \geq 0$. The Rademacher complexity of $\mathcal{F}_{\phi,L}^\circ$ on $(x_i)_{i=1}^n$ is upper bounded by $\frac{2Lc^2}{\sqrt{n}}$ (proof in Sec. 5.B .8).*

5.5 Extensions: Integral Constraints and Output in Convex Cones

In this section we cover two extensions. The first one generalizes the optimization problem in Eq. (5.5) to include linear constraints on the integral of the model, in order to deal with problems like density estimation in Example 5.1. The second formalizes models with outputs in convex cones, which is crucial when dealing with problems like multivariate quantile estimation (Bondell, Reich, and Wang, 2010), detailed in Sec. 5.6 .

Constraints on the integral and other linear constraints. We can extend the definition of the problem in Eq. (5.5) to take into account constraints on the integral of the model. Indeed by linearity of integration and trace, we have the following (proof in Sec. 5.B .9).

Proposition 5.4 (Integrability in closed form, **P4**). *Let $A \in \mathcal{S}(\mathcal{H})$ with A bounded and ϕ uniformly bounded. Let $p : \mathcal{X} \to \mathbb{R}$ be an integrable function. There exists a trace class operator $W_p \in \mathcal{S}(\mathcal{H})$ such that $\int_{\mathcal{X}} f_A(x)p(x)dx = \mathrm{Tr}(AW_p)$ and $W_p = \int_{\mathcal{X}} \phi(x)\phi(x)^\top p(x)dx$.*

The result can be extended to derivatives and more general linear functionals on f_A (see Sec. 5.B .9). In particular, note that if we consider the *empirical feature map* Φ in Eq. (5.8), which characterizes the optimal solution of Eq. (5.5), by Theorem 5.1, we have that W_p is defined explicitly as $W_p = \mathbf{V}^{-\top}\mathbf{M}_p\mathbf{V}^{-1}$ with $(\mathbf{M}_p)_{i,j} = \int k(x, x_i)k(x, x_j)p(x)dx$, for $i, j = 1, \ldots, n$ and it is computable in closed form. Then, assuming an equality and an inequality constraint on the integral w.r.t. two functions p and q and two values $c_1, c_2 \in \mathbb{R}$, the resulting problem takes the following finite-dimensional form

$$\min_{\mathbf{A} \in \mathcal{S}(\mathbb{R}^n)} L(\tilde{f}_{\mathbf{A}}(x_1), \ldots, \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}), \quad \text{s.t. } \mathbf{A} \succeq 0,\ \mathrm{Tr}(\mathbf{A}W_p) = c_1,\ \mathrm{Tr}(\mathbf{A}W_q) \leq c_2. \tag{5.12}$$

Representing function with outputs in convex polyhedral cones. We represent a vector-valued function with our model as the juxtaposition of p scalar valued models, with $p \in \mathbb{N}$, as follows

$$f_{A_1 \cdots A_p}(x) = (f_{A_1}(x), \ldots, f_{A_p}(x)) \in \mathbb{R}^p, \qquad \forall\, x \in \mathcal{X}.$$

We recall that a convex polyhedral cone $\mathcal{Y}$ is defined by a set of inequalities as follows

$$\mathcal{Y} = \{y \in \mathbb{R}^p \mid {c^1}^\top y \geq 0, \ldots, {c^h}^\top y \geq 0\}, \tag{5.13}$$

for some $c^1, \ldots, c^h \in \mathbb{R}^p$ and $h \in \mathbb{N}$. Let us now focus on a single constraint $c^\top y \geq 0$. Note that, by definition of positive operator (i.e., $A \succeq 0$ implies $v^\top Av \geq 0$ for any A), we have that $\sum_{s=1}^p c_s A_s \succeq 0$ implies $\phi(x)^\top (\sum_{s=1}^p c_s A_s)\phi(x) \geq 0$ for any $x \in \mathcal{X}$, which, by linearity of the inner product and the definition of $f_{A_1 \cdots A_p}$ is equivalent to $c^\top f_{A_1 \cdots A_p}(x) \geq 0$. From this reasoning we derive the following proposition (see complete proof in Sec. 5.B .10).

Proposition 5.5. *Let $\mathcal{Y}$ be defined as in Eq. (5.13). Let $A_1, \ldots, A_p \in \mathcal{S}(\mathcal{H})$. Then the following holds*

$$\sum_{s=1}^p c_s^t A_s \succeq 0 \quad \forall t = 1, \ldots, h \qquad \Rightarrow \qquad f_{A_1 \cdots A_p}(x) \in \mathcal{Y} \quad \forall x \in \mathcal{X}.$$

Note that the set of constraints on the l.h.s. of the equation above defines in turn a convex set on $A_1, \ldots, A_p$. This means that we can use it to constrain a convex optimization problem over the space of the proposed vector-valued models as follows

$$\min_{A_1, \ldots, A_p \in \mathcal{S}(\mathcal{H})} L(f_{A_1 \cdots A_p}(x_1), \ldots, f_{A_1 \cdots A_p}(x_n)) + \sum_{s=1}^p \Omega(A_s) \quad \text{s.t.} \sum_{s=1}^p c_s^t A_s \succeq 0, \quad \forall\, t = 1, \ldots, h. \tag{5.14}$$

By proposition 5.5, the function $f_{A_1^* \cdots A_p^*}$, where $(A_1^*, \ldots, A_p^*)$ is the minimizer above, will be a function with output in $\mathcal{Y}$. Moreover, the formulation above admits a finite-dimensional representation analogous to the one for non-negative functions, as stated below (see proof in Sec. 5.B .11)

Theorem 5.6 (Representer theorem for model with output in convex polyhedral cones)**.** *Under the assumptions of Theorem 5.1, the problem in Eq. (5.14) admits a minimizer* $(A_1^*, \cdots, A_p^*)$ *of the form*

$$A_s^* = \sum_{i,j=1}^{n} [\mathbf{B}_s]_{i,j} \phi(x_i)\phi(x_j)^\top \implies (f_*(x))_s = \sum_{i,j=1}^{n} [\mathbf{B}_s]_{i,j} k(x_i, x)k(x_j, x), \quad s = 1, ..., p,$$

where $f_* := f_{(A_1^*,\ldots,A_p^*)}$ *is the corresponding function and the* $\mathbf{B}_s \in \mathcal{S}(\mathbb{R}^n)$ *are symmetric* $n \times n$ *matrices which satisfy the conic constraints* $\sum_{s=1}^{p} c_s^t \mathbf{B}_s \succeq 0,\ t = 1, ..., h.$

Remark 15 (Efficient representations when the ambient space of $\mathcal{Y}$ is high-dimensional)**.** *When* $p \gg h$*, or when* $\mathcal{Y}$ *is a polyhedral cone with* $\mathcal{Y} \subset \mathcal{G}$ *and* $\mathcal{G}$ *an infinite-dimensional space, it is still possible to have an efficient representation of functions with output in* $\mathcal{Y}$ *by using the representation of* $\mathcal{Y}$ *in terms of* conical hull *(Boyd and Vandenberghe, 2004), i.e.,* $\mathcal{Y} = \{\sum_{i=1}^{t} \alpha_i y_i \mid \alpha_i \geq 0\}$ *for some* $y_1, \ldots, y_t$ *and* $t \in \mathbb{N}$*. In particular, given* $A_1, \ldots, A_t \succeq 0$*, the model* $f_{A_1 \ldots A_t}(x) = \sum_{i=1}^{t} f_{A_i}(x) y_i$ *satisfies* $f_{A_1 \ldots A_t}(x) \in \mathcal{Y}$ *for any* $x \in \mathcal{X}$*. Moreover it is possible to derive a representer theorem as Theorem 5.6.*

Remark 16. *By extending this approach, we believe it is possible to model (a) functions with output in the cone of positive semidefinite matrices, (b) convex functions. We leave this for future work.*

5.6 Numerical Simulations

In this section, we provide illustrative experiments on the problems of density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression. We derive the algorithm according to the finite-dimensional formulation in Eq. (5.12) for non-negative functions with constraints on the integral, and to Eq. (5.14) with the finite-dimensional representation suggested by Theorem 5.6. Optimization is performed applying FISTA (Beck and Teboulle, 2009) on the dual of the resulting formulations. More details on implementation and the specific formulations are given below and in Sec. 5.E . The algorithms are compared with careful implementations of Eq. (5.1) with the models presented in Sec. 5.2 .1, i.e., partially non-negative models (PNM), non-negative coefficients models (NCM) and generalized linear models (GLM). For all methods we used $\Omega(A) = \lambda \left(\|A\|_* + \frac{0.01}{2} \|A\|_F^2 \right)$ or $\Omega(w) = \lambda \|w\|^2$. We used the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2))$ with width σ. Full cross-validation has been applied to each model independently, to find the best λ (see Sec. 5.E).

Density estimation. This problem is illustrated in Example 5.1. Here we considered the *log-likelihood loss* as a measure of error, i.e., $L(z_1, \ldots, z_n) = -\frac{1}{n} \sum_{i=1}^{n} \log(z_i)$, which is jointly convex and with an efficient proximal operator (Chaux, Combettes, Pesquet, and Wajs, 2007). We recall that the problems are constrained to output a function whose integral is 1. In Fig. 5.1, we show the experiment on $n = 50$ i.i.d. points sampled from $\rho(x) = \frac{1}{2}\mathcal{N}(-1, 0.3) + \frac{1}{2}\mathcal{N}(1, 0.3)$ and where for all the models we used $\sigma = 1$, to illustrate pictorially the main interesting behaviors. Instead in Sec. 5.E , we perform a multivariate experiment in $d = 10$ and $n = 1000$, where we cross-validated σ for each algorithm and show the same effects more quantitatively. Note that PNM *(left)* is non-negative on the training points, but it achieves negative values on the regions not covered by examples. This effect is worsened by the constraint on the integral that borrows areas from negative regions to reduce the log-likelihood on the dataset. NCM *(center-left)* produces a function whose integral is one and that is non-negative everywhere, but

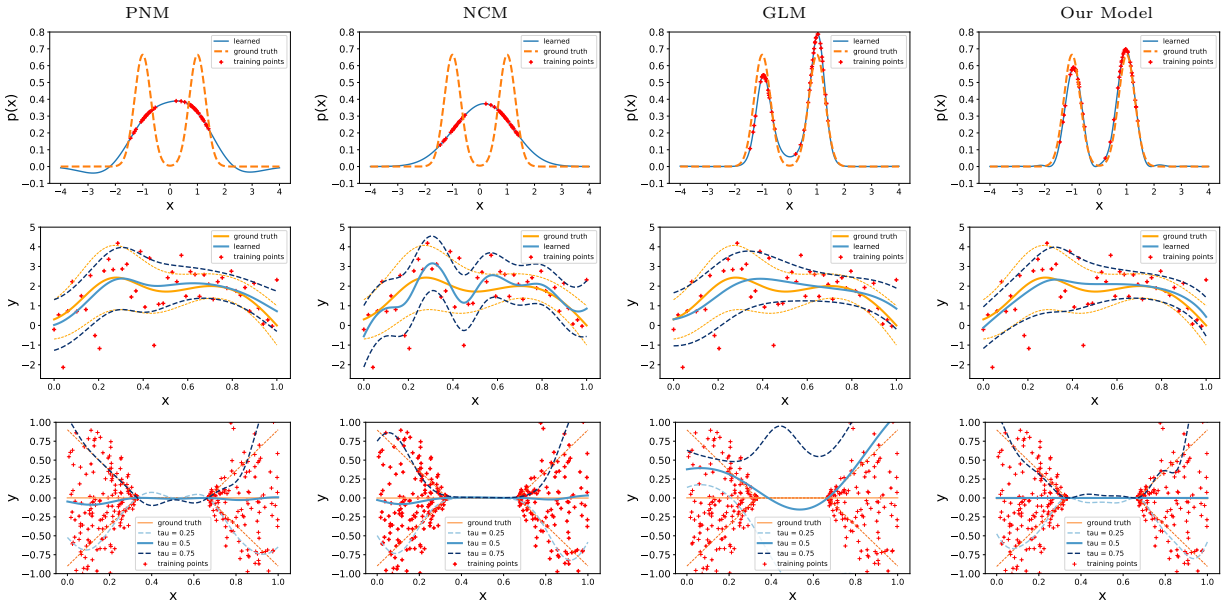

Figure 5.1: Details in Sec. 5.6 . *(top)* density estimation, *(center)* regression with Gaussian heteroscedastic errors. *(bottom)* multiple quantile regression. Shades of blue: estimated curves. Orange: ground truth. Models: *(left)* PNM, *(center-left)* NCM, *(center-right)*, GLM, *(right)* Our model.

the poor approximation properties of the model do not allow to fit the density of interest (see Example 5.2). GLM *(center-right)* produces a function that is non-negative and approximates quite well ρ, however, the obtained function does not sum to one, but to 0.987, since the integral constraint can be enforced only approximately via Monte Carlo sampling (GLM does not satisfy **P4**). Estimating the integral is easy in low dimensions but becomes soon impractical in higher dimensions (Robert and Casella, 2013). Finally the proposed model *(right)* leads to a convex problem and produces a non-negative function whose integral is 1 and that fits the density ρ quite well.

Heteroscedastic Gaussian process estimation. The goal is to estimate $\mu : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}_+$ determining the conditional density ρ of the form $\rho(y|x) = (2\pi v(x))^{-1/2} \exp(-(y - \mu(x))^2/(2v(x)))$ from which the data are sampled. The considered functional corresponds to the negative log-likelihood, i.e., $L = \sum_{i=1}^{n} \frac{1}{2} \log v(x_i) + (y_i - \mu(x_i))^2/(2v(x_i))$ that becomes convex in η, θ via the so called *natural parametrization* $\eta(x) = \mu(x)/v(x)$ and $\theta(x) = 1/v(x)$ (Le, Smola, and Canu, 2005). We used a linear model to parametrize η and the non-negative models for θ. The experiment on the same model of (Le, Smola, and Canu, 2005; Yuan and Wahba, 2004) is reported in Fig. 5.1. Modeling θ via PNM *(left)* leads to a convex problem and reasonable performance. In particular, the fact that $\theta = 0$ corresponds to $v = +\infty$ prevents the model for θ from crossing zero. NCM *(center-left)* leads to a convex problem, but very sensitive to the kernel width σ and with poor approximation properties. GLM *(center-right)* leads to a non-convex problem and we need to restart the method randomly to have a reasonable convergence. Our model *(right)* leads to a convex problem and produces a non-negative function for θ, that fits well the observed data.

Multiple quantile regression. The goal here is to estimate multiple quantiles of a given conditional distribution $P(Y|x)$. Given $\tau \in (0, 1)$, q_τ defined by $P(Y > q_\tau(x)|x) = \tau$ is the τ-quantile of ρ. By construction $0 < \tau_{-h} \leq \cdots \leq \tau_h < 1$ implies $q_{\tau_{-h}}(x) \leq \cdots \leq q_{\tau_h}(x)$. If

we denote by $\mathbf{q} : \mathcal{X} \to \mathbb{R}^{2h+1}$ the list of quantiles, we have by construction $\mathbf{q}(x) \in \mathcal{Y}$ where $\mathcal{Y}$ is a convex cone $\mathcal{Y} = \{y \in \mathbb{R}^h \mid y_{-h} \leq \cdots \leq y_h\}$. To regress quantiles, we used the pinball loss L_τ (convex, non-smooth) considered by Koenker (2005); Steinwart and Christmann (2011), obtaining $L = \sum_{j=-h}^{h} \sum_{i=1}^{n} L_{\tau_j}(f(x_i), y_i)$. In Fig. 5.1, we used $\tau_{-1} = \frac{1}{4}, \tau_0 = \frac{1}{2}, \tau_1 = \frac{3}{4}$. Using PNM, *(left)* the ordering is enforced by explicit constraints on the observed dataset (Takeuchi, Le, Sears, and Smola, 2005; Bondell, Reich, and Wang, 2010). The resulting problem is convex. However, in regions with low density of points, PNM quantiles do not respect their natural order. To enforce the order constraint, a fine grid covering the space would be needed (Takeuchi, Le, Sears, and Smola, 2005). For NCM, GLM and our model, we represented the quantiles as $q_{\tau_{\pm j}} = q_{\tau_0} \pm \sum_{i=1} v_{\pm i}$ where the v's are non-negative functions and q_{τ_0}, with $\tau_0 = \frac{1}{2}$, is the median and is modeled by a linear model. NCM *(center-left)* leads to a convex problem and quantiles that respect the ordering, but the estimation is very sensitive to the chosen σ and has poor approximation properties. GLM *(center-right)* leads to a non-convex non-differentiable problem, with many local minima, which is difficult to optimize with standard techniques (see Sec. 5.E). GLM does not succeed in approximating the quantiles. Our model *(right)* leads to a convex optimization problem that approximates the quantiles relatively well and preserves their natural order everywhere.

Acknowledgments

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19- P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support of the European Research Council (grant SEQUOIA 724063).

5.A Notation and basic definitions

- $\mathcal{H}$ is a separable Hilbert space.
- $\mathcal{X}$ is a Polish space (we will require explicitly compactness in some theorems).
- $\phi : \mathcal{X} \to \mathcal{H}$ is a continous map. We also assume it to be uniformly bounded i.e. $\sup_{x\in\mathcal{X}} \|\phi(x)\| \leq c$ for some $c \in (0, \infty)$, if not differently stated.
- $k(x, x') := \phi(x)^\top \phi(x')$ is the *kernel* function associated to the feature map ϕ (Scholkopf and Smola, 2001; Berlinet and Thomas-Agnan, 2011).

5.B Proofs and additional discussions

5.B .1 Proof of proposition 5.1

In this section, let us extend the definition in Eq. (5.4) to any operator $A \in \mathcal{S}(\mathcal{H})$, without the implied positivity restriction (in Eq. (5.4), we ask that $A \succeq 0$) :

$$\forall A \in \mathcal{S}(\mathcal{H}),\ \forall x \in \mathcal{X},\ f_A(x) := \phi(x)^\top A\phi(x). \tag{4bis}$$

Proof of proposition 5.1. To prove linearity, let $A, B \in \mathcal{S}(\mathcal{H})$ and $\alpha, \beta \in \mathbb{R}$. Since $\mathcal{S}(\mathcal{H})$ is a vector space, $\alpha A + \beta B \in \mathcal{S}(\mathcal{H})$. Let $x \in \mathcal{X}$. By definition for the first equality and linearity for the second,

$$f_{\alpha A+\beta B}(x) = \phi(x)^\top(\alpha A + \beta B)\phi(x) = \alpha\phi(x)^\top A\phi(x) + \beta\phi(x)^\top B\phi(x).$$

Finally, since by definition, $f_A(x) = \phi(x)^\top A\phi(x)$ and $f_B(x) = \phi(x)^\top B\phi(x)$, it holds :

$$f_{\alpha A+\beta B}(x) = \alpha\phi(x)^\top A\phi(x) + \beta\phi(x)^\top B\phi(x) = \alpha f_A(x) + \beta f_B(x).$$

Since this holds for all $x \in \mathcal{X}$, this shows $f_{\alpha A+\beta B} = \alpha f_A + \beta f_B$.

To prove the non-negativity, assume now that $A \succeq 0$. By definition of of positive semi-definiteness,

$$\forall h \in \mathcal{H},\ h^\top Ah \geq 0.$$

In particular, for any $x \in \mathcal{X}$, the previous inequality applied to $h = \phi(x)$ yields

$$f_A(x) = \phi(x)^\top A\phi(x) \geq 0.$$

Hence, $f_A \geq 0$.

□

5.B .2 Proof of proposition 5.2

Recall the definition of f_A for any $A \in \mathcal{S}(\mathcal{H})$ in Eq. (4bis). We have the lemma:

Lemma 5.1 (Linearity of evaluations)**.** *Let $x_1, \ldots, x_n \in \mathcal{X}$. Then the map*

$$A \in \mathcal{S}(\mathcal{H}) \mapsto (f_A(x_i))_{1\leq i\leq n} \in \mathbb{R}^n$$

is linear from $\mathcal{S}(\mathcal{H})$ to $\mathbb{R}^n$.

Proof. This just follows from the fact that the definition of $f_A(x_i)$, $f_A(x_i) := \phi(x_i)^\top A \phi(x_i)$, is linear in A. □

Proof of proposition 5.2. Let $L : \mathbb{R}^n \to \mathbb{R}$ be a jointly convex function and $x_1, \dots, x_n \in \mathcal{X}$. The function $A \in \mathcal{S}(\mathcal{H}) \mapsto L(f_A(x_1), \dots, f_A(x_n))$ can be written $L \circ R$, where

$$R : A \in \mathcal{S}(\mathcal{H}) \mapsto (f_A(x_i))_{1 \leq i \leq n} \in \mathbb{R}^n.$$

Since L is convex, and R is linear by Lemma 5.1, their composition is convex.

Moreover, since $\mathcal{S}(\mathcal{H})_+$ is a convex subset of $\mathcal{S}(\mathcal{H})$, the restriction of $A \in \mathcal{S}(\mathcal{H}) \mapsto L(f_A(x_1), \dots, f_A(x_n))$ on $\mathcal{S}(\mathcal{H})_+$ is also convex. □

5.B .3 Proof of Theorem 5.1

In this section, we prove Theorem 5.1 for a more general class of spectral regularizers.

Compact operators and spectral functions

In this section, we briefly introduce compact self-adjoint operators and the spectral theory of compact self-adjoint operators. For more details, see for instance the work by Gohberg, Goldberg, and Kaashoek (2004). We start by defining a compact self-adjoint operator (Gohberg, Goldberg, and Kaashoek, 2004, Section 2.16) and stating its main properties:

Definition 5.1 (compact operators)**.** *Let $\mathcal{H}$ be a separable Hilbert space. A bounded self-adjoint operator $A \in \mathcal{S}(\mathcal{H})$ is said to be compact if its range is included in a compact set. We denote with $\mathcal{S}_\infty(\mathcal{H})$ the set of compact self adjoint operators on $\mathcal{H}$. It is a closed subspace of $\mathcal{S}(\mathcal{H})$ for the operator norm and the closure of the set of finite rank operators.*

Proposition 5.6 (Gohberg, Goldberg, and Kaashoek (2004, Spectral theorem))**.** *Let $\mathcal{H}$ be a separable Hilbert space and let A be a compact self adjoint operator on $\mathcal{H}$. Then there exists a spectral decomposition of A, i.e., an orthonormal system $(u_k) \in \mathcal{H}$ of eigenvectors of A and corresponding eigenvalues (σ_k) such that for all $h \in \mathcal{H}$, it holds*

$$Ah = \sum_k \sigma_k u_k^\top h \; u_k =: \left(\sum_k \sigma_k u_k u_k^\top \right) h.$$

Moreover, if σ_k is an infinite sequence, it converges to zero.
Furthermore, we say that the orthonormal system (u_k) of eigenvectors of A and the corresponding eigenvalues (σ_k) is a basic system *of eigenvectors of A if all the σ_k are non zero. In this case, if P_0 denotes the orthogonal projection on* $\mathrm{Ker}(A)$*, then it holds*

$$\forall h \in \mathcal{H}, \; h = \Pi_0 \; h + \sum_k u_k u_k^\top \; h$$

In what follows, to simplify notations, we will usually write $A = U \operatorname{diag}(\sigma) U^\top$ in order to denote a basic system of eigenvectors of A. Moreover, if A is positive semi-definite, we will assume that the eigenvalues are sorted in decreasing order, i.e., $\sigma_{k+1} \leq \sigma_k$.

Definition 5.2 (Spectral function on $\mathcal{S}_\infty(\mathcal{H})$ (Gohberg, Goldberg, and Kaashoek, 2004))**.** *Let $q : \mathbb{R} \to \mathbb{R}$ be a lower semi-continuous function such that $q(0) = 0$. Let $\mathcal{H}$ be any separable*

Hilbert space. For any $A \in \mathcal{S}_\infty(\mathcal{H})$ and any basic system $A = U \operatorname{diag}(\sigma) U^\top$, we define the spectral function q

$$q(A) = U \operatorname{diag}(q(\sigma)))U^T = \sum_k q(\sigma_k) u_k u_k^\top.$$

Classes of regularizers

Let us now state our main assumption on regularizers.

Assumption 5.1 (Assumption on regularizers)**.** *Ω is of of the form*

$$\forall A \in \mathcal{S}(\mathcal{H}),\ \Omega(A) = \begin{cases} \operatorname{Tr}(q(A)) = \sum_k q(\sigma_k) & \text{if } A = U \operatorname{diag}(\sigma) U^\top \in \mathcal{S}_\infty(\mathcal{H}),\ \sum_k q(\sigma_k) < \infty \\ +\infty & \text{otherwise,} \end{cases}$$

where $q : \mathbb{R} \to \mathbb{R}_+$ is:

- *non-decreasing on $\mathbb{R}_+$ with $q(0) = 0$;*
- *lower semi-continuous;*
- $q(\sigma) \underset{|\sigma| \to +\infty}{\longrightarrow} +\infty$.

Note that in this case, Ω is defined on $\mathcal{S}(\mathcal{H})$ for any Hilbert space $\mathcal{H}$.

Remark 17. *$\Omega(A) = \lambda_1 \|A\|_\star + \frac{\lambda_2}{2} \|A\|_F^2$ satisfies Assumption 5.1, with $q(\sigma) = \lambda_1 \ |\sigma| + \lambda_2 \ \sigma^2$.*

Lemma 5.2 (Properties of Ω)**.** *Let Ω satisfying Assumption 5.1. Then the following properties hold.*

(i) For any separable Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and any linear isometry $O : \mathcal{H}_1 \to \mathcal{H}_2$, i.e., such that $O^*O = I_{\mathcal{H}_1}$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H}_1),\ \Omega(OAO^*) = \Omega(A).$$

(ii) For any separable Hilbert space $\mathcal{H}$ and any orthogonal projection $\Pi \in \mathcal{S}(\mathcal{H}_1)$, i.e., satisfying $\Pi = \Pi^*,\ \Pi^2 = \Pi$, it holds

$$\forall A \succeq 0,\ \Omega(\Pi A \Pi) \leq \Omega(A).$$

(iii) For any finite dimensional Hilbert space $\mathcal{H}_n$,

$$\Omega \text{ is lower semi-continuous (l.s.c)}, \qquad \Omega(A) \underset{\|A\|_{op} \to +\infty}{\longrightarrow} +\infty$$

where we denoted by $\|\cdot\|_{op}$ the operator norm.

Proof. (i) Write $A = \sum_k \sigma_k u_k u_k^\top$ where the (u_k) form a basic system of eigen-vectors for A. The $(v_k) = (Ou_k)$ form a basic system of eigen-vectors for OAO^*, as

$$OAO^* = \sum_k \sigma_k v_k v_k^\top, \qquad \sigma_k \neq 0.$$

Hence, by definition, $q(OAO^*) = \sum_k q(\sigma_k) v_k v_k^\top$. By definition of the trace, we have

$$\Omega(OAO^*) = \sum_k q(\sigma_k) = \Omega(A).$$

(ii) Let A be a compact self-adjoint semi-definite operator. Let $A = U\,\mathrm{diag}(\sigma)U^\top$ be a basic system of eigenvectors of A, where the σ_k are positive and in decreasing order. Define $B = U\,\mathrm{diag}(\sqrt{\sigma})U^\top$ and note that in this case, $A = B^2 = B^*B$. Using Exercise 23 by Gohberg, Goldberg, and Kaashoek (2004), we have that for any orthogonal projection operator Π and any index k, $\sigma_k(\Pi B^* B \Pi) \leq \sigma_k(B^*B)$ and hence $\sigma_k(\Pi A \Pi) \leq \sigma_k(A)$. Since q is non decreasing, it holds $q(\sigma_k(\Pi A\Pi)) \leq q(\sigma_k(A))$ and hence

$$\Omega(\Pi A \Pi) = \sum_k q(\sigma_k(\Pi A \Pi)) \leq \sum_k q(\sigma_k(A)) = \Omega(A).$$

(iii) Let $\mathcal{H}_n$ be a finite dimensional Hilbert space and let $\|\cdot\|_{op}$ be the operator norm on $\mathcal{S}(\mathcal{H}_n)$. If q is continuous, then $A \in \mathcal{H}_n \mapsto q(A)$ is continuous and hence Ω is continuous (since the trace is continuous in finite dimensions). Now assume q is lower semi-continuous, and define for $n \in \mathbb{N}$, $q_n(t) := \inf_{s\in\mathbb{R}} q(s) + n|t-s|$. We have $q_n \geq 0$, $q_n(0) = 0$ q_n is uniformly continuous and q_n is an increasing sequence of functions such that $q_n \to q$ point-wise. Now it is easy to see that $\mathrm{Tr}(q(A)) = \sup_n \mathrm{Tr}(q_n(A))$ and hence Ω is lower semi-continuous as a supremum of continuous functions.
The fact that Ω goes to infinity is a direct consequence of the fact that q goes to infinity, by Assumption 5.1.

□

Remark 18. *The three conditions of the previous lemma are in fact the only conditions needed in the proof. We could loosen Assumption 5.1 to satisfy only these three properties.*

Finite-dimensional representation and existence of a solution

Fix $n \in \mathbb{N}$, a loss function $L : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, a separable Hilbert space $\mathcal{H}$, a regularizer Ω on $\mathcal{S}(\mathcal{H})$ a feature map $\phi : \mathcal{X} \to \mathcal{H}$ and points $(x_1, ..., x_n) \in \mathcal{X}^n$.

Recall the problem in Eq. (5.5):

$$\inf_{A\succeq 0} L(f_A(x_1), \ldots, f_A(x_n)) + \Omega(A). \tag{5.5}$$

Define $\mathcal{H}_n$ to be the finite-dimensional subset of $\mathcal{H}$ spanned by the $\phi(x_i)$, i.e.,

$$\mathcal{H}_n := \mathrm{span}\,(\phi(x_i))_{1\leq i\leq n} = \left\{ \sum_{i=1}^n \alpha_i \phi(x_i) \;:\; \alpha \in \mathbb{R}^n \right\}.$$

Define Π_n is the orthogonal projection on $\mathcal{H}_n$, i.e.,

$$\Pi_n \in \mathcal{S}(\mathcal{H}),\ \Pi_n^2 = \Pi_n,\ \mathrm{range}(\Pi_n) = \mathcal{H}_n.$$

Define $\mathcal{S}_n(\mathcal{H})_+$ to be the following subspace of $\mathcal{S}(\mathcal{H})_+$:

$$\mathcal{S}_n(\mathcal{H})_+ := \Pi_n \mathcal{S}(\mathcal{H})_+ \Pi_n = \{\Pi_n A \Pi_n \;:\; A \in \mathcal{S}(\mathcal{H})_+\}.$$

Proposition 5.7. *Let L be a lower semi-continuous function which is bounded below, and assume Ω satisfies Assumption 5.1. Then Eq. (5.5) has a solution A^* which is in $\mathcal{S}_n(\mathcal{H})_+$.*

Proof. In this proof, denote by J the function defined by

$$\forall A \in \mathcal{S}(\mathcal{H}),\ J(A) := L(f_A(x_1), ..., f_A(x_n)) + \Omega(A).$$

Our goal is to prove that the problem $\inf_{A\in\mathcal{S}(\mathcal{H})_+} J(A)$ has a solution which is in $\mathcal{S}_n(\mathcal{H})_+$, i.e., of the form $\Pi_n A \Pi_n$ for some $A \in \mathcal{S}(\mathcal{H})_+$.

1. Let us start by fixing $A \in \mathcal{S}(\mathcal{H})_+$.
First note that since Π_n is the orthogonal projection on $\mathrm{span}(\phi(x_i))_{1\le i\le n}$, in particular $\Pi_n\phi(x_i) = \phi(x_i)$ for all $1 \le i \le n$. Thus, for any $1 \le i \le n$,

$$f_A(x_i) = \phi(x_i)^\top A\phi(x_i) = \phi(x_i)^\top \Pi_n A\Pi_n\phi(x_i) = f_{\Pi_n A\Pi_n}(x_i).$$

Here, the first and last equalities come from the definition of f_A and $f_{\Pi_n A\Pi_n}$. Thus,

$$J(A) = L(f_{\Pi_n A\Pi_n}(x_1), ..., f_{\Pi_n A\Pi_n}(x_n)) + \Omega(A).$$

Now since Ω satisfies Assumption 5.1, by the second point of Lemma 5.2, it holds $\Omega(\Pi_n A\Pi_n) \le \Omega(A)$, hence

$$J(\Pi_n A\Pi_n) \le J(A).$$

This last inequality combined with the fact that $\mathcal{S}_n(\mathcal{H})_+ = \Pi_n\mathcal{S}(\mathcal{H})_+\Pi_n \subset \mathcal{S}(\mathcal{H})_+$ show that

$$\inf_{A\in\mathcal{S}_n(\mathcal{H})_+} J(A) = \inf_{A\succeq 0} J(A). \tag{5.15}$$

2. Let us now show that $\inf_{A\in\mathcal{S}_n(\mathcal{H})_+} J(A)$ has a solution. Let us exclude the case where $J = +\infty$, in which case $A = 0$ can be taken to be a solution.

Let V_n be the injection $V_n : \mathcal{H}_n \hookrightarrow \mathcal{H}$. Note that $V_nV_n^* = \Pi_n$ and $V_n^*V_n = I_{\mathcal{H}_n}$. These simple facts easily show that

$$\mathcal{S}_n(\mathcal{H})_+ = V_n\mathcal{S}(\mathcal{H}_n)_+V_n^* = \left\{V_n\tilde{A}V_n^* \ :\ \tilde{A} \in \mathcal{S}(\mathcal{H}_n)_+\right\}.$$

Thus, our goal is to show that $\inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+} J(V_nAV_n^*)$ has a solution.

By the first point of Lemma 5.2, since $V_n^*V_n = I_{\mathcal{H}_n}$, it holds

$$\forall \tilde{A} \in \mathcal{S}(\mathcal{H}_n),\ \Omega(V_n\tilde{A}V_n^*) = \Omega(\tilde{A}) \implies J(V_n\tilde{A}V_n^*) = L(f_{V_n\tilde{A}V_n^*}(x_1), ..., f_{V_n\tilde{A}V_n^*}(x_n)) + \Omega(\tilde{A}).$$

Let $\tilde{A}_0 \in \mathcal{S}(\mathcal{H}_n)_+$ be a point such that $J_0 := J(V_n\tilde{A}_0V_n^*) < \infty$. Let c_0 be a lower bound for L. By the third point of Lemma 5.2, there exists a radius R_0 such that for all $\tilde{A} \in \mathcal{S}(\mathcal{H}_n)$,

$$\|\tilde{A}\|_F > R_0 \implies \Omega(\tilde{A}) > J_0 - c_0.$$

Since c_0 is a lower bound for L, this implies

$$\inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+} J(V_n\tilde{A}V_n^*) = \inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+,\ \|\tilde{A}\|_F\le R_0} J(V_n\tilde{A}V_n^*).$$

Now since L is lower semi-continuous, Ω is lower semi-continuous by the last point of Lemma 5.2, and $\tilde{A} \mapsto (f_{V_n\tilde{A}V_n^*}(x_i))_{1\le i\le n}$ is linear hence continuous, the mapping $A \mapsto J(V_n\tilde{A}V_n^*)$ is lower semi-continuous. Hence, it reaches its minimum on any non empty compact set. Since $\mathcal{H}_n$ is finite dimensional, the set $\left\{\tilde{A} \in \mathcal{S}(\mathcal{H}_n)_+ \ :\ \|\tilde{A}\|_F \le R_0\right\}$ is compact (closed and bounded) and non empty since it contains $\tilde{A}_0$, and hence there exists $\tilde{A}_* \in \mathcal{S}(\mathcal{H}_n)_+$ such that $J(V_n\tilde{A}_*V_n^*) = \inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+,\ \|\tilde{A}\|_F\le R_0} J(V_n\tilde{A}V_n^*)$. Going back up the previous equalities, this shows that $A_* = V_n\tilde{A}_*V_n^* \in \mathcal{S}_n(\mathcal{H})_+$ and $J(A_*) = \inf_{A\succeq 0} J(A)$. □

Proof of Theorem 5.1

We will prove the following Theorem 5.7 whose statement is that of Theorem 5.1 with more general assumptions.

Theorem 5.7. *Let L be lower semi-continuous and bounded below, and Ω satisfying Assumption 5.1. Then Eq. (5.5) has a solution A_* which can be written in the form*

$$\sum_{i,j=1}^{n} \mathbf{B}_{ij}\phi(x_i)\phi(x_j)^\top, \qquad \text{for some matrix } \mathbf{B} \in \mathbb{R}^{n\times n},\ \mathbf{B} \succeq 0.$$

Moreover, if L is convex, and Ω is of the form Eq. (5.6) with $\lambda_2 > 0$, this solution is unique. By Eq. (5.4), A_* corresponds to a function of the form

$$f_*(x) = \sum_{i,j=1}^{n} \mathbf{B}_{ij}k(x, x_i)k(x, x_j).$$

Lemma 5.3. *The set $\mathcal{S}_n(\mathcal{H})_+$ can be represented in the following way*

$$\mathcal{S}_n(\mathcal{H})_+ = \left\{ \sum_{1\le i,j\le n} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top,\ :\ \mathbf{B} \in \mathbb{R}^{n\times n},\ \mathbf{B} \succeq 0 \right\}.$$

In particular, for any $A \in \mathcal{S}_n(\mathcal{H})_+$, there exists a matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \succeq 0$ such that

$$A = \sum_{1\le i,j\le n} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top \implies \forall x \in \mathcal{X},\ f_A(x) = \sum_{1\le i,j\le n} \mathbf{B}_{i,j}k(x_i, x)k(x_j, x).$$

Proof. Define $S_n : \mathcal{H} \to \mathbb{R}^n$ to be the operator such that

$$\forall h,\ S_n(h) = \left(h^\top\phi(x_i)\right)_{1\le i\le n},$$

with adjoint $S_n^* : \mathbb{R}^n \to \mathcal{H}$ such that

$$\forall \alpha \in \mathbb{R}^n,\ S_n^*\alpha = \sum_{i=1}^{n} \alpha_i\phi(x_i).$$

Note that for any $\mathbf{B} \in \mathbb{R}^{n\times n}$, $S_n^*\mathbf{B}S_n = \sum_{i,j} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top$.

1. Proving $\mathcal{S}_n(\mathcal{H})_+ \subset \left\{\sum_{1\le i,j\le n} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top,\ :\ \mathbf{B} \in \mathbb{R}^{n\times n},\ \mathbf{B} \succeq 0\right\}$. Let $\Pi_n A\Pi_n \in \mathcal{S}_n(\mathcal{H})_+$. Using the previous equality, we want to show there exists $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \succeq 0$ such that $\Pi_n A\Pi_n = S_n^*\mathbf{B}S_n$. Using Lemma 5.4, we see that Π_n can be written in the form $S_n^*T_n$ where $T_n : \mathcal{H} \to \mathbb{R}^n$ (write $\Pi_n = O_nO_n^*$ and note that O_n is of the form $S_n^*\tilde{O}_n$). Hence, defining $\mathbf{B}$ to be the matrix associated to the operator $T_nAT_n^* : \mathbb{R}^n \to \mathbb{R}^n$, it holds $\Pi_n A\Pi_n = S_n^*\mathbf{B}S_n$. Moreover, $A \succeq 0$ implies $\mathbf{B} = T_nAT_n^* \succeq 0$.

2. Proving $\left\{\sum_{1\leq i,j\leq n}\mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top,\ :\ \mathbf{B}\in\mathbb{R}^{n\times n},\ \mathbf{B}\succeq 0\right\}\subset\mathcal{S}_n(\mathcal{H})_+$. Let $\mathbf{B}\in\mathbb{R}^{n\times n}$, $\mathbf{B}\succeq 0$. Since $\mathbf{B}\succeq 0$, $A:=S_n^*\mathbf{B}S_n\succeq 0$. Since S_n^* has its range included in $\mathcal{H}_n$, $\Pi_n S_n^*=S_n^*$. Thus, $\Pi_n A\Pi_n=A$ and hence $A\in\mathcal{S}_n(\mathcal{H})_+$.

The second statement comes from the definition of $f_A(x)$. Indeed assume $A\in\mathcal{S}_n(\mathcal{H})_+$. By definition, $f_A(x)=\phi(x)^\top A\phi(x)$. Moreover, by the previous point, there exists $\mathbf{B}\in\mathbb{R}^{n\times n}$, $\mathbf{B}\succeq 0$ such that $A=\sum_{1\leq i,j\leq n}\mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top$. Combining these two facts yields:

$$\forall x\in\mathcal{X},\ f_A(x)=\sum_{1\leq i,j\leq n}\mathbf{B}_{i,j}\phi(x)^\top\phi(x_i)\ \phi(x_j)^\top\phi(x)=\sum_{1\leq i,j\leq n}\mathbf{B}_{i,j}k(x,x_i)\ k(x,x_j).$$

The last equality comes from the definition $k(x,\tilde{x})=\phi(x)^\top\phi(\tilde{x})$. □

Proof of Theorem 5.7. Under the assumptions of Theorem 5.7, one satisfies the assumptions of proposition 5.7. Thus, Eq. (5.5) has a solution A_* which is in $\mathcal{S}_n(\mathcal{H})_+$. Now applying Lemma 5.3, A_* can be written in the form $A_*=\sum_{i,j}\mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top$ for $\mathbf{B}\in\mathbb{R}^{n\times n}$, $\mathbf{B}\succeq 0$, and hence

$$\forall x\in\mathcal{X},\ f_{A_*}(x)=\sum_{i,j}\mathbf{B}_{i,j}k(x,x_i)k(x,x_j).$$

Uniqueness in the case where Ω is of the form Eq. (5.6) with $\lambda_2>0$ comes from the fact that the loss function is strongly convex in this case, and thus the minimizer is unique. □

5.B .4 Proof of proposition 5.3

Recall the definitions of $S_n:\mathcal{H}\to\mathbb{R}^n$ and its adjoint $S_n^*:\mathbb{R}^n\to\mathcal{H}$:

$$\forall h,\ S_n(h)=\left(h^\top\phi(x_i)\right)_{1\leq i\leq n},\ \forall\alpha\in\mathbb{R}^n,\ S_n^*\alpha=\sum_{i=1}^n\alpha_i\phi(x_i).$$

Note that the kernel matrix $\mathbf{K}=(k(x_i,x_j))_{1\leq i,j\leq n}$ can also be written as $\mathbf{K}=S_nS_n^*$.
Let r be the rank of $\mathbf{K}$ and $\mathbf{V}\in\mathbb{R}^{r\times n}$ be a matrix such that

$$\mathbf{V}^\top\mathbf{V}=\mathbf{K}.$$

Note that $\mathbf{V}$ is of rank r and hence $\mathbf{V}\mathbf{V}^\top$ is invertible, making the following definition of $O_n:\mathbb{R}^r\to\mathcal{H}$ valid:

$$O_n=S_n^*\mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-1}.$$

The following result holds :

Lemma 5.4. *$O_nO_n^*=\Pi_n$ and $O_n^*O_n=I_r$.*

Proof. Using the fact that $\mathbf{V}^\top\mathbf{V}=\mathbf{K}=S_nS_n^*$, we have

$$O_n^*O_n=(\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V}S_nS_n^*\mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-1}=(\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-1}=I_r.$$

Now let us show that $O_nO_n^*=\Pi_n$. First of all, $\tilde{\Pi}_n:=O_nO_n^*$ is self adjoint and is a projection operator since $\tilde{\Pi}_n^2=O_n(O_n^*O_n)O_n^*=O_nO_n^*=\tilde{\Pi}_n$ by the previous point. Moreover, its range is included in $\mathrm{span}(\phi(x_i))_{1\leq i\leq n}$ since $O_n=S_n^*\tilde{O}_n$ for a certain $\tilde{O}_n$ and the range of S_n^* is $\mathrm{span}(\phi(x_i))_{1\leq i\leq n}$. Finally since the rank of S_n^* is also the rank of $S_nS_n^*$ which is r, we deduce that the range of $\mathrm{span}(\phi(x_i))_{1\leq i\leq n}$ is of dimension r and hence, since $O_n^*O_n=I_r$ implies that $O_nO_n^*$ is of rank r, putting things together, $\tilde{\Pi}_n=\Pi_n$. □

Remark 19 (Constructing $\mathbf{V}$)**.** *In the case where the kernel matrix* $\mathbf{K}$ *is full rank,* $\mathbf{V} \in \mathbb{R}^{n\times n}$ *and is invertible, and* O_n *can be simply written* $S_n^*\mathbf{V}^{-1}$.
In the case where the kernel matrix $\mathbf{K}$ *is not full-rank, we build* $\mathbf{V}$ *as* $\mathbf{V} = \mathbf{\Sigma}^{1/2}\mathbf{U}^\top$, *where* $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ *is diagonal and* $\mathbf{U} \in \mathbb{R}^{n\times r}$ *is unitary and correspond to the* economy eigendecomposition *of* $\mathbf{K}$ *where* r *is the rank of* $\mathbf{K}$, *i.e.,* $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$.

Consider the following generalization of the finite dimensional model proposed in Eq. (5.8) in the case where $\mathbf{K}$ is not necessarily full rank :

$$\tilde{f}_{\mathbf{A}}(x) = \Phi(x)^\top \mathbf{A}\Phi(x), \qquad \mathbf{A} \in \mathbb{R}^{r\times r},\ \mathbf{A} \succeq 0, \tag{5.8}$$

where $\Phi : \mathcal{X} \mapsto \mathbb{R}^r$ is defined as $\Phi(x) = O_n^*\phi(x) = (\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V}v(x)$, where $v(x) = (k(x_i, x))_{1\leq i\leq n} \in \mathbb{R}^n$.

We are now ready to prove proposition 5.3.

Proof of proposition 5.3. Recall

$$\min_{\mathbf{A}\succeq 0} L(\tilde{f}_{\mathbf{A}}(x_1), \dots, \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}). \tag{5.9}$$

The fact that Eq. (5.9) has a solution, and that this solution is unique if $\lambda_2 > 0$ and L is convex can be seen as a simple consequence of Theorem 5.7 in the case where the model considered is the finite dimensional model defined in Eq. (5.8). Let us now prove the other part of the proposition.

Start by noting that with our definition of O_n, for all $\mathbf{A} \in \mathbb{R}^{r\times r},\ \mathbf{A} \succeq 0$,

$$f_{O_n\mathbf{A}O_n^*} = \tilde{f}_{\mathbf{A}}. \tag{a}$$

Moreover,

$$\left\{O_n\mathbf{A}O_n^* \ :\ \mathbf{A} \in \mathbb{R}^{r\times r},\ \mathbf{A} \succeq 0\right\} = \mathcal{S}_n(\mathcal{H})_+. \tag{b}$$

Finally, since O_n is an isometry which implies $\Omega(O_n\mathbf{A}O_n^*) = \Omega(\mathbf{A})$ and by Eq. (a), for any $\mathbf{A} \in \mathcal{S}(\mathbb{R}^n)_+$, it holds :

$$L(f_{O_n\mathbf{A}O_n^*}(x_1), ..., f_{O_n\mathbf{A}O_n^*}(x_n)) + \Omega(O_n\mathbf{A}O_n^*) = L(\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}). \tag{c}$$

Now combining Eq. (c) and Eq. (b), any solution $\mathbf{A}_*$ to Eq. (5.9) corresponds to a solution $A_* \in \arg\min_{A\in\mathcal{S}_n(\mathcal{H})_+} L(f_A(x_1), ..., f_A(x_n)) + \Omega(A)$, where $A_* = O_n\mathbf{A}_*O_n^*$. Now using Eq. (5.15) in the proof of proposition 5.7, we see that A_* is also a minimizer of Eq. (5.5) hence the result.

Note that the fact that the condition number of the problem, if it exists, is preserved because O_n is an isometry. □

5.B .5 Proof of Theorem 5.2 and algorithmic consequence.

In this section, we prove Theorem 5.2 and explain how to derive an efficient algorithm to solve it in certain cases.

Let us start by proving the following lemma.

Lemma 5.5. *Let $\lambda_1, \lambda_2 \geq 0$ and assume $\lambda_2 > 0$. Let Ω_+ be defined on $\mathcal{S}(\mathbb{R}^r)$ as follows :*

$$\Omega_+(A) = \begin{cases} \lambda_1 \|A\|_\star + \frac{\lambda_2}{2}\|A\|_F^2 & \text{if } A \succeq 0; \\ +\infty & \text{otherwise .} \end{cases}$$

Then Ω_+ is a closed convex function, and its Fenchel conjugate is given for any $B \in \mathcal{S}(\mathbb{R}^r)$ by the formula:

$$\Omega_+^*(B) = \frac{1}{2\lambda_2} \left\| [B - \lambda_1 I]_+ \right\|_F^2 .$$

Moreover, Ω_+ is differentiable at every point, and is $1/\lambda_2$ smooth. Its gradient is given by:

$$\nabla \Omega_+^*(B) = \frac{1}{\lambda_2} [B - \lambda_1 I]_+ .$$

Proof. Write

$$\Omega_+(A) = \iota_{\mathcal{S}(\mathbb{R}^r)_+} + \lambda_1 \|A\|_\star + \frac{\lambda_2}{2}\|A\|_F^2.$$

Here, ι_C stands for the characteristic function of the convex set C, i.e. $\iota_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise. Since $\|\cdot\|_F^2$ and $\|\cdot\|_\star$ are both convex, continuous, and real valued, and since $\iota_{\mathcal{S}(\mathbb{R}^r)_+}$ is closed since $\mathcal{S}(\mathbb{R}^r)_+$ is a closed non-empty convex subset of $\mathcal{S}(\mathbb{R}^r)$, this shows that Ω_+ is indeed convex and closed. Note that it is continuous on its domain $\mathcal{S}(\mathbb{R}^r)_+$. Moreover, it is strongly convex since $\lambda_2 > 0$. Fix $B \in \mathcal{S}(\mathbb{R}^r)$ and consider the problem

$$\sup_{A \in \mathcal{S}(\mathbb{R}^r)} \operatorname{Tr}(AB) - \Omega_+(A) = \sup_{A \succeq 0} \operatorname{Tr}(A(B - \lambda_1 I)) - \frac{\lambda_2}{2}\|A\|_F^2$$

Since Ω_+ is strongly convex, we know there exists a unique solution to this problem.

Note that $A_* = \arg\max \operatorname{Tr}(AB) - \Omega_+(A)$ if and only if

$$A_* = \operatorname*{arg\,min}_{A \in \mathcal{S}(\mathbb{R}^r)_+} \frac{1}{2} \left\| \left(A - \frac{1}{\lambda_2}(B - \lambda_1 I) \right) \right\|^2 .$$

That is A_* is the orthogonal projection of $\frac{B - \lambda_1 I}{\lambda_2}$ on $\mathcal{S}(\mathbb{R}^r)_+$ for the Frobenius scalar product. Hence, $A_* = \left[\frac{B - \lambda_1 I}{\lambda_2} \right]_+$.

Here, for any symetric matrix C, we denote with $[C]_+$ resp $[C]_-$ its positive resp negative part. Given an eigendecomposition $C = U \Sigma U^T$ with Σ diagonal, they are defined by $[C]_+ = U \max(0, \Sigma) U^T$ and $[C]_- = U \max(0, -\Sigma) U^T$. Hence, the Fenchel conjugate of Ω_+ is given by

$$\Omega_+^*(B) = \frac{1}{2\lambda_2} \left\| [B - \lambda_1 I]_+ \right\|_F^2 .$$

Consider $\omega_+^* : \sigma \in \mathbb{R} \mapsto \max(0, \sigma^2) \in \mathbb{R}$. ω_+^* is 1-smooth and differentiable, and $(\omega_+^*)'(\sigma) = \max(0, \sigma)$. Hence, the function

$$B \mapsto \operatorname{Tr}(\omega_+^*(B)) = \|[B]_+\|_F^2$$

is differentiable and 1-smooth, with differential given by the spectral function $(\omega_+^*)'(B) = [B]_+$. Hence, Ω_+ is differentiable and $\nabla \Omega_+^*(B) = \frac{1}{\lambda_2}[B - \lambda_1 I]_+$, and is $1/\lambda_2$ smooth. □

Theorem 5.8 (Convex dual problem)**.** *Let $L : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex closed function and L^* be the Fenchel conjugate of L (see the book by Boyd and Vandenberghe (2004) for the definition of closed and of the dual conjugate). Assume Ω is of the form Eq. (5.6). Assume there exists $\mathbf{A} \in \mathbb{R}^{r\times r}$, $\mathbf{A} \succeq \mathbf{0}$ such that L is continuous in $(\tilde{f}_{\mathbf{A}}(x_i))_{1\leq i\leq n}$.*

Then the problem in Eq. (5.9) has the following dual formulation,

$$\sup_{\alpha\in\mathbb{R}^n} -L^*(\alpha) - \tfrac{1}{2\lambda_2}\|[\mathbf{V}\operatorname{diag}(\alpha)\mathbf{V}^\top + \lambda_1\mathbf{I}]_-\|_F^2, \tag{5.10}$$

and this supremum is atteined. Let $\alpha^* \in \mathbb{R}^n$ be a solution of (5.10). Then, the solution of (5.5) is obtained via (5.7), with $\mathbf{B} \in \mathbb{R}^{r\times r}, \mathbf{B} \succeq 0$ as

$$\mathbf{B} = \mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}}\left(\frac{1}{\lambda_2}\left[\mathbf{V}\operatorname{diag}(\alpha_*)\mathbf{V}^\top + \lambda_1 I\right]_-\right)(\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}}\mathbf{V}. \tag{5.11}$$

Proof of Theorem 5.8. We apply theorem 3.3.1 by Borwein and Lewis (2010) with the following parameters (on le left, the ones in theorem 3.3.1 by Borwein and Lewis (2010) and on the right the ones by which we replace them).

$\mathbf{E}$	$\mathcal{S}(\mathbb{R}^r)$
$\mathbf{Y}$	$\mathbb{R}^n$
$A : \mathbf{E} \to \mathbf{Y}$	$R : \mathbf{A} \in \mathcal{S}(\mathbb{R}^r) \mapsto (\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n)) \in \mathbb{R}^n$
$f : \mathbf{E} \to]-\infty, +\infty]$	$\Omega_+ : \mathcal{S}(\mathbb{R}^r) \to]-\infty, +\infty]$
$g : \mathbf{Y} \to]-\infty, +\infty]$	$L : \mathbb{R}^n \to]-\infty, +\infty]$
$p = \inf_{x\in\mathbf{E}} g(Ax) + f(x)$	$p = \inf_{\mathbf{A}\in\mathcal{S}(\mathbb{R}^r)} L(\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n)) + \Omega_+(\mathbf{A})$
$d = \sup_{\phi\in\mathbf{Y}} -g^*(\phi) - f^*(-A^*\phi)$	$d = \sup_{\alpha\in\mathbb{R}^n} -L^*(\alpha) - \Omega_+^*(-R^*(\alpha))$

Indeed, for all $1 \leq i \leq n$, if Φ is defined in Eq. (5.8), $\Phi(x_i) = \mathbf{V}e_i$ and thus $\tilde{f}_{\mathbf{A}}(x_i) = \Phi(x_i)^\top\mathbf{A}\Phi(x_i) = e_i^\top(\mathbf{V}^\top\mathbf{A}\mathbf{V})e_i$. Thus, for any $\mathbf{A} \in \mathcal{S}(\mathbb{R}^r)$, $R(\mathbf{A}) := \left(\tilde{f}_{\mathbf{A}}(x_i)\right)_{1\leq i\leq n} = \operatorname{diag}(\mathbf{V}^\top\mathbf{A}\mathbf{V})$. The following properties are satisfied :

- L is lower semi-continuous, convex and bounded below hence closed (Borwein and Lewis, 2010);
- similarly, Ω_+ is a non negative closed convex function, with dual Ω_+^* given in Lemma 5.5 which is differentiable and smooth;
- $\operatorname{dom}(\Omega_+) = \mathcal{S}(\mathbb{R}^n)_+$;
- R is linear, and for any $\alpha \in \mathbb{R}^n$, it holds $R^*\alpha = \mathbf{V}\operatorname{diag}(\alpha)\mathbf{V}^\top$;
- The dual d can therefore be re-expressed as Eq. (5.10), using the expressions for Ω_+^* and R^* :

$$\sup_{\alpha\in\mathbb{R}^n} -L^*(\alpha) - \frac{1}{2\lambda_2}\left\|\left[\mathbf{V}\operatorname{diag}(\alpha)\mathbf{V}^\top + \lambda_1 I\right]_-\right\|_F^2 \tag{5.10}$$

- Assume there exists $\mathbf{A} \in \mathbb{R}^{r\times r}$, $\mathbf{A} \succeq \mathbf{0}$ such that L is continuous in $(\tilde{f}_{\mathbf{A}}(x_i))_{1\leq i\leq n}$. Then there exists a point of continuity of g such which is also in $R\operatorname{dom} f$, hence the assumption of theorem 3.3.1 by Borwein and Lewis (2010) is satisfied.

Applying theorem 3.3.1 by Borwein and Lewis (2010), the following properties hold:

- $d = p$,
- d is atteined for a certain $\alpha_* \in \mathbb{R}^n$. Indeed, there exists $\mathbf{A} \in \operatorname{dom} \Omega_+$ such that $R(\mathbf{A}) \in \operatorname{dom}(L)$. Thus , $L(R(\mathbf{A})) + \Omega_+(\mathbf{A}) < +\infty$ and hence $d < +\infty$. Moreover, since L and Ω_+ are lower bounded, this shows that d is lower bounded and hence $d > -\infty$. Hence d is finite and thus is atteined by theorem 3.3.1.

Now using Exercise 4.2.17 by Borwein and Lewis (2010) since L and Ω_+ are closed convex and since Ω_+^* is differentiable, we see that the optimal solution of the primal problem $\mathbf{A}_*$ is given by the following formula:

$$\mathbf{A}_* = \nabla \Omega_+^*(-R^* \alpha^*) = \frac{1}{\lambda_2} \left[\mathbf{V} \operatorname{diag}(\alpha_*) \mathbf{V}^\top + \lambda_1 I \right]_- .$$

Thus, for any $x \in \mathcal{X}$, using the definition of $\Phi(x)$, it holds

$$\tilde{f}_{\mathbf{A}}(x) = \Phi(x)^\top \mathbf{A}_* \Phi(x) = v(x)^\top \mathbf{V}^\top (\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}} \left(\frac{1}{\lambda_2} \left[\mathbf{V} \operatorname{diag}(\alpha_*) \mathbf{V}^\top + \lambda_1 I \right]_- \right) (\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}} \mathbf{V} v(x).$$

Thus, setting

$$\mathbf{B} = \mathbf{V}^\top (\mathbf{V}\mathbf{V}^\top)^{-1} \left(\frac{1}{\lambda_2} \left[\mathbf{V} \operatorname{diag}(\alpha_*) \mathbf{V}^\top + \lambda_1 I \right]_- \right) (\mathbf{V}\mathbf{V}^\top)^{-1} \mathbf{V},$$

it holds $\tilde{f}_{\mathbf{A}}(x) = v(x)^\top \mathbf{B} v(x)$. Since $v(x) = (k(x, x_i))_{1 \le i \le n} \in \mathbb{R}^n$, this shows the result. In particular, note that when $\mathbf{V}$ is invertible (i.e. when $\mathbf{K}$ is full rank) then the equation above is exactly Eq. (5.11), since $\mathbf{V}^\top (\mathbf{V}\mathbf{V}^\top)^{-1} = \mathbf{V}^{-1}$.

□

Proof of Theorem 5.2. It is a direct consequence of the previous theorem.

□

Note that the conditions of theorem Theorem 5.2 are satisfied in many interesting cases, such as the ones described in the following proposition.

Proposition 5.8. *Assume one of the following conditions is satisfied :*

(i) $\operatorname{dom}(L) = \mathbb{R}^n$;

(ii) $\mathbb{R}^n_{++} \subset \operatorname{dom}(L)$ *and* $k(x_i, x_i) > 0$ *for all* $1 \le i \le n$

(iii) $\mathbf{K}$ *is full rank and there exists a continuity point* α_0 *of* L *such that* $\alpha_0 \in \mathbb{R}^n_+$.

Then there exists $\mathbf{A} \in \mathcal{S}(\mathbb{R}^{\mathbf{n}})_+$ *such that* L *is continuous in* $(\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n))$.

Proof. Let us prove these points.

- if $\operatorname{dom}(L) = \mathbb{R}^n$, since L is convex, L is continuous everywhere. Taking $\mathbf{A} = \mathbf{0}$, the result holds.
- if $k(x_i, x_i) > 0$ for all $i > 0$, then taking $\mathbf{A} = I_r$, we have $(\tilde{f}_{\mathbf{A}}(x_i))_{1 \le i \le n} = (k(x_i, x_i))_{1 \le i \le n} \in \mathbb{R}^n_{++}$. Since $\mathbb{R}^n_{++} \subset \operatorname{dom}(L)$ and $\mathbb{R}^n_{++}$ is open, L is continuous on $\mathbb{R}^n_{++}$ and hence, $\mathbf{A}$ satisfies the desired property.

- Let α_0 be a continuity point of L in $\mathbb{R}^n_+$. If we assume $\mathbf{K}$ is full rank, then in particular, $\mathbf{V} \in \mathbb{R}^{n\times n}$ is of rank n and invertible. Thus, there exists $\mathbf{A} \in \mathcal{S}(\mathbb{R}^r)_+$ such that

$$\mathbf{V}^\top \mathbf{A}\mathbf{V} = \operatorname{diag}(\alpha_0) \implies (\tilde{f}_{\mathbf{A}}(x_i))_{1\leq i\leq n} = \alpha_0.$$

□

Discussion on how to solve Eq. (5.10) Proximal splitting methods can be applied to solve Eq. (5.10) such as FISTA (Beck and Teboulle, 2009), provided the proximal operator of L^* can be computed (see the work by Parikh and Boyd (2014) for the definition of the proximal operator). Indeed, Eq. (5.10) can be written as

$$\min_{\alpha\in\mathbb{R}^n} F(\alpha) = f(\alpha) + g(\alpha), \qquad f(\alpha) = \Omega^*_+(-\mathbf{V}\operatorname{diag}(\alpha)\mathbf{V}^\top),\ g(\alpha) = L^*(\alpha).$$

where Ω^*_+ has been defined in Lemma 5.5 and has been shown to be smooth and differentiable. Thus, since $\alpha \mapsto \mathbf{V}\operatorname{diag}(\alpha)\mathbf{V}^\top$ is linear, f is smooth and differentiable. Moreover, one can have access to the gradient of f by performing an eigenvalue decomposition of $\mathbf{V}\operatorname{diag}(\alpha)\mathbf{V}^\top$ whose complexity is bounded above by $\mathcal{O}(r^3)$. Thus, one can apply one of the algorithms in section 4 by Beck and Teboulle (2009) in order to compute an optimal solution to Eq. (5.10). Moreover, a bound on the performance of the algorithm is given in theorem 4.4 of this same work. Note that if L is of the form $L(\alpha) = \sum_{i=1}^n \ell_i(\alpha_i)$, it suffices to be able to compute the proximal operator of the ℓ_i to get a proximal operator for L^* (Parikh and Boyd, 2014).

5.B .6 Proof and additional discussion of Theorem 5.3

We recall the notion of universality (Micchelli, Xu, and Zhang, 2006), in particular *cc-universality* (Sriperumbudur, Fukumizu, and Lanckriet, 2011), here explicited in the context of non-negative functions. A set $\mathcal{F}$ is a *universal approximator* for non-negative functions on $\mathcal{X}$ if, for any compact subset $\mathcal{Z}$ of $\mathcal{X}$, we have that the set $\mathcal{F}|_{\mathcal{Z}}$ of restrictions on $\mathcal{Z}$, defined as $\mathcal{F}|_{\mathcal{Z}} = \{f|_{\mathcal{Z}} \mid f \in \mathcal{F}\}$, is dense in the set $C^+(\mathcal{Z})$ of non-negative continuous functions over $\mathcal{Z}$ in the maximum norm. In the following theorem we prove the cc-universality of the proposed model

Theorem 5.9. *Let $\mathcal{X}$ be a locally compact Hausdorff space, $\mathcal{H}$ a separable Hilbert space and $\phi : \mathcal{X} \to \mathcal{H}$ a cc-universal feature map. Let $\|\cdot\|_\circ$ be a norm for $\mathcal{S}(\mathcal{H})$ such that $\|\cdot\|_\star \geq \|\cdot\|_\circ$. Then $\mathcal{F}^\circ_\phi$ is a cc-universal approximator for the non-negative functions on $\mathcal{X}$.*

Proof. Proving that the proposed model is a cc-universal approximator for non-negative functions, is equivalent to require that given a compact set $\mathcal{Z} \subseteq \mathcal{X}$, a non-negative function $g : \mathcal{Z} \to \mathbb{R}_+$ and $\epsilon > 0$, there exists $f_{A_{g,\mathcal{Z},\epsilon}} \in \mathcal{F}^\circ_\phi$ such that $\|g - f_{A_{g,\mathcal{Z},\epsilon}}\|_{C(Z)} \leq \epsilon$. In particular, let $Q = 2\|g\|^{1/2}_{C(\mathcal{Z})} + \epsilon^{1/2}$, since ϕ is *cc*-universal, given $\mathcal{Z}, g, \epsilon$, there exists $w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}$ such that $\|\sqrt{g} - \phi(\cdot)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}\|_{C(Z)} \leq \frac{\epsilon}{Q}$. Define $A_{g,\mathcal{Z},\epsilon} = w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \otimes w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}$. Note that for any $x \in \mathcal{X}$,

$$f_{A_{g,\mathcal{Z},\epsilon}}(x) = \phi(x)^\top A_{g,\mathcal{Z},\epsilon}\phi(x) = \phi(x)^\top \left(w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \otimes w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \right) \phi(x) = (\phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}})^2. \quad (5.16)$$

Then, by denoting with $h(x) = \sqrt{g(x)} - \phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}$, we have

$$\begin{aligned}
\|g - f_{A_{g,\mathcal{Z},\epsilon}}\|_{C(Z)} &= \sup_{x\in\mathcal{Z}} |g(x) - (\phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}})^2| && (5.17)\\
&= \sup_{x\in\mathcal{Z}} \left| \left(\sqrt{g(x)} - \phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}\right) \left(\sqrt{g(x)} + \phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}\right)\right| && (5.18)\\
&= \sup_{x\in\mathcal{Z}} |h(x)(2\sqrt{g(x)} - h(x))| && (5.19)\\
&\leq \|h\|_{C(\mathcal{Z})}(2\|\sqrt{g}\|_{C(\mathcal{Z})} + \|h\|_{C(\mathcal{Z})}) && (5.20)\\
&\leq \frac{\epsilon}{Q}\left(2\|g\|^{1/2}_{C(\mathcal{Z})} + \frac{\epsilon}{Q}\right) \leq \epsilon. && (5.21)
\end{aligned}$$

The last step is due to the fact that $\epsilon/Q \leq \sqrt{\epsilon}$, then $2\|g\|^{1/2}_{C(\mathcal{Z})} + \frac{\epsilon}{Q} \leq Q$. □

5.B .7 Proof and additional discussion of Theorem 5.4

In Theorem 5.10, stated below, we prove that $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi$ under the very general assumption that $\mathcal{G}_\phi$ is a multiplication algebra, i.e.. if $\mathcal{G}_\phi$ is closed under pointwise product of the functions. In Theorem 5.11 we specify this result when $\mathcal{G}_\phi$ is a Sobolev space, proving that $\mathcal{E}_\phi \subsetneq \mathcal{F}^\circ_\phi$. Theorem 5.4 is a direct consequence of the latter theorem.

General result when $\mathcal{G}_\phi$ is a multiplication algebra. First we endow $\mathcal{G}_\phi$ with a Hilbertian norm. Define $\|\cdot\|_{\mathcal{G}_\phi}$ as $\|f_w\|_{\mathcal{G}_\phi} = \|w\|_{\mathcal{H}}$, for any $w \in \mathcal{H}$.

Definition 5.3. *$\mathcal{G}_\phi$ is a multiplication algebra, when there exists a constant C such that the unit function $u : \mathcal{X} \to \mathbb{R}$ that maps $x \mapsto 1$ for any $x \in \mathcal{X}$ is in $\mathcal{G}_\phi$ and*

$$\|f \cdot g\|_{\mathcal{G}_\phi} \leq C\|f\|_{\mathcal{G}_\phi}\|g\|_{\mathcal{G}_\phi}, \qquad \forall\ f, g \in \mathcal{G}_\phi, \tag{5.22}$$

where we denote by $f \cdot g$ the pointwise multiplication, i.e., $(f \cdot g)(x) = f(x)g(x)$ for all $x \in \mathcal{X}$.

Remark 20 (Renormalizing the constant)**.** *Note that when $\mathcal{G}_\phi$ is a multiplication algebra for a constant C, it is always possible to define an equivalent norm $\|\cdot\|'_{\mathcal{G}_\phi}$ as $\|\cdot\|'_{\mathcal{G}_\phi} = C\|\cdot\|_{\mathcal{G}_\phi}$ for which $\mathcal{G}_\phi$ is a multiplication algebra with constant 1.*

Theorem 5.10 (General version when $\mathcal{G}_\phi$ is an algebra)**.** *Let $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Let $\mathcal{X}$ be a compact space and ϕ be a bounded continuous map such that $\mathcal{G}_\phi$ is a multiplication algebra, then $\mathcal{E}_\phi \subseteq \mathcal{F}^\circ_\phi$.*

Proof. Let $g \in \mathcal{E}_\phi$ and take $f \in \mathcal{G}_\phi$ such that $g(x) = e^{f(x)}$ for all $x \in \mathcal{X}$. First we prove that $\mathcal{E}_\phi \subseteq \mathcal{F}^\circ_\phi$. With this goal, first we prove that $\sqrt{g} \in \mathcal{G}_\phi$ and then we construct a rank one positive operator such that $f_{A_g}(x) = g(x)$ for every $x \in \mathcal{X}$. We start noting that, given $f \in \mathcal{G}_\phi$ and $t \in \mathbb{N}$, f^t defined by $f \cdot f^{t-1}$ for $t \in \mathbb{N}$ satisfies $f^t \in \mathcal{G}_\phi$, with $\|f^t\|_{\mathcal{G}_\phi} \leq C^t\|f\|^t_{\mathcal{G}_\phi}$, by repeated application of the Eq. (5.22). Moreover note that the function $s = \sum_{t\in\mathbb{N}} \frac{1}{2^t t!} f^t$, satisfies $s \in \mathcal{G}_\phi$, indeed

$$\|s\|_{\mathcal{G}_\phi} \leq \sum_{t\in\mathbb{N}} \frac{1}{2^t t!}\|f^t\|_{\mathcal{G}_\phi} \leq \sum_{t\in\mathbb{N}} \frac{1}{2^t t!} C^t\|f\|^t_{\mathcal{G}_\phi} \leq e^{C\|f\|_{\mathcal{G}_\phi}/2}.$$

Moreover s satisfies $s(x) = \sqrt{g(x)}$ for all $x \in \mathcal{X}$, indeed for $x \in \mathcal{X}$ we have

$$s(x) = \phi(x)^\top s = \sum_{t\in\mathbb{N}} \frac{1}{2^t t!}\phi(x)^\top f^t = \sum_{t\in\mathbb{N}} \frac{1}{2^t t!} f^t(x) = e^{f(x)/2} = \sqrt{g(x)}.$$

Now let $A_g = s \otimes s$, we have that $\|A_g\|_\circ \leq \|A_g\|_\star$ by assumption, and $\|A_g\|_\star = \|s\|^2_{\mathcal{G}_\phi} < \infty$, so the function $f_{A_g} \in \mathcal{F}^\circ_\phi$ and for any $x \in \mathcal{X}$

$$f_{A_g}(x) = \phi(x)^\top A_g \phi(x) = \phi(x)^\top (s \otimes s)\phi(x) = (\phi(x)^\top s)^2 = g(x).$$

Since for any $g \in \mathcal{E}_\phi$ there exists $f_{A_g} \in \mathcal{F}^\circ_\phi$ that is equal to g on their domain of definition, we have that $\mathcal{E}_\phi \subseteq \mathcal{F}^\circ_\phi$. □

Now we are going to specialize the result above for Sobolev spaces.

Result for Sobolev spaces The result below is based on the general result in Theorem 5.10, however it is possible to do a proof based only on norm inequalities for compositions of functions in Sobolev space (see for example the work by Brezis and Mironescu (2001)). While more technical, this second approach would allow to derive also a more quantitative analysis on the norms of the functions in $\mathcal{G}_\phi$ and $\mathcal{F}^\circ_\phi$. We will leave this for a longer version of this work.

Theorem 5.11. *Let $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{X}$ compact with locally Lipschitz boundary and let $\mathcal{G}_\phi = W^m_2(\mathcal{X})$. Let $x_0 \in \mathcal{X}$. Then the following holds:*

(a) $\mathcal{E}_\phi \subsetneq \mathcal{F}^\circ_\phi$. (b) The function $f_{x_0}(x) = e^{-\|x-x_0\|^{-2}} \in C^\infty(\mathcal{X})$ satisfies $f_{x_0} \in \mathcal{F}^\circ_\phi$ and $f_{x_0} \notin \mathcal{E}_\phi$.

Proof. First we prove that $\mathcal{E}_\phi \subseteq \mathcal{F}^\circ_\phi$, via Theorem 5.10, then we. To apply this result we need first to prove that $\mathcal{G}_\phi = W^m_2(\mathcal{X})$ is a multiplication algebra when $W^m_2(\mathcal{X})$ is a RKHS as in our case.

Step 1, $m > d/2$. First note that $\mathcal{G}_\phi$ satisfies $m > d/2$ since $W^m_2(\mathcal{X})$ admits a representation in terms of a separable Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$, i.e., it is a *reproducing Kernel Hilbert space* and for the same reason $\|\cdot\|_{\mathcal{G}_\phi}$ is equivalent to $\|\cdot\|_{W^m_2(\mathcal{X})}$ (Wendland, 2004).

Step 2. $\mathcal{G}_\phi$ is a multiplication algebra. Applying Theorem 5.10. Since $\mathcal{G}_\phi = W^m_2(\mathcal{X})$ with $m > d/2$, then it is a multiplication algebra. This result is standard (e.g. see pag. 106 by Adams and Fournier (2003) for $m \in \mathbb{N}$ and $\mathcal{X} = \mathbb{R}^d$) and we report it in Lemma 5.8 in Sec. 5.C . Then we apply Theorem 5.10 obtaining $\mathcal{E}_\phi \subseteq \mathcal{F}^\circ_\phi$.

Step 3. Proving that $f_{x_0} \in \mathcal{F}^\circ_\phi$ and not in $\mathcal{E}_\phi$. By construction the function $v(x) = e^{-1/(2\|x-x_0\|^2)}$ is in $C^\infty(\mathcal{X})$ and so in $W^m_2(\mathcal{X})$ for any $m \geq 0$. Since $\mathcal{G}_\phi = W^m_2(\mathcal{X})$, then $v \in \mathcal{G}_\phi$, i.e., there exists $w \in \mathcal{H}$ such that $w^\top \phi(\cdot) = v(\cdot)$. Define $A_v = w \otimes w$, then

$$f_{A_v}(x) = \phi(x)^\top A_v \phi(x) = (w^\top \phi(x))^2 = v^2(x) = f_{x_0}(x), \quad \forall x \in \mathcal{X}.$$

Then $f_{x_0} = f_{A_v}$ on $\mathcal{X}$, i.e., $f_{x_0} \in \mathcal{F}^\circ_\phi$. To conclude note that, f_{x_0} does not belong to $\mathcal{E}_\phi$, since $x_0 \in \mathcal{X}$ and $f_{x_0}(x_0) = 0$, while for any $g \in \mathcal{E}_\phi$ we have $\inf_{x \in X} g(x) > 0$. Indeed, we have that for any $f \in \mathcal{G}_\phi$, $\|f\|_{C(\mathcal{X})} = \sup_{x \in \mathcal{X}} |f(x)| < \infty$, since $\mathcal{G}_\phi = W^m_2(\mathcal{X}) \subset C(\mathcal{X})$. Moreover, given $g \in \mathcal{G}_\phi$, and denoting by $f \in \mathcal{G}_\phi$ the function such that $g = e^f$, we have that $\inf_{x \in \mathcal{X}} g(x) \geq e^{-\|f\|_{C(\mathcal{X})}} > 0$. Finally, since $\mathcal{E}_\phi \subseteq \mathcal{F}^\circ_\phi$, but there exists $f_{x_0} \in \mathcal{F}^\circ_\phi$ and not in $\mathcal{E}_\phi$, then $\mathcal{E}_\phi \subsetneq \mathcal{F}^\circ_\phi$. □

Proof of Theorem 5.4. This result is a direct application of Theorem 5.11, since $\mathcal{X} = [-R, R]^d$, with $R \in (0, \infty)$ is a compact set with Lipschitz boundary.

5.B .8 Proof of Theorem 5.5

We recall here the Rademacher complexity and prove Theorem 5.5. This latter theorem is obtained from the following Theorem 5.12 that bounds the *empirical Rademacher complexity* introduced below. First we recall that the function class $\mathcal{F}^{\circ}_{\phi,L}$ is defined as

$$\mathcal{F}^{\circ}_{\phi,L} = \{f_A \mid A \succeq 0, \|A\|_{\circ} \leq L\},$$

for a given norm $\|\cdot\|_{\circ}$ on operators, a feature map $\phi : \mathcal{X} \to \mathcal{H}$ and $L > 0$. Now we define the empirical Rademacher complexity and the Rademacher complexity (Bartlett and Mendelson, 2002). Given $x_1, \ldots, x_n \in \mathcal{X}$, the empirical Rademacher complexity for a class $\mathcal{F}$ of functions mapping $\mathcal{X}$ to $\mathbb{R}$, is defined as

$$\widehat{R}_n(\mathcal{F}) = 2\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right|,$$

where σ_i independent Rademacher random variables, i.e., $\sigma_i = -1$ with probability $1/2$ and $+1$ with probability $1/2$ and the expectation is on $\sigma_1, \ldots, \sigma_n$. Let ρ be a probability distribution on $\mathcal{X}$ and $x_1, \ldots, x_n$ sampled independently according to ρ. The *Rademacher complexity* $R_n(\mathcal{F})$ is defined as

$$R_n(\mathcal{F}) = \mathbb{E}\widehat{R}_n(\mathcal{F}),$$

where the last expectation is on $x_1, \ldots, x_n$. In the following theorem we bound $\widehat{R}_n$.

Theorem 5.12. *Let $\|\cdot\|_{\circ} \trianglerighteq \|\cdot\|_F$. Let $x_1, \ldots, x_n \in \mathcal{X}$, $L \geq 0$.*

$$\widehat{R}_n(\mathcal{F}^{\circ}_{\phi,L}) \leq \frac{2L}{n} \sqrt{\sum_{i=1}^{n} \|\phi(x_i)\|^4}.$$

Proof. Given $f_A \in \mathcal{F}^{\circ}_{\phi,L}$, since $\|\cdot\|_{\circ}$ is stronger or equivalent to Hilbert-Schmidt norm, we have that $\|A\|_F \leq \|A\|_{\circ} \leq L$. Since A is bounded and $\phi(\cdot) \in \mathcal{H}$, by linearity of the trace we have $f_A(x) = \phi(x)^{\top} A \phi(x) = \operatorname{Tr}(A\ \phi(x) \otimes \phi(x))$ for any $x \in \mathcal{X}$. Then, by linearity of the trace

$$\hat{R}_n(\mathcal{F}^{\circ}_{\phi,L}) = 2\mathbb{E} \sup_{f \in \mathcal{F}^{\circ}_{\phi,L}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| = 2\mathbb{E} \sup_{A \succeq 0, \|A\|_{\circ} \leq L} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(x_i)^{\top} A \phi(x_i) \right| \tag{5.23}$$

$$= 2\mathbb{E} \sup_{A \succeq 0, \|A\|_{\circ} \leq L} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \operatorname{Tr}(A\ (\phi(x_i) \otimes \phi(x_i))) \right| \tag{5.24}$$

$$= 2\mathbb{E} \sup_{A \succeq 0, \|A\|_{\circ} \leq L} \left| \operatorname{Tr}\left(A\ \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(x_i) \otimes \phi(x_i) \right) \right) \right| \tag{5.25}$$

Now since $\|\cdot\|_{\circ}$ is stronger or equivalent to $\|\cdot\|_F$ this means that $\{A \in \mathcal{S}(\mathcal{H}) \mid \|A\|_{\circ} \leq L\} \subseteq$

$\{A \in \mathcal{S}(\mathcal{H}) \mid \|A\|_F \leq L\}$, then

$$2\mathbb{E} \sup_{A\succeq 0, \|A\|_\circ \leq L} \left| \operatorname{Tr}\left(A \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(x_i) \otimes \phi(x_i) \right) \right) \right| \tag{5.26}$$

$$\leq 2\mathbb{E} \sup_{A\succeq 0, \|A\|_F \leq L} \left| \operatorname{Tr}\left(A \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(x_i) \otimes \phi(x_i) \right) \right) \right| \tag{5.27}$$

$$\leq 2\mathbb{E} \sup_{A\succeq 0, \|A\|_F \leq L} \|A\|_F \left\| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(x_i) \otimes \phi(x_i) \right\|_F \tag{5.28}$$

$$\leq 2L\ \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(x_i) \otimes \phi(x_i) \right\|_F. \tag{5.29}$$

To conclude denote by ζ_i the random variable $\sigma_i \phi(x_i) \otimes \phi(x_i)$. Then

$$\mathbb{E} \left\| \tfrac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(x_i) \otimes \phi(x_i) \right\|_F^2 = \mathbb{E}\ \| \tfrac{1}{n} \sum_{i=1}^{n} \zeta_i \|_F$$

$$= \mathbb{E} \sqrt{\operatorname{Tr}\left(\left(\tfrac{1}{n} \sum_{i=1}^{n} \zeta_i \right)^* \left(\tfrac{1}{n} \sum_{i=1}^{n} \zeta_i \right) \right)} = \mathbb{E} \sqrt{\operatorname{Tr}\left(\tfrac{1}{n^2} \sum_{i,j=1}^{n} \zeta_i \zeta_j \right)}.$$

By Jensen inequality, the concavity of the square root, and the linearity of the trace

$$\mathbb{E} \sqrt{\operatorname{Tr}\left(\tfrac{1}{n^2} \sum_{i,j=1}^{n} \zeta_i \zeta_j \right)} \leq \sqrt{\mathbb{E}\ \operatorname{Tr}\left(\tfrac{1}{n^2} \sum_{i,j=1}^{n} \zeta_i \zeta_j \right)} = \sqrt{\tfrac{1}{n^2} \sum_{i,j=1}^{n} \operatorname{Tr}(\mathbb{E} \zeta_i \zeta_j)}.$$

Now note that for $i \in \{1, \ldots, n\}$, we have $\mathbb{E}_{\sigma_i} \zeta_i = 0$, moreover $\mathbb{E}\sigma_i^2 = \|\phi(x_i)\|^2 \phi(x_i) \otimes \phi(x_i)$. Finally, given $x_1, \ldots, x_n$, we have that ζ_i is independent from ζ_j, when $i \neq j$. Then when $i \neq j$ we have $\operatorname{Tr}(\mathbb{E}\zeta_i \zeta_j) = \operatorname{Tr}((\mathbb{E}_{\sigma_i} \zeta_i)(\mathbb{E}_{\sigma_j} \zeta_j)) = 0$. When $i = j$ we have $\operatorname{Tr}(\mathbb{E}\zeta_i \zeta_j) = \operatorname{Tr}(\mathbb{E}\zeta_i^2) = \|\phi(x_i)\|^4$. So

$$\frac{1}{n^2} \sum_{i,j=1}^{n} \operatorname{Tr}(\mathbb{E}\zeta_i \zeta_j) = \frac{1}{n^2} \sum_{i=1}^{n} \|\phi(x_i)\|^4.$$

From which we obtain the desired result. □

Now we are ready to bound R_n as follows

Proof Theorem 5.5. The proof is obtained by applying Theorem 5.12 and considering that $\|\phi(x)\|$ is uniformly bounded by c on $\mathcal{X}$. □

5.B .9 Proof of proposition 5.4

See Sec. 5.A for the basic technical assumptions on $\mathcal{X}$, $\mathcal{H}$ and ϕ. In particular $\mathcal{X}$ is Polish and ϕ is continuous and uniformly bounded by a constant c.

Proof of proposition 5.4. In the following we will consider integrability and measurability with respect to a measure dx on $\mathcal{X}$. In particular $p : \mathcal{X} \to \mathbb{R}$ is an integrable function on $\mathcal{X}$ with respect to the measure dx. Now define $\Psi(x) = p(x)\phi(x)\phi(x)^\top$. We have that Ψ is measurable,

since ϕ and p are measurable. Since p is integrable, p is finite almost everywhere, and hence $\Psi(x) = p(x)\phi(x)\phi(x)^\top$ is defined and trace class almost everywhere, and satisfies

$$\|\Psi(x)\|_\star = |p(x)|\ \|\phi(x)\|_{\mathcal{H}}^2 \leq |p(x)|c^2 \text{ almost everywhere.}$$

Since the space of trace class operators is separable, this shows that Ψ is Bochner integrable and thus that the operator $W_p = \int_{x\in\mathcal{X}} \phi(x)\phi(x)^\top p(x)dx$ is well defined and trace class, with trace norm bounded by $\kappa^2\|p\|_{L^1(\mathcal{X})}$. Moreover, by linearity of the integral, for any $A \in \mathcal{S}(\mathcal{H})$,

$$\mathrm{Tr}(AW_p) = \int_{\mathcal{X}} \mathrm{Tr}(A\phi(x)\phi(x)^\top)p(x)dx = \int_{\mathcal{X}} f_A(x)p(x)dx,$$

where the last equality follows from the definition of f_A and the fact that

$$\mathrm{Tr}(A\phi(x)\phi(x)^\top) = \mathrm{Tr}(\phi(x)^\top A\phi(x)) = \phi(x)^\top A\phi(x) = f_A(x).$$

□

Remark 21 (Extension to more general linear functionals.)**.** *Note that the linearity of the model in A allows to generalize very easily the construction above to any linear functional that we want to apply to the model. This is especially true when the model has a finite dimensional representation as Eq. (5.7), i.e. $f_{\mathbf{B}} = \sum_{ij=1}^n \mathbf{B}_{i,j}k(x,x_i)k(x,x_j)$ with $\mathbf{B} \succeq 0$. In this case, given a linear functional $\mathcal{L} : C(\mathcal{X}) \to \mathbb{R}$, we have*

$$\mathcal{L}(f_{\mathbf{B}}) = \sum_{i,j=1}^n \mathbf{B}_{i,j}\mathcal{L}(k(x,x_i)k(x,x_j)) = \mathrm{Tr}(\mathbf{B}\mathbf{W}_{\mathcal{L}}),$$

where $(\mathbf{W}_{\mathcal{L}})_{i,j} = \mathcal{L}(k(x,x_i)k(x,x_j))$ for $i,j = 1,\dots,n$.

5.B .10 Proof of proposition 5.5

In Sec. 5.B .10 and Sec. 5.B .11, we will use the following notations.

Let $h, p \in \mathbb{N}$ and $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2$ be separable Hilbert spaces.

- $A = (A_s)_{1\leq s\leq p} \in \mathcal{S}(\mathcal{H})^p$ will denote a family of self-adjoint operators;
- Given a feature map $\phi : \mathcal{X} \to \mathcal{H}$ and $A = (A_s)_{1\leq s\leq p} \in \mathcal{S}(\mathcal{H})^p$ we will define the function f_A as follows

$$\forall x \in \mathcal{X},\ f_A(x) = (f_{A_s}(x))_{1\leq s\leq p} = \left(\phi(x)^\top A_s\phi(x)\right)_{1\leq s\leq p} \in \mathbb{R}^p, \qquad f_A : \mathcal{X} \to \mathbb{R}^p$$

- Given a matrix $C \in \mathbb{R}^{p\times h}$ which corresponds to a list of column vectors $(c^t)_{1\leq t\leq h} \in (\mathbb{R}^p)^h$, we define

$$K^C(\mathcal{H}) := \left\{ A = (A_s)_{1\leq s\leq p} \in \mathcal{S}(\mathcal{H})^p \ :\ \sum_{s=1}^p c_s^t A_s \succeq 0,\ 1 \leq t \leq h \right\}$$

- For any $A = (A_s)_{1\leq s\leq p} \in \mathcal{S}(\mathcal{H}_1)^p$ and any bounded linear operator $L : \mathcal{H}_1 \to \mathcal{H}_2$, LAL^* will be a slight abuse of notation to denote the family $(LA_sL^*)_{1\leq s\leq p} \in \mathcal{S}(\mathcal{H}_2)^p$.

Proof of proposition 5.5. Let $p, h \in \mathbb{N}$ and let $C \in \mathbb{R}^{p\times h}$ be a matrix representing the column vectors $c^1 ... c^h$.
Let $\mathcal{Y}$ be the polyhedral cone defined by C, i.e. $\mathcal{Y} = \{y \in \mathbb{R}^p \ : \ C^\top y \geq 0\}$.
Let $\mathcal{H}$ be a separable Hilbert space and $\phi : \mathcal{X} \to \mathcal{H}$ be a fixed feature map.
With our previous notations, our goal is to prove that for any $A = (A_s)_{1\leq s\leq p} \in \mathcal{S}(\mathcal{H})^p$,

$$A \in K^C(\mathcal{H}) \implies \forall x \in \mathcal{X}, \ f_A(x) \in \mathcal{Y}.$$

Assume $A \in K^C(\mathcal{H})$ and let $x \in \mathcal{X}$. By definition, $f_A(x) = (\phi(x)^\top A_s \phi(x))_{1\leq s\leq p} \in \mathbb{R}^p$. Hence,

$$C^\top f_A(x) = \left(\sum_{s=1}^p c_s^t \phi(x)^\top A_s \phi(x)\right)_{1\leq t\leq h} = \left(\phi(x)^\top \left(\sum_{s=1}^p c_s^t A_s\right) \phi(x)\right)_{1\leq t\leq h}.$$

Since $A \in K^C(\mathcal{H})$, for all $1 \leq t \leq h$, it holds $\sum_{s=1}^p c_s^t A_s \succeq 0$. In particular, this implies $\phi(x)^\top \sum_{s=1}^p c_s^t A_s \phi(x) \geq 0$ for all $1 \leq t \leq h$. Hence

$$C^\top f_A(x) \geq 0 \implies f_A(x) \in \mathcal{Y}.$$

□

5.B .11 Proof of Theorem 5.6

Using the notations of the previous section, the goal of this section is to solve a problem of the form

$$\inf_{A\in K^C(\mathcal{H})} L(f_A(x_1), ..., f_A(x_n)) + \Omega(A), \tag{5.14}$$

for given $p, h \in \mathbb{N}$, $C \in \mathbb{R}^{p\times h}$, separable Hilbert space $\mathcal{H}$, feature map $\phi : \mathcal{X} \to \mathcal{H}$, regularizer Ω, loss function $L : \mathbb{R}^n \to \mathbb{R} \cup +\infty$ and $x_1, ..., x_n \in \mathcal{X}$.

We start by stating the form of the regularizers we will be using.

Assumption 5.2. *Let $p \in \mathbb{N}$. For any separable Hilbert space $\mathcal{H}$ and any $A = (A_s)_{1\leq s\leq p} \in \mathcal{S}(\mathcal{H})^p$, Ω is of the form*

$$\Omega(A) = \sum_{s=1}^p \Omega_s(A_s), \qquad \Omega_s(A_s) = \lambda_{s,1}\|A_s\|_\star + \frac{\lambda_{s,2}}{2}\|A_s\|_F^2,$$

where $\lambda_{s,1}, \lambda_{s,2} \geq 0$ and $\lambda_{s,1} + \lambda_{s,2} > 0$.

Lemma 5.6 (Properties of Ω)**.** *Let Ω be a regularizer such that Ω satisfies Assumption 5.2. Then Ω satisfies the following properties.*

(i) For any separable Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and any linear isometry $O : \mathcal{H}_1 \to \mathcal{H}_2$, i.e., such that $O^*O = I_{\mathcal{H}_1}$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H}_1)^p, \ \Omega(OAO^*) = \Omega(A).$$

(ii) For any separable Hilbert space $\mathcal{H}$ and any orthogonal projection $\Pi \in \mathcal{S}(\mathcal{H}_1)$, i.e. satisfying $\Pi = \Pi^*$, $\Pi^2 = \Pi$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H})^p, \ \Omega(\Pi A \Pi) \leq \Omega(A).$$

(iii) For any finite dimensional Hilbert space $\mathcal{H}_n$, taking $||A_s||_{op}$ to be the operator norm on $\mathcal{H}_n$,

$$\Omega \text{ is continuous}, \qquad \Omega(A) \underset{\sup_s ||A_s||_{op} \to +\infty}{\longrightarrow} +\infty$$

Proof. Note that since

$$\Omega(A) = \sum_{s=1}^{p} \Omega_s(A_s), \qquad \Omega_s(A_s) = \lambda_{s,1}\|A_s\|_\star + \frac{\lambda_{s,2}}{2}\|A_s\|_F^2,$$

where $\lambda_{s,1}, \lambda_{s,2} \geq 0$ and $\lambda_{s,1} + \lambda_{s,2} > 0$, it is actually sufficient to prove the following result. Let $\lambda_1, \lambda_2 \geq 0$ and assume $\lambda_1 + \lambda_2 > 0$. Let for any $A \in \mathcal{S}(\mathcal{H})$, $\Omega(A) = \lambda_1\|A\|_\star + \frac{\lambda_2}{2}\|A\|_F^2$. Then the following hold:

(i) For any separable Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and any linear isometry $O : \mathcal{H}_1 \to \mathcal{H}_2$, i.e., such that $O^*O = I_{\mathcal{H}_1}$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H}_1)^p, \ \Omega(OAO^*) = \Omega(A).$$

(ii) For any separable Hilbert space $\mathcal{H}$ and any orthogonal projection $\Pi \in \mathcal{S}(\mathcal{H}_1)$, i.e. satisfying $\Pi = \Pi^*, \ \Pi^2 = \Pi$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H})^p, \ \Omega(\Pi A \Pi) \leq \Omega(A).$$

(iii) For any finite dimensional Hilbert space $\mathcal{H}_n$,

$$\Omega \text{ is continuous}, \qquad \Omega(A) \underset{\|A\|_{op} \to +\infty}{\longrightarrow} +\infty,$$

where we denote by $\|\cdot\|_{op}$ the operatorial norm.

1. (i) has already been proven in Lemma 5.2.

2. Let us prove (ii). Let $\mathcal{H}$ be a separable Hilbert space, Π an orthogonal projection on $\mathcal{H}$ and $A \in \mathcal{S}(\mathcal{H})$.

Using the fact that $\|B\|_\star = \sup_{\|C\|_{op} \leq 1} \mathrm{Tr}(BC)$, where $\|C\|_{op}$ denotes the operator norm on $\mathcal{S}(\mathcal{H})$, we have by property of the trace

$$\|\Pi A \Pi\|_\star = \sup_{\|C\|_{op} \leq 1} \mathrm{Tr}(\Pi A \Pi C) = \sup_{\|C\|_{op} \leq 1} \mathrm{Tr}(A(\Pi C \Pi)).$$

Now since $\|\Pi C \Pi\|_{op} \leq \|C\|_{op} \leq 1$, it holds $\sup_{\|C\|_{op} \leq 1} \mathrm{Tr}(A(\Pi C \Pi)) \leq \sup_{\|C\|_{op} \leq 1} \mathrm{Tr}(AC) = \|A\|_\star$. Thus:

$$\|\Pi A \Pi\|_\star \leq \|A\|_\star.$$

Moreover, since $\Pi \preceq I$, it holds $\Pi A \Pi A \Pi \preceq \Pi A^2 \Pi$. Hence,

$$\|\Pi A \Pi\|_F^2 = \mathrm{Tr}(\Pi A \Pi \Pi A \Pi) \leq \mathrm{Tr}(\Pi A^2 \Pi)$$

Now using the fact that $\mathrm{Tr}(\Pi A^2 \Pi) = \mathrm{Tr}(A \Pi A)$, we can once again use the fact that $\Pi \preceq I$ to show that $A \Pi A \preceq A^2$ and hence $\mathrm{Tr}(A \Pi A) \leq \mathrm{Tr}(A^2)$. Putting things together, we have shown

$$\mathrm{Tr}(\Pi A \Pi \Pi A \Pi) \leq \mathrm{Tr}(A^2) \implies \|\Pi A \Pi\|_F^2 \leq \|A\|_F^2.$$

Thus, by summing the inequalities, $\Omega(\Pi A \Pi) \leq \Omega(A)$.

3. The proof of (iii) is straightforward. The continuity of Ω comes from the fact that it is a norm on any finite dimensional Hilbert space. Moreover, since $\lambda_1 > 0$ or $\lambda_2 > 0$, Ω goes to infinity. □

Remark 22. *As in the previous sections, the fact that Ω satisfies these three properties is actually sufficient to complete the proof.*

Recall that $\mathcal{H}_n$ is the finite dimensional subset of $\mathcal{H}$ spanned by the $\phi(x_i)$. Recall that Π_n is the orthogonal projection on $\mathcal{H}_n$, i.e.

$$\Pi_n \in \mathcal{S}(\mathcal{H}),\ \Pi_n^2 = \Pi_n,\ \text{range}(\Pi_n) = \mathcal{H}_n.$$

Define $K_n^C(\mathcal{H})$ to be the following subspace of $K^C(\mathcal{H})$:

$$K_n^C(\mathcal{H}) := \left\{\Pi_n A \Pi_n \ :\ A \in K^C(\mathcal{H})\right\}.$$

It is straightforward to show that $K_n^C(\mathcal{H}) \subset K^C(\mathcal{H})$ since projecting left and right preserves the linear inequalities.

Proposition 5.9. *Let L be a lower semi-continuous function which is bounded below, and assume Ω satisfies Assumption 5.2. Then Eq. (5.14) has a solution A^* which is in $K_n^C(\mathcal{H})$.*

Proof. In this proof, denote with J the function defined by

$$\forall A \in \mathcal{S}(\mathcal{H})^p,\ J(A) := L(f_A(x_1), ..., f_A(x_n)) + \Omega(A).$$

Our goal is to prove that the problem $\inf_{A \in K^C(\mathcal{H})} J(A)$ has a solution which is in $K_n^C(\mathcal{H})$, i.e. of the form $\Pi_n A \Pi_n$ for some $A \in K_n^C(\mathcal{H})$.

1. Let us start by fixing $A \in K^C(\mathcal{H})$.
First note that since Π_n is the orthogonal projection on $\text{span}(\phi(x_i))_{1 \le i \le n}$, in particular $\Pi_n \phi(x_i) = \phi(x_i)$ for all $1 \le i \le n$. Thus, for any $1 \le i \le n$,

$$f_A(x_i) = (\phi(x_i)^\top A_s \phi(x_i))_{1 \le s \le p} = (\phi(x_i)^\top \Pi_n A_s \Pi_n \phi(x_i))_{1 \le s \le p} = f_{\Pi_n A \Pi_n}(x_i).$$

Here, the first and last equalities come from the definition of f_A and $f_{\Pi_n A \Pi_n}$. Thus,

$$J(A) = L(f_{\Pi_n A \Pi_n}(x_1), ..., f_{\Pi_n A \Pi_n}(x_n)) + \Omega(A).$$

Now since Ω satisfies Assumption 5.2, by the second point of Lemma 5.6, it holds $\Omega(\Pi_n A \Pi_n) \le \Omega(A)$, hence

$$J(\Pi_n A \Pi_n) \le J(A).$$

This last inequality combined with the fact that $K_n^C(\mathcal{H}) = \left\{\Pi_n A \Pi_n \ :\ A \in K^C(\mathcal{H})\right\} \subset K^C(\mathcal{H})$ show that

$$\inf_{A \in K_n^C(\mathcal{H})} J(A) = \inf_{K^C(\mathcal{H})} J(A).$$

2. Let us now show that $\inf_{A\in K_n^C(\mathcal{H})} J(A)$ has a solution. Let us exclude the case where $J = +\infty$, in which case $A = 0$ can be taken to be a solution.

Let V_n be the injection $V_n : \mathcal{H}_n \hookrightarrow \mathcal{H}$. Note that $V_nV_n^* = \Pi_n$ and $V_n^*V_n = I_{\mathcal{H}_n}$. These simple facts easily show that

$$K_n^C(\mathcal{H}) = V_nK^C(\mathcal{H}_n)V_n^* = \left\{V_n\tilde{A}V_n^* \ : \ \tilde{A} \in K_n^C(\mathcal{H}_n)\right\}.$$

Thus, our goal is to show that $\inf_{\tilde{A}\in K_n^C(\mathcal{H}_n)} J(V_nAV_n^*)$ has a solution.

By the first point of Lemma 5.6, since $V_n^*V_n = I_{\mathcal{H}_n}$, it holds

$$\forall \tilde{A} \in \mathcal{S}(\mathcal{H}_n),\ \Omega(V_n\tilde{A}V_n^*) = \Omega(\tilde{A}) \implies J(V_n\tilde{A}V_n^*) = L(f_{V_n\tilde{A}V_n^*}(x_1), ..., f_{V_n\tilde{A}V_n^*}(x_n)) + \Omega(\tilde{A}).$$

Let $\tilde{A}_0 \in K^C(\mathcal{H}_n)$ be a point such that $J_0 := J(V_n\tilde{A}_0V_n^*) < \infty$. Let c_0 be a lower bound for L. By the third point of Lemma 5.6, there exists a radius R_0 such that for all $\tilde{A} \in \mathcal{S}(\mathcal{H}_n)$,

$$\|\tilde{A}\|_F > R_0 \implies \Omega(\tilde{A}) > J_0 - c_0.$$

Since c_0 is a lower bound for L, this implies

$$\inf_{\tilde{A}\in K^C(\mathcal{H}_n)} J(V_n\tilde{A}V_n^*) = \inf_{\tilde{A}\in K^C(\mathcal{H}_n),\ \|\tilde{A}\|_F\le R_0} J(V_n\tilde{A}V_n^*).$$

Now since L is lower semi-continuous, Ω is continuous by the last point of Lemma 5.6, and $\tilde{A} \mapsto (f_{V_n\tilde{A}V_n^*}(x_i))_{1\le i\le n}$ is linear hence continuous, the mapping $A \mapsto J(V_n\tilde{A}V_n^*)$ is lower semi-continuous. Hence, it reaches its minimum on any non empty compact set. Since $\mathcal{H}_n$ is finite dimensional, the set $\left\{\tilde{A} \in K^C(\mathcal{H}_n) \ : \ \|\tilde{A}\|_F \le R_0\right\}$ is compact (closed and bounded) and non empty (it contains $\tilde{A}_0$), and hence there exists $\tilde{A}_* \in K_n^C(\mathcal{H})$ such that $J(V_n\tilde{A}_*V_n^*) = \inf_{\tilde{A}\in K^C(\mathcal{H}_n),\ \|\tilde{A}\|_F\le R_0} J(V_n\tilde{A}V_n^*)$. Going back up the previous equalities, this shows that $A_* := V_n\tilde{A}_*V_n^* \in K_n^C(\mathcal{H})$ and $J(A_*) = \inf_{A\succeq 0} J(A)$.

□

Lemma 5.7. *The set $K_n^C(\mathcal{H})$ can be represented in the following way*

$$K_n^C(\mathcal{H}) = \left\{(S_n^*\mathbf{B}_sS_n)_{1\le s\le p} \in \mathcal{S}(\mathcal{H})^p \ : \ \mathbf{B} = (\mathbf{B}_s)_{1\le s\le p} \in K^C(\mathbb{R}^n)\right\}$$

In particular, for any $A \in K_n^C(\mathcal{H})$, there exists p symmetric matrices $\mathbf{B} = (\mathbf{B}_s)_{1\le s\le p} \in K^C(\mathbb{R}^n)$ such that

$$\forall x \in \mathcal{X},\ f_A(x) = \left(\sum_{1\le i,j\le n} [\mathbf{B}_s]_{i,j}k(x_i, x)k(x_j, x)\right)_{1\le s\le p}.$$

Proof. The proof is exactly analoguous to the proof of Lemma 5.3. □

We will prove the following Theorem 5.13 which statement is that of Theorem 5.6 with more precise assumptions.

Theorem 5.13. *Let L be lower semi-continuous and bounded below, and Ω satisfying Assumption 5.2. Then Eq. (5.5) has a solution of the form*

$$f_*(x) = \left(\sum_{i,j=1}^n [\mathbf{B}_s]_{i,j} k(x,x_i)k(x,x_j)\right)_{1\leq s\leq p}, \qquad \text{for some family } \mathbf{B} = (\mathbf{B}_s)_{1\leq s\leq p} \in K^C(\mathbb{R}^n).$$

Moreover, if L is convex, this solution is unique.

Proof of Theorem 5.13. The proof is completely analoguous to that of Theorem 5.7, combining Lemma 5.7 and proposition 5.9. □

5.C Additional proofs

Lemma 5.8. *Let $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a compact set with Lipschitz boundary. Let $m > d/2$. Then $W_2^m(\mathcal{X})$ is a multiplication algebra (see definition 5.3).*

Proof. When $m \in \mathbb{N}$ and $m > d/2$, then $W_2^m(\mathbb{R}^d)$ is a multiplication algebra (Adams and Fournier, 2003). When $m \notin \mathbb{N}$, by Eq. 2.69 pag. 138 by Triebel (2006) we have that $F_{2,2}^m(\mathbb{R}^d)$ is a multiplication algebra when $m > d/2$, where $F_{2,2}^m$ is the Triebel-Lizorkin space of smoothness m and order $2, 2$ and corresponds to $W_2^m(\mathbb{R}^d)$, i.e., $F_{2,2}^m(\mathbb{R}^d) = W_2^m(\mathbb{R}^d)$ (Triebel, 2006).

So far we have that $m > d/2$ implies that $W_2^m(\mathbb{R}^d)$ is a multiplication algebra, now we extend this result to $W_2^m(\mathcal{X})$. Note that since $\mathcal{X}$ is compact and with Lipschitz boundary, for any $f \in W_2^m(\mathcal{X})$ there exists an *extension* $\tilde{f} \in W_2^m(\mathbb{R}^d)$ such that $\tilde{f}|_{\mathcal{X}} = f$ and $\|\tilde{f}\|_{W_2^m(\mathbb{R}^d)} \leq C_1 \|f\|_{W_2^m(\mathcal{X})}$ with C_1 depending only on $m, d, \mathcal{X}$ (see Thm. 5.24 pag. 154 for $m \in \mathbb{N}$ and 7.69 when $m \notin \mathbb{N}$ pag. 256 by Adams and Fournier (2003)). Then, since for any $f : \mathbb{R}^d \to \mathbb{R}$, by construction we have $\|f|_X\|_{W^m(\mathcal{X})} \leq \|f\|_{W^m(\mathbb{R}^d)}$ (Adams and Fournier, 2003). Then, for any $f, g \in W_2^m(\mathcal{X})$, denoting by $\tilde{f}, \tilde{g}$ the extensions of f, g, we have

$$\|f \cdot g\|_{W_2^m(\mathcal{X})} = \|\tilde{f}|_{\mathcal{X}} \cdot \tilde{g}|_{\mathcal{X}}\|_{W_2^m(\mathcal{X})} \leq \|\tilde{f} \cdot \tilde{g}\|_{W_2^m(\mathbb{R}^d)} \tag{5.30}$$

$$\leq C\|\tilde{f}\|_{W_2^m(\mathbb{R}^d)}\|\tilde{g}\|_{W_2^m(\mathbb{R}^d)} \leq CC_1^2\|f\|_{W_2^m(\mathcal{X})}\|g\|_{W_2^m(\mathcal{X})}. \tag{5.31}$$

To conclude $u : \mathcal{X} \to \mathbb{R}$ that maps $x \mapsto 1$ has bounded norm corresponding to $\|u\|^2_{W_2^m(\mathcal{X})} = \int_{\mathcal{X}} dx$. So $W_2^m(\mathcal{X})$ when $m > d/2$ and $\mathcal{X}$ is compact with Lipschitz boundary is a multiplication algebra. □

5.D Additional details on the other models

Recall that the goal is to solve a problem of the form Eq. (5.1), i.e.

$$\min_{f\in\mathcal{F}} L(f(x_1), ..., f(x_n)) + \Omega(f).$$

In this section, $\phi : \mathcal{X} \to \mathcal{H}$ will always denote a feature map, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive semi definite kernel on $\mathcal{X}$ ($k(x, x') = \phi(x)^\top \phi(x')$ if k is the positive semi-definite kernel associated to ϕ). Given a kernel k, $\mathbf{K} \in \mathbb{R}^{n\times n}$ will always denote the positive semi-definite kernel matrix with coefficients $\mathbf{K}_{i,j} = k(x_i, x_j)$, $1 \leq i, j \leq n$.

Generalized linear models (GLM). Consider generalized linear models of the form, $f_w(x) = \psi(w^\top \phi(x))$. Assume the regularizer is of the form $\Omega(f_w) = \frac{\lambda}{2}\|w\|^2$. Using the representer theorem

by Cheney and Light (2009), any solution to Eq. (5.1) is of the form $w = \sum_{i=1}^{n} \alpha_i \phi(x_i)$ and thus Eq. (5.1) becomes the following finite dimensional problem in α:

$$\min_{\alpha \in \mathbb{R}^n} L(\psi(\mathbf{K}\alpha)) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{5.32}$$

In the case where one wishes to learn a density function with respect to a basis measure ν, a common choice of model is functions of the form

$$p_\alpha(x) = \frac{\exp(g(x))}{\int_{\tilde{x} \in \mathcal{X}} \exp(g(\tilde{x})) d\nu(\tilde{x})}, \qquad g(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

where k is a positive semi-definite kernel on $\mathcal{X}$. The prototypical problem one solves to find the best p_α is

$$\min_{\alpha \in \mathbb{R}^n} L(p_\alpha(x_1), ..., p_\alpha(x_n)) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{5.33}$$

In the specific case where the loss function is the negative log likelihood $L(z_1, ..., z_n) = \frac{1}{n} \sum_{i=1}^{n} -\log(z_i)$, it can be shown that Eq. (5.33) is convex in α.

In practice, we solve Eq. (5.32) by applying standard gradient descent with restarts, as the problem is non convex.

To solve Eq. (5.33), since the problem is convex, the algorithm is guaranteed to converge. However, since we can only estimate the quantity $\int_{\tilde{x} \in \mathcal{X}} \exp(g(\tilde{x})) d\nu(\tilde{x})$; we do so by taking a measure ν from which we can sample. However, this becomes intractable as the dimension grows, as the experiments on density estimation will put into light.

Non-negative coefficients models (NCM). Recall the definition of an NCM. It represent non-negative functions as $f_\alpha(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$, with $\alpha_1, \dots \alpha_n \geq 0$, given a kernel $k(x, x') \geq 0$ for any $x, x' \in \mathcal{X}$. In this case, the prototypical problem is of the form :

$$\min_{\alpha \geq 0} L(\mathbf{K}\alpha) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{5.34}$$

If we are performing density estimation with respect to the measure ν, one wishes to impose $\int_{\mathcal{X}} f_\alpha(x) d\nu(x) = 1$, which can be seen as an affine constraint over α, since

$$\int_{\mathcal{X}} f_\alpha(x) d\nu(x) = \mathbf{u}^\top \alpha, \qquad \mathbf{u} = \left(\int_{\mathcal{X}} k(x, x_i) d\nu(x) \right)_{1 \leq i \leq n} \in \mathbb{R}^n.$$

In this case, the prototypical problem will be of the form

$$\min_{\substack{\alpha \geq 0 \\ \mathbf{u}^\top \alpha = 1}} L(\mathbf{K}\alpha) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{5.35}$$

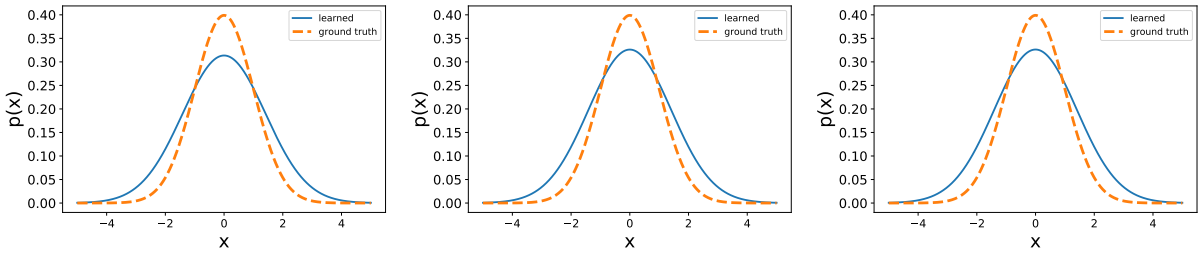

Figure 5.2: Best approximation of g using NCM with *(left)* $n = 100$ *(center)* $n = 1000$ *(right)* $n = 10000$ points.

If L is a convex smooth function, both problems Eq. (5.34) and Eq. (5.35) can be solved using projected gradient descent, since the projections on the set $\alpha \geq 0$ and the simplex $\left\{\alpha \in \mathbb{R}^n \ : \ \alpha \geq 0, \ \mathbf{u}^\top \alpha = 1\right\}$ can be computed in closed form.

In the main paper, we mention that NCM models do not satisfy **P2** i.e. that they cannot approximate any function arbitrarily well. We implement Example 5.2 in the following way. Let $g(x) = e^{-\|x\|^2/2}$. Take $k(x, x') = e^{-\|x-x'\|^2}$, n points $(x_1, ..., x_n)$ taken uniformly in the interval $[-5, 5]$. To find the function f_α which best approximates g, we perform least squares regression, i.e. solve the prototypical problem Eq. (5.34) with the square loss function

$$L(y) = \frac{1}{2n} \sum_{i=1}^{n} |y_i - g(x_i)|^2.$$

We perform cross validation to select the value of λ for each value of n. In Fig. 5.2, we show the obtained function f_α for $n = 100, 1000, 10000$. This clearly illustrates that with this model, we cannot approximate g in a good way, no matter how many points n we have.

Partially non-negative linear models (PNM). Consider partially non negative models of the form $f_w(x) = w^\top \phi(x)$, with $w \in \{w \in \mathcal{H} \mid w^\top \phi(x_1) \geq 0, \ldots, w^\top \phi(x_n) \geq 0\}$ (that is we impose $f_w(x_i) \geq 0$). Take Ω to be of the form $\frac{\lambda}{2}\|w\|^2$ in Eq. (5.1). Using the representer theorem by Cheney and Light (2009), we can show that there is a solution of this problem of the form $f_\alpha = \sum_{i=1}^n \alpha_i k(x, x_i)$, leading to the following optimization problem in α to recover the optimal solution:

$$\min_{\mathbf{K}\alpha \geq 0} L(\mathbf{K}\alpha) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha \tag{5.36}$$

If we want to impose that the resulting f_α sums to one for a given measure ν on $\mathcal{X}$, we proceed as in Eq. (5.35) and solve

$$\min_{\substack{\mathbf{K}\alpha \geq 0 \\ \mathbf{u}^\top \alpha = 1}} L(\mathbf{K}\alpha) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{5.37}$$

However, there is no guarantee that the resulting f_α will be a density, as will be made clear in the next section on density estimation.

In the experiments, we solve Eq. (5.36) and Eq. (5.37) in the following way. We first compute a cholesky factor of $\mathbf{K}$: $\mathbf{K} = \mathbf{V}^\top \mathbf{V}$. Changing variables by setting $\mathbf{V}\beta = \alpha$, the objective functions become strongly convex in β. We then compute the dual of these problems and apply a proximal algorithm like FISTA, since the proximal operator of L is always known in our experiments.

5.E Additional details on the experiments

In this section, we provide additional details on the experiments. The code will be available online. Recall that we consider four different models for functions with non-negative outputs : GLM, PNM, NCM and our model.

Kernels. All the models we consider depend on certain positive semi definite kernels k. In all the experiments, we have taken the kernels to be Gaussian kernels with width σ:

$$\forall x, x' \in \mathbb{R}^d,\ k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right).$$

Regularizers. For GLM, PNM and NCM, the regularizer for the underlying linear models are always of the form $\frac{\lambda}{2}\|w\|^2$ where w is the parameter of the linear model, which translates to $\frac{\lambda}{2}\alpha^\top \mathbf{K}\alpha$ where the α are the coefficients of the finite dimensional representation. For our model, we always take the regularizer to be of the form $\lambda\left(\|A\|_\star + 0.05\|A\|_F^2\right)$.

Parameter selection. In all experiments except for the one on density estimation in the main paper (in which we fix $\sigma = 1$ and select λ), we select the parameters σ of the kernels involved as well as the parameters λ for the regularizers using K fold cross validation with $K = 7$. This means that once the data set has been generated, we randomly divide it into two sets : the training set containing 70% of the data and the test set containing 30% of the data. We then train our model for the given σ, λ and report the performance on the test set. We repeat this operation $K = 7$ times and consider the mean performance on the test set to be a good indicator of the performance of our model for a given set of parameters. We then select the best parameters by doing a grid search. The code for this cross-validation will be available online.

Formulations and algorithms. The formulations of our three problems : density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression, have been expressed in the main paper in a generic way involving functions with unconstrained outputs, and functions with outputs constrained to be non negative and sometimes summing to one. We always model functions with unconstrained outputs with a linear model with gaussian kernel, and model the functions with constrained outputs with the four models for non-negative functions we consider: ours, PNM, GLM and NCM.

In practice, we implement the methods PNM, GLM and NCM as explained in Sec. 5.D . In particular, we use FISTA for PNM, and our model, dualizing the equality constraints for density estimation. This relies on the fact that the proximal operators of the log likelihood, the objective function for heteroscedastic regression as well as the pinball loss can be computed in closed form, and that the regularization is smooth in the right coordinates.

Details on the experiments of the main text. Here, we add a few precisions on the toy distributions we have used to sample data and the number of sampled used when not specified in the main text.

- For heteroscedastic regression, the data was generated as the toy data in section 5 by Le, Smola, and Canu (2005), with $n = 80$ points.

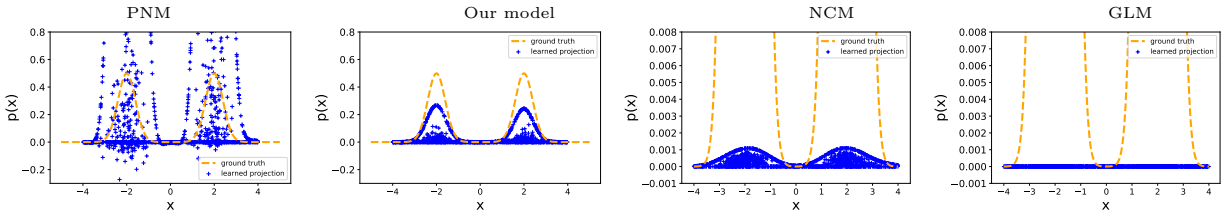

Figure 5.3: Representation of the densities learned by the different models.

- For quantile regression, the data points (x_i, y_i) were generated according to the following distribution for (X, Y) : $X \sim \frac{1}{2}U(0, 1/3) + \frac{1}{2}U(2/3, 1)$ and $Y|x \sim \mathcal{N}(0, \sigma(x))$ where

$$\sigma(x) = \begin{cases} -x + 1/3 & \text{for } 0 \leq x \leq 1/3 \\ x - 2/3 & \text{for } 2/3 \leq x \leq 1 \\ 0 & \text{otherwise .} \end{cases}$$

  Here, U stands for the uniform distribution. Moreover, in order to perform the experiments in the main paper, we have used 500 sample points.

Density estimation in dimension 10 with $n = 1000$. In this paragraph, we consider the following experiment. Let $d = 10$, $X \in \mathbb{R}^d$ be a random variable distributed as a mixture of Gaussians :

$$X \sim \frac{1}{2}\mathcal{N}(-2e_1, 1/\sqrt{2\pi}I_d) + \frac{1}{2}\mathcal{N}(2e_1, 1/\sqrt{2\pi}I_d)$$

where e_1 is the first vector of the canonical basis of $\mathbb{R}^d$.

Let $n = 1000$ and let $(x_1, ..., x_n)$ be n iid samples of X. We perform the four different methods, cross validating both the regularization parameter λ and the kernel parameter σ at each time. We learn the density in the form

$$p(x) = f(x)\nu(x), \qquad \nu \text{ is the density associated with } \mathcal{N}(0, 5I_d).$$

We then use our models for densities to compute the best f in its class using the negative log-likelihood as a loss function. It is crucial that we can sample from ν in order to approximate the integral in the case of GLMs.

In order to visualize the results of the different algorithms in Fig. 5.3, we compute the learnt distribution p, and then sample randomly $n_0 = 500$ points from a uniform distribution on the box centered at 0 and of width 5 in order to explore regions where the density is close to zero, n_0 points sampled from the true distribution of the data, in order to explore points where the density is representative, and n_0 points on the line $[-4, 4] \times \{0\}^{d-1}$ where the density is at its highest. We then project onto the first coordinate, i.e. given a point $x = (x_i)_{1 \leq i \leq d}$ and the associated predicted density $p(x)$, we plot the point $(x_1, p(x))$. Note that for readability, we have used the same scale for our model and the PNM, and a smaller scale for the two others since the learnt density is much flatter.

Let us now analyse the results in Fig. 5.3. Note that in terms of performance, i.e. log likelihood on the test set, the first two models (PNM and our model) are quite close and are better than the two others.

- **PNM**. As in $d = 1$ we see that for $d = 10$ the problems of non-negativity for PNM are exacerbated, making it not suitable to learn a probability distribution. Indeed there are low density regions where the optimization problem pushes the model to be negative. Since by constraint we have $\int f d\nu = 1$, the volume of the negative regions is used to push up the function in the regions with high density. So $\int |f| d\nu \gg 1$, while it should be $\int |f| d\nu = 1$. This is confirmed by the behavior of the cross validation.
- **Our model** Our model seems to perform reasonably well.
- **NCM**. This problem is particularly difficult for NCM. Indeed, as the width of the kernel decreases, the model is unable to learn since it overfits in the direction e_1 and it would require way more points than $n = 1000$. However, as soon as the width of the kernel is good for e_1, the learnt distribution becomes too heavy tailed in the direction orthogonal to e_1.
- **GLM**. It is interesting to note that GLM completely fails, because the measure ν which we take as a reference measure has a support which has only double variance compared to p, but in 10 dimensions it corresponds to a support with way larger volume compared to the one of the target distribution. In particular, the estimation of the integral, which was possible in $d = 1$ with 10000 i.i.d. points from ν, in 10 dimensions becomes almost impossible (it would require way more sampling points). Note that we sample the points from ν to simulate the real-world situation where p is a measure from which it is difficult to sample from, while ν is an simple measure to sample from which contains the support of p. Further experiments show that if one takes the target distribution to sample, one obtains a good model, which reassures us in the fact that this is not a coding error but a real phenomenon.

5.F Relationship to the work by Bagnell and Farahmand (2015)

As mentioned in the main paper, the model in Eq. (5.4) has already been considered by Bagnell and Farahmand (2015) with a similar goal as ours. This paper is a workshop publication that has only be lightly peer-reviewed and contains fundamental flaws. In particular, they provide an incorrect characterization of the solution of Eq. (5.5), that limits the representation power of the model to the one of non-negative coefficients models, that, as we have seen in Sec. 5.2 .1 and in Example 5.2, has poor approximation properties and cannot be universal. This severe limitation affects also the optimization framework (which also only relies on general-purpose toolboxes such as CVX (`http://cvxr.com/cvx/`), which are not scalable to large n).

Indeed, in their main result, the representer theorem incorrectly characterizes A^* the solution of Eq. (5.5) as

$$A^* \in R_n \cap \mathcal{S}(\mathcal{H})_+, \qquad R_n = \left\{ \sum_{i=1}^n \alpha_i \phi(x_i) \otimes \phi(x_i) \mid \alpha \in \mathbb{R}^n \right\},$$

and $\mathcal{S}(\mathcal{H})_+ = \{A \in \mathcal{S}(\mathcal{H}) \mid A \succeq 0\}$. Note, however that $R_n \subseteq \mathcal{S}(\mathcal{H}_n) \subset \mathcal{S}(\mathcal{H})$ by construction, where $\mathcal{H}_n = \operatorname{span}\{\phi(x_1), \dots, \phi(x_n)\}$. So their characterization corresponds to

$$A^* \in \left\{ A = \sum_{i=1}^n \alpha_i \phi(x_i) \otimes \phi(x_i) \mid \alpha \in \mathbb{R}^n, A \succeq 0 \right\}.$$

Now, for simplicity, consider the interesting case where ϕ is universal and $x_1, \dots, x_n$ are distinct points. Then $(\phi(x_i))_{i=1}^n$ forms a basis for $\mathcal{H}_n$ and the only $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ that guarantee $A \succeq 0$

are $\alpha_1 \geq 0, \ldots, \alpha_n \geq 0$, i.e.,

$$R_n = \left\{ A = \sum_{i=1}^{n} \alpha_i \phi(x_i) \otimes \phi(x_i) \mid \alpha_1 \geq 0, \ldots, \alpha_n \geq 0 \right\}.$$

Note that this class of operators leads only to *non-negative coefficients models*. Indeed, let $A \in R_n$ and denote by $k(x, x')$ the function $k(x, x') = (\phi(x)^\top \phi(x'))^2$, then

$$f_A(x) = \phi(x)^\top A \phi(x) = \sum_{i=1}^{n} \alpha_i (\phi(x)^\top \phi(x_i))^2 = \sum_{i=1}^{n} \alpha_i k(x, x_i), \quad \forall\, x \in \mathcal{X}.$$

Since k is a kernel (it is an integer power of $\phi(x)^\top \phi(x')$ that is a kernel (Scholkopf and Smola, 2001)) and $\alpha_1 \geq 0, \ldots, \alpha_n \geq 0$, then f_A belongs to the non-negative coefficients models.

Instead, we know by our Theorem 5.1 that $A^* \in \mathcal{S}(\mathcal{H}_n)_+$ and more explicitly, by Theorem 5.2 that A^*, the solution of Eq. (5.5) is characterized by the non-positive part operator of a symmetric matrix $[\cdot]_+$. By Theorem 5.3 we already know that our model is universal while NCM is not and thus the characterization by (Bagnell and Farahmand, 2015) cannot be universal.

Chapter 6

Sampling from arbitrary functions via PSD models

This chapter is a verbatim of the work :

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Sampling from arbitrary functions via psd models. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2823–2861. PMLR, 28–30 Mar 2022a. URL https://proceedings.mlr.press/v151/marteau-ferey22a.html.

Contents

6.1 Introduction **265**
6.2 Backround on Positive Semi-Definite (PSD) models **266**
6.3 A sampling algorithm for PSD models **268**
6.4 Sampling from arbitrary distributions using PSD models **273**
6.5 Experiments **279**
6.6 Extensions, future work **280**
6.A Definitions and notations **284**
6.B Properties of the Gaussian RKHS **289**
6.C Properties of Gaussian PSD models **294**
6.D The sampling algorithm **297**
6.E A general method of approximation and sampling **302**
6.F Approximation and sampling using a rank one PSD model **309**
6.G Additional experimental details **314**

6.1 Introduction

In many fields such as biochemistry, statistical mechanics and machine learning, effectively sampling arbitrary numbers of independent and identically distributed (i.i.d.) samples from probability distributions is a key task (Gelman et al., 2004; Liu, 2008; Lelièvre et al., 2010).

Basic sampling methods include rejection sampling and gridding, and rely on simple properties of the density. However, they are suitable only in small dimensions, except for very structured cases.

Moreover, they are hard to adapt to probabilities which are known up to their renormalization constant, which is often the case when dealing with exponential models that are common in applications (Robert and Casella, 2013).

More involved methods have been developed to address these dimensionality and renormalization issues, in the class of so-called Markov chain Monte Carlo (MCMC) methods. However, they are complex to set up: in particular, independence between samples is not directly guaranteed, convergence can be slow and hard to measure non-asymptotically (Lelièvre et al., 2010; Robert and Casella, 2013).

In this work, we address the problem in a different way, by incorporating a modeling step. Instead of sampling directly from the target density, we first model this density using a positive semi-definite (PSD) model (Marteau-Ferey et al., 2020; Rudi and Ciliberto, 2021), and then sample from this PSD model.

PSD models have been introduced by Marteau-Ferey, Bach, and Rudi (2020) and their relevance for modeling probability distributions has been further established by Rudi and Ciliberto (2021), showing that i) they are stable under key operations for probabilistic inference, such as marginalization, integration (also called "sum-rule"), and product, which can be done efficiently in practice, and ii) they concisely approximate a large class of probability distributions. We present these models in Sec. 6.2 . Building on this work, we show that these models are also relevant in the context of sampling, making the following main contributions.

(1) In Sec. 6.3 , we derive an algorithm that is easy to implement and which can generate an arbitrary number of i.i.d. samples from a given PSD model, with any given precision. This answers one of the open questions outlined by Rudi and Ciliberto (2021) and shows that one can indeed efficiently sample from a PSD model.

(2) In Sec. 6.4 we show that we can sample an arbitrary number of i.i.d. samples from a target probability distribution that is regular enough, with any given precision. The algorithm consists in (a) approximating the un-normalized density p via a PSD model, using evaluations of p, and (b) extracting i.i.d. samples from the PSD model. We show that for sufficiently regular densities the resulting PSD model is concise and avoids the curse of dimensionality: to achieve error ε, the PSD model requires a number of parameters and a number of evaluations of p that are in the order $\varepsilon^{-2-d/\beta}$, where d is the dimension of the space and β is the order of differentiability of the density. For regular probabilities, i.e., when $\beta \geq d$, the rate does not depend exponentially on d and is bounded by $O(\varepsilon^{-3})$ (the constant term instead may depend exponentially on d).

In Sec. 6.5 , we also present numerical simulations which demonstrate the quality of both our sampling technique and approximation results.

6.2 Backround on Positive Semi-Definite (PSD) models

Denote by $\mathbb{R}^d_{++}$ the vectors of $\mathbb{R}^d$ with positive components and $\mathbb{S}^m_+$ the set of positive semi-definite m by m matrices. Following Marteau-Ferey, Bach, and Rudi (2020); Rudi and Ciliberto (2021), a Gaussian PSD model is parametrized by a triplet $(A, X, \eta) \in \mathbb{S}^m_+ \times \mathbb{R}^{m\times d} \times \mathbb{R}^d_{++}$, and is defined for any $x \in \mathbb{R}^d$ as

$$f(x;\ A, X, \eta) = \sum_{i,j=1}^{m} A_{ij} k_\eta(x, x_i) k_\eta(x, x_j), \tag{6.1}$$

where, with $\mathrm{diag}(\eta)$ being the diagonal matrix with diagonal η, $k_\eta(x, x') = e^{-(x-x')^\top \mathrm{diag}(\eta)(x-x')}$ is the Gaussian kernel of parameter η , $X \in \mathbb{R}^{n\times d}$ is the matrix whose rows corresponds to the

centers $x_1, \dots, x_n$ of the Gaussian PSD model, and A is a matrix of coefficients which is positive semi-definite, to guarantee the non-negativity of f.

Note that when $A = aa^\top,\ a \in \mathbb{R}^m$, is a rank-1 operator, a Gaussian PSD model is simply the square of a linear model $f(x;\ A, X, \eta) = g(x;\ a, X, \eta)^2$ of the form,

$$g(x;\ a, X, \eta) = \sum_{i=1}^{m} a_i k_\eta(x, x_i), \tag{6.2}$$

for any $x \in \mathbb{R}^d$. This particular case of PSD model will appear when approximating an arbitrary probability density p in Sec. 6.4 .2.

6.2 .1 Main properties of PSD models

As explained in the introduction, PSD models show properties that make them particularly well suited to model non-negative functions and probability distributions. Such properties are analyzed by Marteau-Ferey, Bach, and Rudi (2020) and Rudi and Ciliberto (2021), here we recall the ones that are important for our purpose.

Non-negativity. Since A is positive semidefinite, then the PSD model $f(x;\ A, X, \eta)$ satisfies $f(x;\ A, X, \eta) \geq 0$ for all $x \in \mathbb{R}^d$.

Preservation of convex functionals. Using the PSD model to represent non-negative functions in a problem of the form $\min_{f \geq 0} L(f)$, where L is a convex functional, leads to a convex problem $\min_{A \in \mathbb{S}_+(\mathbb{R}^m)} L(f(\cdot;\ A, X, \eta))$. Indeed, the constraint $A \in \mathbb{S}_+(\mathbb{R}^m)$ is convex, the PSD model $f(\cdot;\ A, X, \eta)$ is linear in the parameter matrix A and a composition of a convex function L with a linear function is convex. This allows, e.g., to perform empirical risk minimization for the square and logarithmic losses.

Conciseness of the representation. under mild conditions, recalled in Assumption 6.1, a PSD model can approximate a probability density that is β-times differentiable with error ε, using a number of centers $m = O(\varepsilon^{-d/\beta})$ (which is minimax optimal). Rudi and Ciliberto (2021) provide also an algorithm to learn the PSD model given i.i.d. samples from the probability. However, we cannot use this result in our context since we do not assume to have samples from our density.

Integration over hyper-rectangles in closed form. As integration of PSD models will play a key role in the algorithm developed for sampling in Sec. 6.3 , both for theoretical an computational reasons, we develop this integration aspect in greater detail.

A hyper-rectangle $Q \subset \mathbb{R}^d$ can be parametrized with its corners $a, b \in \mathbb{R}^d,\ a \leq b$, by writing $Q = \prod_{k=1}^{d} [a_k, b_k[$; a corresponds to the "bottom left" corner and b to the "top right" one.

For $X \in \mathbb{R}^{m \times d}$ and $\eta \in \mathbb{R}^d_{++}$, we denote with $K_{X,\eta} \in \mathbb{R}^{m \times m}$ the *kernel matrix* such that $[K_{X,\eta}]_{ij} = k_\eta(x_i, x_j)$. The integral of a PSD model in Eq. (6.1) over a hyper-rectangle can be expressed with simple matrices, leveraging the fact that for any pair (x_i, x_j), it holds

$k_\eta(x, x_i)k_\eta(x, x_j) = k_{\eta/2}(x_i, x_j)k_{2\eta}(x, (x_i + x_j)/2)$. Then we have

$$\begin{aligned} I(Q; A, X, \eta) &:= \int_Q f(x; A, X, \eta)\ dx \\ &= \sum_{i,j=1}^m A_{ij} k_{\frac{\eta}{2}}(x_i, x_j) \int_Q k_{2\eta}(x, \tfrac{x_i+x_j}{2})\ dx \\ &= \sum_{i,j=1}^m A_{ij}[K_{X,\eta/2}]_{ij}[G_{X,2\eta,Q}]_{ij}, \end{aligned} \tag{6.3}$$

where $[G_{X,\eta,Q}]_{ij} = \int_{Q_{ij}} k_\eta(x, 0)\ dx$, and $Q_{ij} = Q - (x_i + x_j)/2$. These integrals can be computed by $2d$ calls to the erf function, as, for any $i, j \in \{1, ..., m\}$:

$$[G_{X,\eta,Q}]_{ij} = c_\eta \prod_{k=1}^d [\mathrm{erf}(\sqrt{\eta_k}\mathcal{B}_{ijk}) - \mathrm{erf}(\sqrt{\eta_k}\mathcal{A}_{ijk})], \tag{6.4}$$

where $c_\eta = (\pi/4)^{d/2} \det \mathrm{diag}(\eta)^{-1/2}$, $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{d\times m\times m}$, $\mathcal{A}$ is the tensor of bottom left corners and $\mathcal{B}$ is the tensor of top right corners, defined formally from the means tensor $\overline{X}_{ijk} = \frac{1}{2}(X_{ik} + X_{jk})$ as

$$\mathcal{A}_{ijk} = a_k - \overline{X}_{ijk}, \quad \mathcal{B}_{ijk} = b_k - \overline{X}_{ijk}. \tag{6.5}$$

This shows that, for any hyper-rectangle Q, we can compute $G_{X,\eta,Q}$ with exactly $2dm^2$ calls to the erf function and dm^2 arithmetic operations (so there is no dependence on the dimension of the hyper-rectangle).

6.3 A sampling algorithm for PSD models

In this section, we fix a Gaussian PSD model on $\mathbb{R}^d$ parametrized by $(A, X, \eta) \in \mathbb{S}^m_+ \times \mathbb{R}^{m\times d} \times \mathbb{R}^d_{++}$ for a given $m \in \mathbb{N}$. To simplify notations, we will omit the parameters of the PSD model using $f(x)$ as a shorthand for $f(x; A, X, \eta)$ and $I(Q)$ as a shorthand of $I(Q) = I(Q; A, X, \eta)$.

Given a bounded hyper-rectangle Q (see Sec. 6.3 .1), denote by p_Q the function

$$p_Q(x) = f(x)\mathbf{1}_Q(x)/I(Q), \tag{6.6}$$

where $\mathbf{1}_Q(x) = 1$ when $x \in Q$ and 0 otherwise. In Sec. 6.3 .2, we explain that even in the case of an infinite hyper-rectangle (e.g., $Q = \mathbb{R}^d$), we can easily find a finite hyper-rectangle $\tilde{Q}$ on which the whole mass of f is essentially concentrated, and thus approximately sample in this case as well. We end this section with a discussion on the main elements needed to sample, and which could allow to generalize this approach to PSD models with different kernels.

6.3 .1 A sampling algorithm on a finite hyper-rectangle

Given the function f, the algorithm will take three inputs (Q, N, ρ): the hyper-rectangle Q (with sides parallel to the axes) from which we would like to sample, the number of i.i.d. samples N which we would like to obtain, and a parameter ρ which defines the quality of the approximation of p_Q from which the algorithm generates samples. The effect of ρ on the precision of the algorithm is formally established in Theorem 6.2.

We start with the case $N = 1$. Starting from Q, we cut Q in half in its longest direction forming two sub-rectangles Q_1, Q_2. If X_Q were a random variable following the law of p_Q, then $X_Q \in Q_i$

with probability $p_i = I(Q_i)/I(Q)$, and $X_Q|\{X_Q \in Q_i\}$ follows the law of p_{Q_i}. Therefore, when looking for a sample from p_Q, we randomly choose with probability p_i one of the two smaller sub-rectangles Q_i in which to look for the sample and then call the algorithm recursively to get a sample from p_{Q_i}. Of course, we need a stopping criterion: when the maximal side of Q has length smaller than ρ then we stop and we return a point sampled uniformly at random in Q. The complete algorithm is presented in algorithm 2 and is explained below.

Details for algorithm 2. In line 1, we define the recursive function SampleRec which will generate samples recursively. The main algorithm Sample in line 15 simply calls the function SampleRec and randomly reshuffles the samples in order to guarantee independence (see RandomPerm line 17). In line 4, the function MaxLen applied to Q returns the maximum of the lengths of the sides of Q; the condition can therefore be translated as "if all sides of Q are smaller than ρ". If it is the case, in line 5, we return N i.i.d. samples from the uniform distribution on Q using SampleUniform. If it is not, in line line 7 we cut the hyper-rectangle Q in half along its largest side with minimal index (i.e., along side $k = \min \arg\max (b_i - a_i)$), yielding two sub hyper-rectangles Q_1, Q_2. This is the purpose of the function SplitLargestSide. In line 8, we compute the probability q that a given sample from p_Q belongs to Q_1 using the fact that we can integrate the PSD model exactly. Since we have to generate N samples, we will select k of them from Q_1 and $N - k$ from Q_2 where k is a sample from a binomial law of paramter q: this is the purpose of the function SampleBinomial and line 9. We then call the algorithm recursively to generate the k samples from Q_1 using p_{Q_1} and the $N - k$ samples from Q_2 from p_{Q_2} (lines 10 and 11).

Algorithm 2 Approximately sampling from p_Q

```
1:  function SampleRec(Q, N, ρ)
2:      if N = 0 then
3:          return EmptyList
4:      else if MaxLen(Q) ≤ ρ then
5:          return SampleUniform(Q, N)
6:      else
7:          Q_1, Q_2 = SplitLargestSide(Q)
8:          q = I(Q_1)/I(Q)
9:          k = SampleBinomial(N, q)
10:         L_1 = SampleRec(Q_1, k, ρ)
11:         L_2 = SampleRec(Q_2, N − k, ρ)
12:         return Concatenate(L_1, L_2)
13:     end if
14: end function
15: function Sample(Q, N, ρ)
16:     L = SampleRec(Q, n, ρ)
17:     return RandomPerm(L)
18: end function
```

Guarantees of the algorithm. Given (Q, N, ρ), algorithm 2 does not sample N i.i.d. samples from the exact distribution p_Q but rather from an approximation $p_{Q,\rho}$ of p_Q, controlled by the parameter ρ. More formally, let $\mathcal{D}_{Q,\rho}$ be the set of dyadic sub-rectangles of Q with largest possible size smaller than ρ (see Appendix 6.D for a formal definition). Our algorithm will

effectively sample from a piece-wise constant approximation of p on the elements of $\mathcal{D}_{Q,\rho}$:

$$p_{Q,\rho} = \frac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \frac{I(Q_\rho)}{|Q_\rho|} \mathbf{1}_{Q_\rho}, \tag{6.7}$$

where $\mathbf{1}_{Q_\rho}$ is the indicator function of Q_ρ. The guarantees of the algorithm are established in the following theorem, proved formally in Appendix 6.D .2.

Theorem 6.1. *Given (Q, N, ρ) where Q is a bounded hyper-rectangle of $\mathbb{R}^d$, $\rho > 0$ and $N \in \mathbb{N}$, the function* SAMPLE *in algorithm 2 returns N i.i.d. samples from the distribution $p_{Q,\rho}$ defined in Eq. (6.7). Moreover, the number of integral computations of the form $I(\widetilde{Q})$ performed during the algorithm is bounded by $N \log_2(|Q|) + Nd \log_2 \frac{2}{\rho} + 1$, and the number of* erf *computations is $O(N\ m^2\ d\ (\log_2(2|Q|) + d \log_2(2/\rho)))$, where m is the dimension of the PSD model.*

Note that the theorem gives us that the complexity is essentially $O(Nm^2d^2 \log(1/\rho))$. This quadratic dependence in the dimension d is verified in practice and the slicing procedure does not yield any time or computational difficulties. Note however that in our two step procedure detailed in the next section, the number m will a priori depend on the dimension, but this is confined to the learning phase; once the m centers are set, the complexity is quadratic. Moreover, note that we verify the claim that computing integrals is the computational bottleneck in practice in Sec. 6.D .4.

Approximation error of the algorithm. Since by Theorem 6.1, the algorithm does not generate samples exactly from p_Q but rather from the piecewise constant approximation $p_{Q,\rho}$ defined in Eq. (6.7), it is necessary to quantify the distance between p_Q and its approximation $p_{Q,\rho}$. We do so in Theorem 6.2 for three different distances.

The weakest distance will be the Wasserstein-1 distance (also called earth mover's distance) (Santambrogio, 2015). It quantifies the discrepancies in the allocation of mass between two distributions, and is defined as

$$\mathbb{W}_1(p_1, p_2) = \sup_{\mathrm{Lip}(f) \leq 1} \left| \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x))dx \right|, \tag{6.8}$$

where $\mathrm{Lip}(f)$ is the Lipschitz constant of f for the Euclidean norm. It is structurally the most adapted to the approximation $p_{Q,\rho}$ since on each hyper-rectangle of $\mathcal{D}_{Q,\rho}$, $p_{Q,\rho}$ has the same mass as p_Q but distributes it uniformly. Hence, the discrepancy in mass allocation will be confined to small hyper-rectangles whose sides are of size at most ρ.

We will also use two stronger distances: the total variation (TV) distance $d_{TV}(p_1, p_2) = \|p_1 - p_2\|_{L^1(\mathcal{X})}$, and the Hellinger distance $H(p_1, p_2) = \|\sqrt{p_1} - \sqrt{p_2}\|_{L^2(\mathcal{X})}$, which is particularly relevant for exponential models (Lucien Le Cam, 1990), and, in our paper, when using rank-1 PSD models (see Sec. 6.4 .2). These distances will naturally appear in Sec. 6.4 to quantify the discrepancy between a given probability density and its approximation as a Gaussian PSD model. For more details on these distances, see Appendix 6.A .2. Theorem 6.2 provides bounds on these distances between the target density $p_Q = f\mathbf{1}_Q/I(Q)$ and $p_{Q,\rho}$ as a function of ρ, and some Lipschitz constant (where $\mathrm{Lip}_\infty(g)$ denotes the Lipschitz constant of g for the norm $\|x\|_\infty = \sup |x_i|$). A more general theorem is proved in Appendix 6.D .3 as Theorem 6.7.

Theorem 6.2 (Variation bounds)**.** *Let Q be a hyper-rectangle, $\rho > 0$, $p_Q = f\mathbf{1}_Q/I(Q)$ and $p_{Q,\rho}$*

defined in Eq. (6.7). *It holds:*

$$H(p_Q, p_{Q,\rho}) \leq \sqrt{\tfrac{|Q|}{I(Q)}}\,\mathrm{Lip}_\infty(\sqrt{f})\;\rho \tag{6.9}$$

$$d_{TV}(p_Q, p_{Q,\rho}) \leq \tfrac{|Q|}{I(Q)}\,\mathrm{Lip}_\infty(f)\rho \tag{6.10}$$

$$\mathbb{W}_1(p_Q, p_{Q,\rho}) \leq \sqrt{d}\rho. \tag{6.11}$$

Combining the result of Theorems 6.1 and 6.2, we have that, given a PSD model on m centers, an hyper-rectangle of interest Q and an error ρ, algorithm 2 provides N i.i.d.samples whose distribution is distant $\sqrt{d}\rho$ in terms of $\mathbb{W}_1$ from the density represented by the PSD model over the hyper-rectangle. In particular, algorithm 2 computes the N i.i.d. samples with a cost of $O(N\ m^2\ d\ (\log_2(2|Q|) + d\log_2(2/\rho)))$.

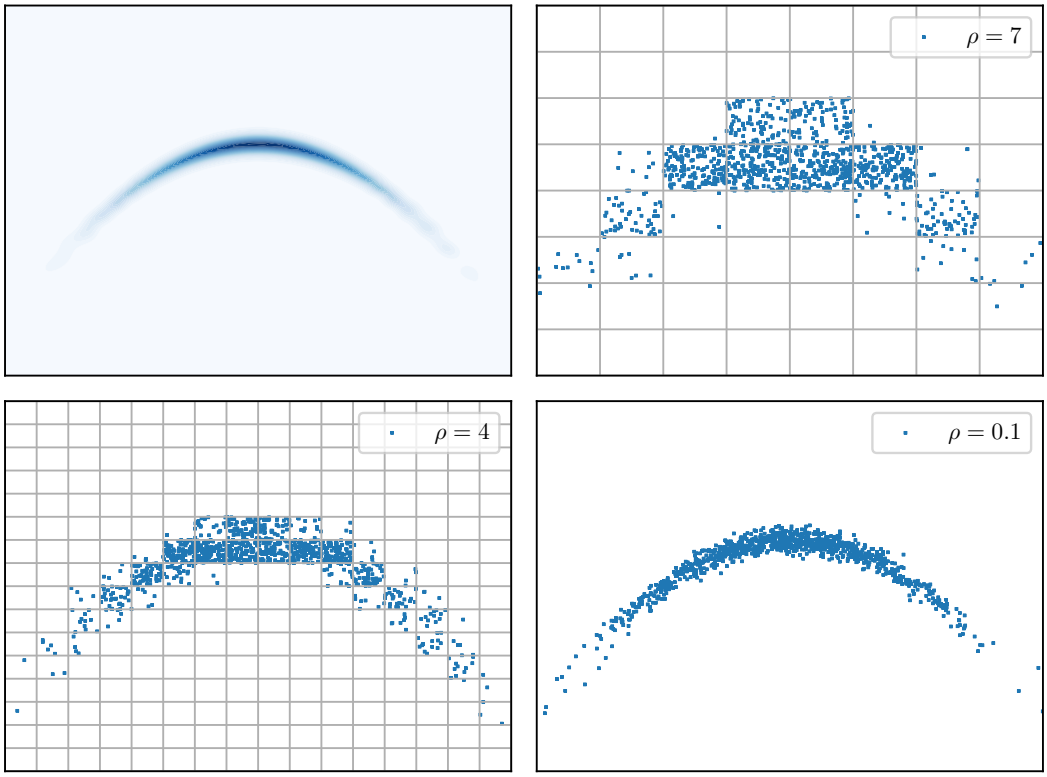

Figure 6.1: Samples obtained from algorithm 2 using different values for ρ

Selection of ρ. In Fig. 6.1, we observe the effect of ρ on the quality of sampling, when sampling from a PSD model whose distribution is illustrated by the heat map defined on the top left figure. We highlight the fact that decreasing ρ corresponds to refining the dyadic decomposition of the hyper-rectangle and hence sampling more precisely. In practice, one can therefore choose ρ manually (for instance $\rho = 10^{-4}, 10^{-6}$) and have an upper bound on the distance between $p_{Q,\rho}$ and p_Q from Theorem 6.2. If one wishes to select ρ in a more principled way to bound the total variation or Hellinger distance, this can also be done using only accessible quantities. If f is a PSD model with parameters (A, X, η) for $\eta = \tau\mathbf{1}_d$, and K is a shorthand for $K_{X,\eta}$, the Lipschitz constants can be bounded using only τ, K and A (or a s.t. $A = aa^\top$ in the case of a rank one

PSD model). More precisely, it holds

$$\mathrm{Lip}_\infty(f) \leq \sqrt{8\tau}d\|K^{1/2}AK^{1/2}\| =: \widetilde{\mathrm{Lip}}(A) \tag{6.12}$$

$$\mathrm{Lip}_\infty(\sqrt{f}) \leq \sqrt{2\tau}d\|K^{1/2}a\| =: \widetilde{\mathrm{Lip}}(a), \tag{6.13}$$

where for Eq. (6.13), $A = aa^\top$ is assumed to be a rank-1 operator[1]. These quantities only depend on a, A, K and can be computed explicitly. Combining these bounds with Eqs. (6.9) and (6.10), ρ can be selected in an adaptive way in algorithm 2.

Remark 23 (Adaptive selection of ρ)**.** *Let $\varepsilon > 0$. Let f be a PSD model with matrix of coefficients A. Define*

$$\rho_\varepsilon^{TV} = \frac{I(Q)\varepsilon}{|Q|\widetilde{\mathrm{Lip}}(A)}, \qquad \rho_\varepsilon^{H} = \frac{\sqrt{I(Q)}\varepsilon}{\sqrt{|Q|\widetilde{\mathrm{Lip}}(a)}}, \tag{6.14}$$

where ρ_ε^H is defined if $A = aa^\top$ is a rank one matrix. If $\rho = \rho_\varepsilon^{TV}$ (resp. $\rho = \rho_\varepsilon^H$), then algorithm 2 applied to (Q, N, ρ) returns N i.i.d. samples from a distribution $p_{Q,\varepsilon}$ which satisfies $d_{TV}(p_Q, p_{Q,\varepsilon}) \leq \varepsilon$ (resp. $H(p_Q, p_{Q,\varepsilon}) \leq \varepsilon$).

6.3 .2 Discussion

Sampling from the distribution on $\mathbb{R}^d$. It is possible to approximately sample from an infinite hyper-rectangle. To do so, one has to find a large enough hyper-rectangle Q such that almost all the mass is contained on Q and then apply the previous algorithm to this hyper-rectangle. One can, for instance, use algorithm 3.

Algorithm 3 Finding an approximate support Q

```
function FindApproximateSupport(f(·; A, X, η), δ)
    Q = ∏_{1≤k≤d} [min_{1≤i≤n} X_ik, max_{1≤i≤n} X_ik]
    I = I(ℝ^d)
    while I(Q)/I ≤ 1 − ε do
        Q = DoubleSize(Q)
    end while
end function
```

Note that one can also concentrate f a priori using only its parameters (X, A, η), using Eq. (6.56) of Lemma 6.4 in Appendix 6.C .1. One can use this bound to bound the number of steps in algorithm 3.

Generality of the algorithm. algorithm 2 only relies on the fact that one can compute integrals on hyper-cubes of the model f. If we were to replace the Gaussian kernel k_η by a kernel k, and therefore have a PSD model of the form $\sum_{ij} A_{ij}k(x, x_i)k(x, x_j)$ with another positive definite kernel and $A \in \mathbb{S}_+^m$, then one would be able to run the algorithm as soon as computations of the form $\int_Q k(x, x_i)k(x, x_j)dx$ were tractable. This would extend this framework to more general PSD models, described by Marteau-Ferey, Bach, and Rudi (2020).

[1] See Lemma 6.5 in Appendix 6.C .1 for a proof of a the bound on $\mathrm{Lip}_\infty(f)$ when f is a PSD model and Lemma 6.2 in Appendix 6.B .1 for a proof of of a bound on $\mathrm{Lip}_\infty(\sqrt{f})$ in the case where f is a rank one PSD model.

6.4 Sampling from arbitrary distributions using PSD models

The previous section provides an algorithm to approximately sample from a distribution in the form of a PSD model. In this section, we show how to leverage that fact to be able to generate N approximate i.i.d. samples from a very general class of probability distributions on a hyper-rectangle $\mathcal{X} \subset \mathbb{R}^d$. The strategy is simple : a) approximate the target distribution p with a PSD model $\widehat{p}$, and b) approximately sample from the PSD model $\widehat{p}$ using the algorithm presented in Sec. 6.3 . The main challenge is to quantify the distance between the target distribution p and its approximation $\widehat{p}$ as a PSD model.

Approaching a distribution by a PSD model by accessing the distribution through samples has been done in Sec. 3. by Rudi and Ciliberto (2021). Instead, in this work, we access the distribution through function evaluations, as our goal is to be able to generate samples. However, a similar algorithm can be implemented to learn a PSD model from function evaluations. Moreover, it can be analysed under the same conditions (see Assumption 6.1 and Sec. 6.4 .1). This algorithm is based on the solving of a semi-definite program to find the matrix A to form a good approximation $f(x;\ A, \widetilde{X}_m, \eta)$ of the density p. In Sec. 6.4 .2, we instead learn a rank-one PSD model, solving a least-squares problem (and not a semi-definite program) using tools by Rudi, Camoriano, and Rosasco (2015); Rudi, Carratino, and Rosasco (2017); Meanti, Carratino, Rosasco, and Rudi (2020). This algorithm, faster than the one based on the solving of a semi-definite program, requires a stronger assumption to be analysed, and is naturally adapted to densities of the form $p(x) \propto e^{-V(x)}$.

Main hyper-parameters. The two methods presented in this section (see Sec. 6.4 .1 and Sec. 6.4 .2) will have hyper-parameters $n, m, \tau, \lambda, \rho$.

The parameters n and m are integer; moreover, we will take two sequences of i.i.d. samples uniformly from $\mathcal{X}$: $x_1, ..., x_n$ represented by $X \in \mathbb{R}^{n \times d}$ and $\widetilde{x}_1, ..., \widetilde{x}_m$ represented by $\widetilde{X}_m \in \mathbb{R}^{m \times d}$. We will use an isotropic $\eta = \tau \mathbf{1}_d$ in the Gaussian linear and PSD models for a strictly positive τ. To simplify notation, take $K_{mm} := K_{\widetilde{X}_m, \eta}$ and $K_{nm} := K_{X, \widetilde{X}_m, \eta}$. The parameter λ will always be a strictly positive real number.

The parameters m and τ will define the PSD model: m will control the number of points, also called *Nyström centers*, which we use to represent our PSD model (as n and m increase, the quality of the approximation increases); and τ will control the width of the Gaussian kernel. The parameter n and λ control the learning phase of the algorithm, i.e., the approximation of p by a PSD model. n is the number of points at which we evaluate our probability density to estimate it; λ will control the strength of the regularization. Finally, ρ will control the scale at which we apply algorithm 2.

6.4 .1 A general method

In this section, we present a method to approximately sample from the density by a) approximating it by a PSD model solving a semi-definite program (SDP) and b) use algorithm 2 to sample from that PSD model. More precisely, we assume that p is known up to a constant, i.e., that we have a function f_p which is proportional to p which we can evaluate.

Step a): approximation of p. To fit a PSD model to p, we use an method similar to the one presented in Section 3 of Rudi and Ciliberto (2021), and construct a Gaussian PSD model

$\widehat{f} = f(\bullet;\ \widehat{A}, \widetilde{X}_m, \eta)$, where $\widehat{A} \in \mathbb{S}_+^m$ is the solution to the empirical semi-definite problem

$$\widehat{A} = \underset{A \in \mathbb{S}_+^m}{\arg\min} \int_{\mathcal{X}} f(x;\ A)^2 dx$$
$$-2\sum_{i=1}^{n} f_p(x_i) f(x_i;\ A) + \lambda \|K_{mm}^{1/2} A K_{mm}^{1/2}\|_F^2, \tag{6.15}$$

where $f(x;\ A) := f(x;\ A, \widetilde{X}_m, \eta)$. This problem is a quadratic problem in A and can be solved in polynomial time in m using semi-definite programming. We then define $\widehat{Z} = \int_{\mathcal{X}} \widehat{f}(x)\ dx$ which can be computed in closed form as the integral over a hyper-cube of a PSD model, and $\widehat{p} = \widehat{f}/\widehat{Z}$, which is our approximation of p.

Problem Eq. (6.15) can be seen as a variation of empirical risk minimization for the square loss, with an additional regularization term $\lambda \|K_{mm}^{1/2} A K_{mm}^{1/2}\|_F$ which is the equivalent of the classical kernel regularization term in the setting of PSD models. Indeed, the function of A being minimized is a proxy of $\|f(\cdot;\ A) - f_p(\cdot)\|_{L^2(\mathcal{X})}^2 = \|f(\cdot;\ A)\|_{L^2(\mathcal{X})}^2 + \int_{\mathcal{X}} f_p(x) f(x;\ a)\ dx + C$. In Eq. (6.15), $\int_{\mathcal{X}} f_p(x) f(x;\ a)\ dx$ is approximated by its empirical version, using uniform samples $X = (x_1, ..., x_n)$ (plus the regularization term). The first term $\|f(\cdot;\ A)\|_{L^2(\mathcal{X})}^2$ is kept as such as it is a quadratic function of A which can be explicitly computed, using the same techniques as those to compute integrals of PSD models, and described by Rudi and Ciliberto (2021). Note that here, $X, \widetilde{X}_m, \tau, \lambda$ are hyper-parameters; n and m will be taken as large as possible with a given computational budget, and λ and τ can be selected by validation on a newly generated test data set (since we assume we can generate samples from $\mathcal{X}$).

Step b): sampling from the approximation $\widehat{p}$. We apply algorithm 2 to $\widehat{p}$ with a parameter ρ and on the hyper-rectangle $\mathcal{X}$. We denote with $p_{\texttt{sample}}$ the density $\widehat{p}_{\mathcal{X},\rho}$ given by Eq. (6.7), from which algorithm 2 effectively samples N i.i.d. samples by Theorem 6.1. This two step strategy is detailed in algorithm 4. SolveSDP simply solves Eq. (6.15).

Algorithm 4 Approximately sampling from p using a SDP

Input $p, \mathcal{X}, N$
Hyper-parameters (approximation) n, m, τ, λ
Hyper-parameters (sampling) ρ
Output N approximate samples from $p|_{\mathcal{X}}$

function ApproximateSamples($p, \mathcal{X}, N, n, m, \tau, \lambda, \rho$)
  X_n = UniformSamples($n, \mathcal{X}$)
  X_m =UniformSamples($m, \mathcal{X}$)
  A = SolveSDP($p, X_n, X_m, \tau, \lambda$)
  $\widehat{p}(\cdot) = f(\cdot \mid A, X_m, \tau)$
  X_N = Sample($\mathcal{X}, N, \rho$) from $\widehat{p}$
  return X_N
end function

Theoretical analysis. Recall that p is the target density, proportional to f_p and that $\widehat{p}$ is the approximation of p obtained by solving Eq. (6.15) and $p_{\texttt{sample}}$ is the distribution from which we effectively sample when applying algorithm 2 to $\widehat{p}$. In proposition 6.1 and Theorem 6.3, we show that under certain regularity assumptions on p, given $\varepsilon > 0$, we can find hyper-parameters

n, m, τ, λ and ρ such that $d_{TV}(p, p_{\texttt{sample}}) \leq C\varepsilon$, i.e. that algorithm 4 generates N i.i.d. samples from a distribution $C\varepsilon$ close to p.

For simplicity, we will assume $\mathcal{X} = (-1, 1)^d$, as is done by Rudi and Ciliberto (2021). In principle, we could approximate p on any bounded domain $\mathcal{X}$ from which we can sample uniformly, and still obtain analogous results. In that case, we would apply algorithm 2 on a hyper-rectangle containing the domain, and reject a sample outside of it. Our main assumption on p will be that p can be written as a sum of squares of functions belonging to the space $\widetilde{W}^\beta(\mathcal{X}) = W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ which is the space of bounded functions whose derivatives of order less or equal to β are square integrable, and which can be equipped with the norm $\|\cdot\|_{\widetilde{W}^\beta(\mathcal{X})} = \|\cdot\|_{W_2^\beta(\mathcal{X})} + \|\cdot\|_{L^\infty(\mathcal{X})}$ (see Appendix 6.A .1 for more precise definitions). The key quantities here are the dimension d and the regularity of the density β. This summarized in the following assumption.

Assumption 6.1 (Sum of squares distribution)**.** *There exists $J \in \mathbb{N}$ and functions $q_1, ..., q_J$ belonging to $\widetilde{W}^\beta(\mathcal{X})$ such that $p = \sum_{j=1}^J q_j^2$. Moreover, we have access to p only through function evaluations of the form $f_p(x)$ where $f_p \geq 0$ is given, is proportional to p, and where the proportionality constant is unknown. We define $\|p\|_{\texttt{sos},\mathcal{X},\beta} = \inf \sum_{j=1}^J \|q_j\|^2_{\widetilde{W}^\beta(\mathcal{X})}$ where the infimum is taken over all such decompositions of p.*

The approximation properties of $\widehat{p}$ w.r.t. p are bounded in total variation distance in the following proposition, proved as proposition 6.10 in Appendix 6.E .

Proposition 6.1 (Performance of $\widehat{p}$)**.** *There exist constants $\varepsilon_0 > 0$ depending only on d, β, and $\|p\|_{\texttt{sos},\mathcal{X},\beta}$ and C_1, C_1', C_2', C_3' depending only on d, β such that the following holds. Let $\delta \in (0, 1]$ and $\varepsilon \leq \varepsilon_0$, and assume n and m satisfy*

$$m \geq C_1' \varepsilon^{-d/\beta} \log^d\left(\frac{C_2'}{\varepsilon}\right) \log\left(\frac{C_3'}{\varepsilon\delta}\right), \tag{6.16}$$

$$n \geq \varepsilon^{-2-d/\beta} \log^d\left(\frac{1}{\varepsilon}\right) \log\left(\frac{2}{\delta}\right). \tag{6.17}$$

Let $\lambda = \varepsilon^{2+2d/\beta}$ and $\tau = \varepsilon^{-2/\beta}$. With probability at least $1 - 2\delta$, it holds

$$d_{TV}(\widehat{p}, p) \leq C_1 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon. \tag{6.18}$$

The key takeaway from this proposition is that the number of samples n, m needed to perform the first step of the algorithm (approximation) is polynomial in the quantities $O(\varepsilon^{-1}), O(\varepsilon^{-d/\beta})$, thus leveraging the regularity β of p. When this is the case, we can find λ, τ such that the distance $d(p, \widehat{p})$ is of order ε. We provide a choice for ρ for the second step of the algorithm (sampling), in order to guarantee a bound for the total variation distance between the sampling distribution and the original distribution in the following theorem. It is proved as Theorem 6.8 in Appendix 6.E . In particular, it bounds the total complexity of the algorithm in terms of erf computations, as a function of N and the desired error ε.

Theorem 6.3 (Performance of $p_{\texttt{sample}}$)**.** *Under the assumptions and notations of proposition 6.1, there exists a constant C_2 depending only on d, β, such that the following holds. If ρ is set either as $\varepsilon^{1+(d+1)/\beta}$ or adaptively as ρ_ε^{TV}, then with probability at least $1 - 2\delta$,*

$$d_{TV}(p, p_{\texttt{sample}}) \leq C_2 \ \|p\|_{\texttt{sos},\mathcal{X},\beta} \ \varepsilon. \tag{6.19}$$

Moreover, the adaptive ρ_ε^{TV} is lower bounded by $\varepsilon^{1+(d+1)/\beta}/(C_3 \ \|p\|_{\texttt{sos},\mathcal{X},\beta})$. In both cases, this guarantees that the complexity in terms of erf *computations is of order $O(Nm^2 \log(1/\rho))$, which in terms of ε yields $O\left(N \ \varepsilon^{-2d/\beta} \log^{2d+1}\left(\frac{1}{\varepsilon}\right) \log^2\left(\frac{1}{\delta\varepsilon}\right)\right)$, where the O notations is taken with constants depending on $d, \beta, \|p\|_{\texttt{sos},\mathcal{X},\beta}$.*

6.4 .2 Efficient method with a rank one model

In this section, we present a method to approximately sample from the density p by approximating it by a PSD model solving a linear system (as opposed to a SDP). This simpler and faster method comes at the expense of the stronger Assumption 6.2 needed to provide guarantees. As for algorithm 4, we first approximate the density with a PSD model and then sample from it using algorithm 2. The difference lies in the approximation step. We assume that we can evaluate a function g_p such that $g_p^2 \propto p$ (usually, this function will be proportional to the square root of p). We then approximate g_p with a Gaussian linear model Eq. (6.2) by solving a regularized empirical least squares problem, which is much faster than the solving of a SDP. Taking the square of that linear model, we obtain a PSD approximation of p from which we can sample using algorithm 2.

Step a): approximation of p. To fit a PSD model to p, we start by approximating g_p by a linear model $\widehat{g} = g(\bullet;\ \widehat{a}, \widetilde{X}_m, \eta)$ (see Eq. (6.2)), where $\widehat{a} \in \mathbb{R}^m$ is the solution to the empirical problem

$$\min_{a \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} |g(x_i;\ a) - g_p(x_i)|^2 + \lambda a^\top K_{mm} a, \tag{6.20}$$

where $g(x;\ a) := g(x;\ a, \widetilde{X}_m, \eta)$ and $g_n = (g_p(x_i))_{1 \leq i \leq n}$. $\widehat{a}$ is the solution to the system :

$$\left(K_{nm}^\top K_{nm} + (\lambda n) K_{mm}\right) a = K_{nm}^\top g_n, \tag{6.21}$$

which can be solved either directly in time $O(nm^2 + m^3)$ (Rudi et al., 2015) or using a preconditioned conjugate gradient method in time $O(m^3 + nm)$ (Rudi et al., 2017; Meanti et al., 2020; Marteau-Ferey et al., 2019). We then define $\widehat{f} = \widehat{g}^2$ which is a rank-1 PSD model with coefficients $\widehat{A} = \widehat{a}\widehat{a}^\top$, $\widehat{Z} = \int_{\mathcal{X}} \widehat{f}(x)dx = \|\widehat{g}\|^2_{L^2(\mathcal{X})}$ which is computable in closed form as the integral of a PSD model (see Eq. (6.3)), and our approximation $\widehat{p} = \widehat{f}/\widehat{Z}$ of p.

Solving Eq. (6.20) can be seen as solving a regularized empirical risk minimization problem for the Hellinger distance (see Eq. (6.32) in Sec. 6.3 .1); the regularization term $\lambda a^\top K_{mm} a$ being a regularization in the norm of the RKHS associated to the Gaussian kernel (see Appendix 6.B). The Hellinger distance is particularly adapted to exponential models of the form $\exp(-V(x))$ for a real-valued potential V, as the square root is simply $\exp(-V(x)/2)$.

Step b): sampling from the approximation $\widehat{p}$. We apply algorithm 2 to $\widehat{p}$ with a parameter ρ and on the hyper-rectangle $\mathcal{X}$. We denote with $p_{\texttt{sample}}$ the density $\widehat{p}_{\mathcal{X},\rho}$ given by Eq. (6.7), from which algorithm 2 effectively samples N i.i.d. samples by Theorem 6.1. This two step strategy is detailed in algorithm 5. SolveHellinger simply solves Eq. (6.20).

Theoretical Analysis We use the same notation as introduced in Sec. 6.4 .1. Once again, we assume that $\mathcal{X} = (-1, 1)^d$ for simplicity. In order to obtain good learning rates for algorithm 5, we make the following assumption, which is stronger than Assumption 6.1: it assumes that p can be written as a *single* square q^2, where q belongs to $\widetilde{W}^\beta(\mathcal{X})$.

Assumption 6.2 (Square distribution)**.** *There exists a function q belonging to $\widetilde{W}^\beta(\mathcal{X})$ such that $p = q^2$. Moreover, we have access to p only through function evaluations of the form $g_p(x)$, where $g_p \propto q$ and where the proportionality constant is unknown.*

Algorithm 5 Sampling from p using a rank-1 model

Input $p, \mathcal{X}, N$
Hyper-parameters (approximation) n, m, τ, λ
Hyper-parameters (sampling) ρ
Output N approximate samples from $p|_{\mathcal{X}}$

function ApproximateSamples($p, \mathcal{X}, N, n, m, \tau, \lambda, \rho$)
  X_n = UniformSamples($n, \mathcal{X}$)
  X_m =UniformSamples($m, \mathcal{X}$)
  A = SolveHellinger($p, X_n, X_m, \tau, \lambda$)
  $\widehat{p}(\cdot) = f(\cdot \,|\, A, X_m, \tau)$
  X_N = Sample($\mathcal{X}, N, \rho$) from $\widehat{p}$
  return X_N
end function

Note that this assumption is satisfied if $p \propto e^{-V(x)}$ for a potential V which is β times continuously differentiable which we can evaluate.

In proposition 6.2 and Theorem 6.4, we show that under certain regularity assumptions on p, given $\varepsilon > 0$, we can find hyper-parameters n, m, τ, λ and ρ such that $H(p, p_{\texttt{sample}}) \leq C\varepsilon$, i.e., that algorithm 5 generates N i.i.d. samples from a distribution $C\varepsilon$ close to p.

Proposition 6.2 (Performance of $\widehat{p}$)**.** *Let $\widetilde{\nu} > \min(1, d/(2\beta))$. There exists a constant ε_0 depending only on $\|q\|_{\widetilde{W}^{\beta}(\mathcal{X})}, \beta, d$, constants C_1, C_2, C_3, C_4 depending only on β, d and a constant C_1' depending only on $\beta, d, \widetilde{\nu}$ such that the following holds.*

Let $\delta \in (0, 1]$ and $\varepsilon \leq \varepsilon_0$, and assume m and n satisfy

$$m \geq C_1 \varepsilon^{-d/\beta} \log^d\left(\tfrac{C_2}{\varepsilon}\right) \log \tfrac{C_3}{\delta\varepsilon} \tag{6.22}$$

$$n \geq C_1' \varepsilon^{-2\widetilde{\nu}} \log \tfrac{8}{\delta}. \tag{6.23}$$

Let $\tau = \varepsilon^{-2/\beta}$ and $\lambda = \varepsilon^{2+d/\beta}$. With probability at least $1 - 3\delta$, it holds

$$H(\widehat{p}, p) \leq C_4 \|q\|_{\widetilde{W}^{\beta}(\mathcal{X})}\, \varepsilon. \tag{6.24}$$

Once again, the key takeaway from this proposition is that the number of samples n, m needed to perform the first step of the algorithm (approximation) is polynomial in the quantities $O(\varepsilon^{-1}), O(\varepsilon^{-d/\beta})$, thus leveraging the regularity β of q s.t. $q^2 = p$. When this is the case, we can find λ, τ such that the distance $H(p, \widehat{p})$ is of order ε. We provide a choice for ρ for the second step of the algorithm (sampling), in order to guarantee a bound for the Hellinger distance between the sampling distribution and the original distribution in the following theorem. It is proved as Theorem 6.10 in Appendix 6.F . In particular, it bounds the total complexity of the algorithm in terms of erf computations, as a function of N and the desired error ε.

Theorem 6.4 (Performance of $p_{\texttt{sample}}$)**.** *Under the assumptions and notations of proposition 6.2, there exists a constant C_5 depending only on d, β, such that the following holds. If on the one hand ρ is set either as $\varepsilon^{1+(d+2)/(2\beta)}$ or adaptively as ρ_ε^H (see Remark 23), then with probability at least $1 - 3\delta$,*

$$H(p, p_{\texttt{sample}}) \leq C_5 \|q\|_{\widetilde{W}^{\beta}(\mathcal{X})}\, \varepsilon. \tag{6.25}$$

Moreover, the adaptive ρ_ε^H is lower bounded by $\varepsilon^{1+(d+2)/\beta}/(C_5\, \|q\|_{\widetilde{W}^{\beta}(\mathcal{X})})$. In both cases, this guarantees that the complexity in terms of erf *computations is bounded by $O(Nm^2 \log \frac{1}{\rho})$, which, in*

terms of ε*, yields* $O\left(N\ \varepsilon^{-2d/\beta}\log^{2d+1}\left(\frac{1}{\varepsilon}\right)\log^2\left(\frac{1}{\delta\varepsilon}\right)\right)$ *where the* O *notation incorporates constants depending on* d, β*,* $\|q\|_{\widetilde{W}^\beta(\mathcal{X})}$*.*

6.4 .3 Discussion

The two methods presented in Sec. 6.4 .1 and Sec. 6.4 .2 share many interesting properties, both from a practical and theoretical viewpoint.

On the theoretical side, even though we only have access to the distribution up to a re-normalizing constant, this does not influence the theoretical results, i.e., the bounds we get only depend on the density p through its norm $\|p\|$. Moreover, the number of samples n, m needed (and hence the complexity of the sampling and of the approximation algorithm) is polynomial in the quantities $O(\varepsilon^{-1}), O(\varepsilon^{-d/\beta})$, showing that as soon as $\beta \geq d$, the dimension plays no role in the exponents of these error terms and thus *breaking the curse of dimensionality* in the rates. However, the constants in the $O(\cdot)$ term can be exponential in d, and without more hypotheses, **they are unimprovable** (Novak, 2006). We therefore keep a form of "curse of dimensionality" in the constants, but not in the rate. Concretely this means that we need a number of points in the order of the constants before having a reasonable error (i.e., $\varepsilon = 1$). However, as soon as this number is reached, one can rapidly gain in precision, if the function is regular. Moreover, in practice, we do not always pay this exponential constant, owing to some additional regularity of the function. Interestingly, this phenomenon is shared with approximation, learning and optimization problems over a wide family of functions (see (Novak, 2006) for more details).

On the practical side, note that both algorithm 4 and algorithm 5 can be run for any hyper-parameter (even though this might not have statistical sense), making it easy to use. More importantly, we can evaluate the learnt model a posteriori using empirical metrics (like the empirical total variation distance or the empirical Hellinger distance for instance) on a new data set generated uniformly from $\mathcal{X}$. We could also evaluate it using certain empirical divergences since we are able to sample from $p_{\texttt{sample}}$. This can help in both selecting τ and λ by validation, as well as in simply evaluating the performance of the learnt model, with error bars if needed. In Fig. 6.2 for example, we evaluate the performance of learnt PSD models for the empirical Hellinger distance. We perform 5 different tests and plot the associated error bars: this methods seems very robust for evaluation.

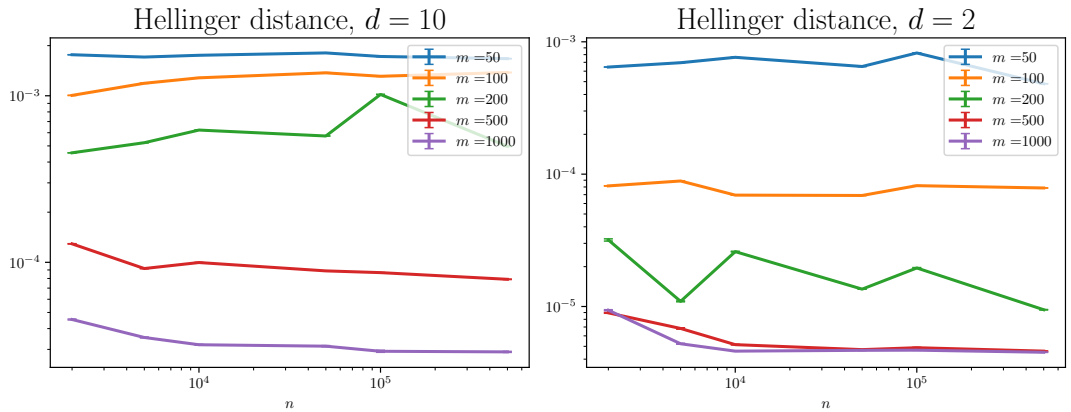

Figure 6.2: Evolution of the empirical Hellinger distance on a test set, between learnt distribution $\widehat{p}$ and target distribution p when increasing the number of evaluation points, for fixed values of m. We learn $\widehat{p}$ as a rank one PSD model through Eq. (6.20). *(left)* Learning p_2 with $d = 10$ defined in Sec. 6.5 . *(right)* Learning p_1 defined in Sec. 6.5 .

6.5 Experiments

The experiments in this work were executed on a MacBook Pro equipped with a 2.8 GHz Quad-Core Intel Core i7 processor and 16Gb of RAM[2].

Influence of m and n. In Fig. 6.2, we show how m and n interact in order to set the precision of our approximation in the learning phase (step a)). For $m = 50, 100, 200$, m is so small that increasing n beyond 1000 does not yield better performance (the variations are due to the fact that points are always resampled accross experiments). However, when $m = 500, 1000$, we see that increasing n yields better performance, before arriving at a plateau. This plateau corresponds to the transition from the phase where n is the limiting statistical factor to the phase where m is.

Qualitative performance of our algorithm. In Fig. 6.3, we show an example of the way our algorithm approximates a certain target density p_1 known up to a renormalization constant: $p_1(x) \propto 0.08k_{0.7}(x, -(1,1)) - 0.4k_{0.6}(x, (1,1)) + 0.4k_{0.7}(x, (1,1))$. In the top left figure, a heat map of p_1 is plotted. Note that p_1 is not a Gaussian PSD model, as the widths of the Gaussian kernels are not the same. We then use algorithm 5 to approximate p_1 by a rank one PSD model $\widehat{p}_1$ (whose heat-map is plotted on the top right figure) and then sample $N = 1000$ samples from this approximation (plotted in the bottom left figure). Note that in order to approximate p_1 by $\widehat{p}_1$, $n = 10^5, m = 300$ were fixed and $\tau = 2, \lambda = 10^{-9}$ were selected on a test set. In Sec. 6.G , we perform and comment another experiment when trying to learn a density which is not smooth (and therefore out of the scope of Theorems 6.3 and 6.4).

Quantitative performance of our algorithm. To further demonstrate the promising nature of our sampling algorithm, we tried learning the density $p_2(x) \propto (k_{1/5}(x, (1, ..., 1)) - k_{1/5}(x, -(1, ..., 1)))^2$ on $Q = [-1, 1]^d$, for $d = 5$. Contrary to p_1, this is a PSD model, we can sample from it with very high precision (here, we chose $\rho = 10^{-6}$). Our goal here is to be able to compare methods through the generated samples.

We compared the performance of our model to the naive gridding algorithm which, if allowed n function evaluations, computes a grid G of side $n^{1/d}$, which we identify to the set of centers of the tiles of the grid, and evaluates p at each point in the grid. To sample a point, one chooses a point $g \in G$ with probability $p(g)/\sum_{h\in G} p(h)$, and then draws a sample uniformly in that tile. It is the algorithm called 'grid' in the bottom right figure of Fig. 6.3.

We compare our algorithm with the gridding algorithm by fixing the number n of function evaluations of p each method is allowed, and computing the distance between each method and the ground truth. The distance we use between distributions is the empirical version of the Maxmium Mean Discrepancy distance (MMD) (Sriperumbudur et al., 2010, 2011), which is defined, for the Gaussian kernel k_η of parameter η, as $d_\eta(p, \tilde{p}) = \|\mathbb{E}_{X\sim p}[\phi_\eta(X)] - \mathbb{E}_{X\sim\tilde{p}}[\phi_\eta(X)]\|_{\mathcal{H}_\eta}$ where ϕ_η is the embedding associated to the Gaussian kernel k_η (for more details, see Appendix 6.A). This distance can be approximated using N samples $(x_i)_{1\leq i\leq N}$ from p and N samples $(\tilde{x}_j)_{1\leq j\leq N}$ from $\tilde{p}$ as $\widehat{d}_\eta(p, \tilde{p}) = \left\| \frac{1}{N}\sum_{i=1}^N \phi_\eta(x_i) - \frac{1}{N}\sum_{j=1}^N \phi_\eta(\tilde{x}_j) \right\|_{\mathcal{H}_\eta}$. This quantity can be computed explicitly using kernel matrices (Sriperumbudur et al., 2010). However, Tolstikhin, Sriperumbudur, and Schölkopf (2016) show that the minimax rate cannot exceed $1/\sqrt{N}$, i.e., that $\widehat{d}_\eta$ approximates d_η only with precision of order $1/\sqrt{N}$.

[2]The code is available at https://github.com/umarteau/sampling_psd_models

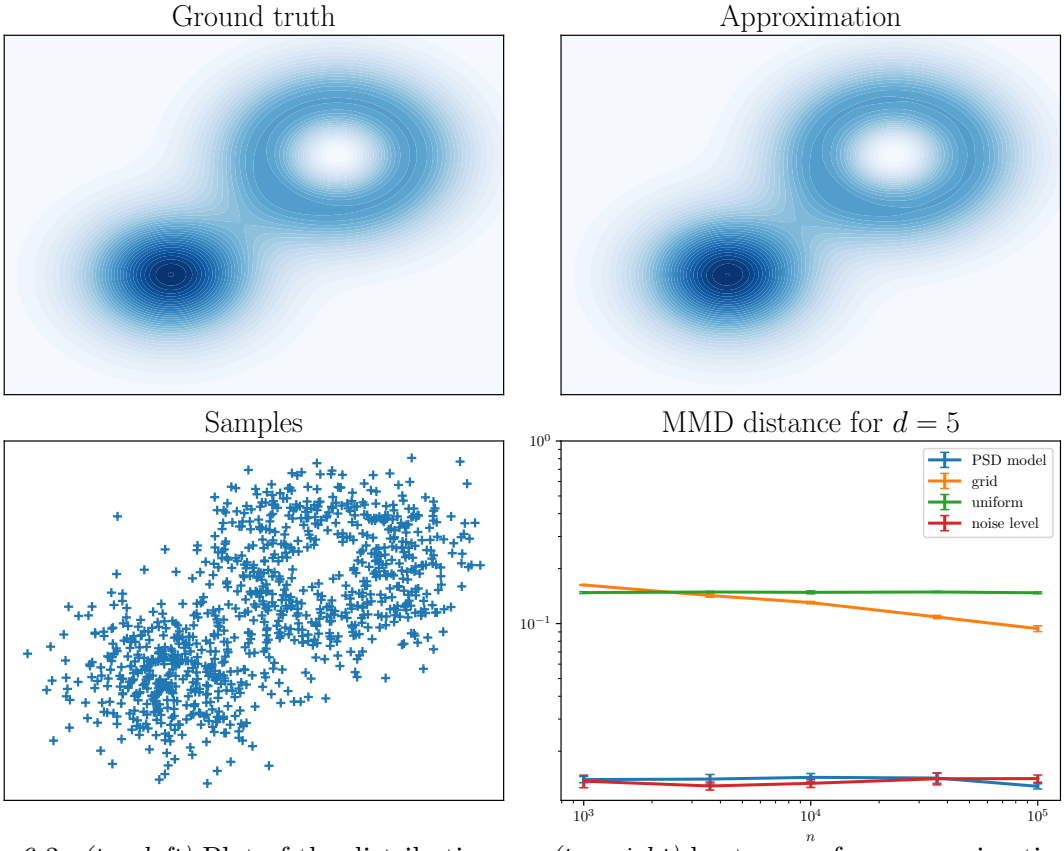

Figure 6.3: *(top left)* Plot of the distribution p_1, *(top right)* heat map of an approximation $\widehat{p}_1$ of p_1. *(bottom left)* samples generated from $\widehat{p}_1$, *(bottom right)* performance of our method in MMD distance.

In our experiments, we take $N = 10^4$. We compute the empirical distances $\widehat{d}_\eta$ five times using newly generated samples from each distribution, and compute an empirical mean and standard deviation, reported as error bars on the plot. When approximating p_2 by a PSD model using algorithm 5, we take $m = 50$, as there is no need to increase m to reach better precision than the target distribution for $\widehat{d}_\eta$. We take $\rho = 10^{-3}$ and select τ, λ by using half of the evaluation points as a test set.

The results reported on the bottom-right plot of Fig. 6.3 show that in dimension 5, the 'grid' method is not competitive anymore, and is close to the uniform distribution in performance for $\eta = 2$. Note that the choice of η in a wide range from 0.1 to 10 does not change these results. They also show that when taking only $N = 10^4$ to approximate the MMD distance, our method is below the noise level.

6.6 Extensions, future work

In this paper, we have introduced a method for sampling any distribution from function values by first approximating it with a so-called PSD model and then sampling from this PSD model using the algorithm introduced in Sec. 6.3 .

Natural extensions of this work include the fact that while we cast a least squares problem in Sec. 6.4 .1, we can actually minimize more general convex losses adapted to distributions, such as maximum log-likelihood estimation. Moreover, as mentioned in Sec. 6.3 , the proposed algorithm

only relies on integral computations, and could therefore be extended to other kernels, provided they can easily be integrated on hyper-rectangles.

Future work will start with trying to scale the sampling method up in terms of generation of samples, by both theoretical means (to make computation saving approximations) and computational means (use of GPUs, parallelization).

Acknowledgements. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001(PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grants SEQUOIA 724063 and REAL 947908), and support by grants from Région Ile-de-France.

Organization of the Supplementary Material

6.A . Definitions and notations

We set the main notations and tools of the appendix (Fourier transform, vector and matrix notations, notations concerning hyper-rectangles, RKHS and specifically the Gaussian kernel).

6.A .1. Sobolev spaces

In this section, we focus more on notations and basic results concerning Sobolev spaces, as they will be our main tool to measure the regularity of a function.

6.A .2. Measuring distances between probability densities

In this section, we define and compare the basic distances we will be using to compare probability distributions in the paper, since we are always "approximating" a certain distribution with another. In particular, we define the total variation, Hellinger and Wasserstein distances.

6.A .3. General PSD models

We define PSD models in general (Marteau-Ferey et al., 2020; Rudi and Ciliberto, 2021). They will be our main tool for approximation and sampling, and relates to the more restrictive definition in Sec. 6.2 .

6.B . Properties of the Gaussian RKHS

Throughout the paper the Gaussian kernel k_η and the associated Gaussian RKHS will be central objects. We introduce different properties and results.

6.B .1. Properties of the Gaussian kernel k_η

We introduce certain properties of the Gaussian kernel involving products, as well as a bound on the derivative of the associated embedding in Lemma 6.2.

6.B .2. Useful Matrices and Linear Operators on the Gaussian RKHS

We introduce the most important theoretical objects of the paper. We introduce kernel matrices, matrices which will appear in the integration of Gaussian PSD models, operators which relate L^2 to the RKHS $\mathcal{H}_\eta$, operators which allow to discretize using samples and "compression" operators which allow concise representations.

6.B .3. Approximation properties of the Gaussian kernel

We prove two important results concerning the approximation properties of the Gaussian RKHS in proposition 6.7 and the concise representation of models in Lemma 6.3.

6.C . Properties of Gaussian PSD models

We present the results specific to Gaussian PSD models. These results are often reformulations of theorems presented by Rudi and Ciliberto (2021).

6.C .1. Bounds on the support and the derivatives

We present result to understand how the mass of a Gaussian PSD model is concentrated (Lemma 6.4) and how the derivative of a Gaussian PSD model can be bounded using only its parameters (Lemma 6.5).

6.C .2. Compression as a Gaussian PSD model

We restate Theorem C.4 of Rudi and Ciliberto (2021) as Theorem 6.5 on the effect of a compression operator on a PSD model.

6.C .3. Approximation properties of Gaussian PSD model

We refine Theorem D.4 of Rudi and Ciliberto (2021) in Theorem 6.6 in order to approximate a sum of squares using a PSD model on the Gaussian RKHS $\mathcal{H}_\eta$.

6.D . The sampling algorithm

We prove that the sampling algorithm indeed returns N i.i.d. samples from the right distribution, and characterize the distance between the sampling distribution and the original PSD distribution.

6.D .1. Dyadic decompositions and convergence of algorithm 2

We formally prove that algorithm 2 finishes and returns N samples from a distribution characterized by a structural induction formula (see Lemma 6.6).

6.D .2. Proof of Theorem 6.1

We prove Theorem 6.1 by structural induction, showing that when the samples are randomly shuffled, we end up with N i.i.d. samples from the distribution defined in Eq. (6.7). This is done by matching the distribution with the one from the previous section using a structural induction.

6.D .3. Evaluating the error of the sampling algorithm : proof of Theorem 6.2

We prove Theorem 6.2 in Theorem 6.7, bounding the distance between the distribution of the PSD model and the actual distribution from which algorithm 2 samples (see Eq. (6.7)). This is done in different distances, all related to the problem in different way (Wasserstein is the most adapted in spirit, but we also need stronger distances such as total variation and Hellinger, which can be bounded using Lipschitz constants of the PSD models).

6.D .4. Time complexity

We illustrate that the time complexity of the algorithm is indeed taken up by the integral computations.

6.E . A general method of approximation and sampling

We prove that we can approximate any probability distribution satisfying Assumption 6.1 using non necessarily normalized function values, by solving Eq. (6.15) with the right parameters in proposition 6.10 which is labeled in the main text as proposition 6.1. We then show that applying algorithm 2 with the right value of ρ yields a good sampling algorithm from a good approximation of the distribution. This proves Theorem 6.3 and is proved here as Theorem 6.8.

6.F . Approximation and sampling using a rank one PSD model

We prove that we can approximate any probability distribution satisfying Assumption 6.2 using non necessarily normalized function values, by solving Eq. (6.20) with the right parameters in proposition 6.11 which is labeled in the main text as proposition 6.2. This has an advantage compared to the previous method which is that the approximation phase is much faster (it solves a linear system instead of an SDP). We then show that applying algorithm 2 with the right value of ρ yields a good sampling algorithm from a good approximation of the distribution. This proves Theorem 6.4 and is proved here as

Theorem 6.10.

6.G . Additional experimental details

6.A Definitions and notations

In this section we recall results by Rudi and Ciliberto (2021) which will be useful in the different statements and proofs.

Basic vector and matrix notations. Let $n, d \in \mathbb{N}$. We denote by $\mathbb{R}^d_{++}$ the space vectors in $\mathbb{R}^d$ with positive entries, $\mathbb{R}^{n\times d}$ the space of $n \times d$ matrices, $\mathbb{S}^n_+ = \mathbb{S}_+(\mathbb{R}^n)$ the space of positive semidefinite $n \times n$ matrices. Given a vector $\eta \in \mathbb{R}^d$, we denote $\mathrm{diag}(\eta) \in \mathbb{R}^{d\times d}$ the diagonal matrix associated to η. We denote by $A \circ B$ the entry-wise product between two matrices A and B. We denote by $\|A\|_F, \|A\|, \det(A), \mathrm{vec}(A)$ and $A^\top$ respectively the Frobenius norm, the operator norm (i.e. maximum singular value), the determinant, the (column-wise) vectorization of a matrix and the (conjugate) transpose of A. With some abuse of notation, where clear from context we write element-wise products and division of vectors $u, v \in \mathbb{R}^d$ as $uv, u/v$. The term $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector with all entries equal to 1.

Hyper-rectangles Define a hyper-rectangle Q as a product of the form $\prod_{k=1}^d [a_k, b_k[$, where $a \le b$. Given a hyper-rectangle Q we denote its extremities with $a(Q) \le b(Q) \in \mathbb{R}^d$ (i.e. $Q = \prod_{k=1}^d [a_k(Q), b_k(Q)[$), and its side-lengths $\rho(Q) = b(Q) - a(Q)$. We sometimes omit Q when it is implied by the context.

We will also use the so-called *error function*, which is defined as follows :

$$\mathrm{erf}(x) = \tfrac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

This function is implemented as an elementary function in most libraries.

Multi-index notation Let $\alpha \in \mathbb{N}^d$, $x \in \mathbb{R}^d$ and f be an infinitely differentiable function on $\mathbb{R}^d$, we introduce the following notation

$$|\alpha| = \sum_{j=1}^d \alpha_i, \quad \alpha! = \prod_{j=1}^d \alpha_j!, \quad x^\alpha = \prod_{j=1}^d x_j^{\alpha_j}, \quad \partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

We introduce also the notation D^α that corresponds to the multivariate distributional derivative of order α and such that

$$D^\alpha f = \partial^\alpha f$$

for functions that are differentiable at least $|\alpha|$ times (Adams and Fournier, 2003).

Fourier Transform Given two functions $f, g : \mathbb{R}^d \to \mathbb{R}$ on some set $\mathbb{R}^d$, we denote by $f \cdot g$ the function corresponding to *pointwise product* of f, g, i.e.,

$$(f \cdot g)(x) = f(x)g(x), \quad \forall x \in \mathbb{R}^d.$$

Let $f, g \in L^1(\mathbb{R}^d)$ we denote the *convolution* by $f \star g$

$$(f \star g)(x) = \int_{\mathbb{R}^d} f(y)g(x-y)dy.$$

We now recall some basic properties, that will be used in the rest of the appendix.

Proposition 6.3 (Basic properties of the Fourier transform (Wendland, 2004), Chapter 5.2.)**.**

(a) *There exists a linear isometry* $\mathcal{F} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ *satisfying*

$$\mathcal{F}[f] = \int_{\mathbb{R}^d} e^{-2\pi i\, \omega^\top x}\, f(x)\, dx \quad \forall f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d),$$

where $i = \sqrt{-1}$*. The isometry is uniquely determined by the property in the equation above.*

(b) *Let* $f \in L^2(\mathbb{R}^d)$*, then* $\|\mathcal{F}[f]\|_{L^2(\mathbb{R}^d)} = \|f\|_{L^2(\mathbb{R}^d)}$*.*

(c) *Let* $f \in L^2(\mathbb{R}^d), r > 0$ *and define* $f_r(x) = f(\frac{x}{r}), \forall x \in \mathbb{R}^d$*, then* $\mathcal{F}[f_r](\omega) = r^d \mathcal{F}[f](r\omega)$*.*

(d) *Let* $f, g \in L^1(\mathbb{R}^d)$*, then* $\mathcal{F}[f \cdot g] = \mathcal{F}[f] \star \mathcal{F}[g]$*.*

(e) *Let* $\alpha \in \mathbb{N}^d$*,* $f, D^\alpha f \in L^2(\mathbb{R}^d)$*, then* $\mathcal{F}[D^\alpha f](\omega) = (2\pi i)^{|\alpha|}\omega^\alpha \mathcal{F}[f](\omega)$*,* $\forall \omega \in \mathbb{R}^d$*.*

(f) *Let* $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$*, then* $\|\mathcal{F}[f]\|_{L^\infty(\mathbb{R}^d)} \leq \|f\|_{L^1(\mathbb{R}^d)}$*.*

(g) *Let* $f \in L^\infty(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$*, then* $\|f\|_{L^\infty(\mathbb{R}^d)} \leq \|\mathcal{F}[f]\|_{L^1(\mathbb{R}^d)}$*.*

Reproducing kernel Hilbert spaces for translation invariant kernels. We now list some important facts about reproducing kernel Hilbert spaces in the case of translation invariant kernels on $\mathbb{R}^d$. For this paragraph, we refer to the works by Steinwart and Christmann (2008); Wendland (2004). For the general treatment of positive kernels and Reproducing kernel Hilbert spaces, see the works by Aronszajn (1950); Steinwart and Christmann (2008). Let $v : \mathbb{R}^d \to \mathbb{R}$ such that its Fourier transform $\mathcal{F}[v] \in L^1(\mathbb{R}^d)$ and satisfies $\mathcal{F}[v](\omega) \geq 0$ for all $\omega \in \mathbb{R}^d$. Then, the following hold.

(a) The function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as $k(x, x') = v(x - x')$ for any $x, x' \in \mathbb{R}^d$ is a positive kernel and is called *translation invariant kernel*.

(b) The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ and its norm $\|\cdot\|_{\mathcal{H}}$ are characterized by

$$\mathcal{H} = \{f \in L^2(\mathbb{R}^d) \mid \|f\|_{\mathcal{H}} < \infty\}, \quad \|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[v](\omega)} d\omega, \tag{6.26}$$

(c) $\mathcal{H}$ is a separable Hilbert space, whose inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is characterized by

$$\langle f, g \rangle_{\mathcal{H}} = \int_{\mathbb{R}^d} \frac{\mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}}{\mathcal{F}[v](\omega)} d\omega.$$

In the rest of the paper, when clear from the context we will simplify the notation of the inner product, by using $f^\top g$ for $f, g \in \mathcal{H}$, instead of the more cumbersome $\langle f, g \rangle_{\mathcal{H}}$.

(d) The feature map $\phi : \mathbb{R}^d \to \mathcal{H}$ is defined as $\phi(x) = k(x - \cdot) \in \mathcal{H}$ for any $x \in \mathbb{R}^d$.

(e) The functions in $\mathcal{H}$ have the *reproducing property*, i.e.,

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}, x \in \mathbb{R}^d,$$

in particular $k(x', x) = \langle \phi(x'), \phi(x) \rangle_{\mathcal{H}}$ for any $x', x \in \mathbb{R}^d$.

We now introduce the main tool of our analysis, the Gaussian RKHS, which will be further explored in Appendix 6.C .

Example 6.1 (Gaussian Reproducing Kernel Hilbert Space)**.** *Let $\eta \in \mathbb{R}^d_{++}$ and $k_\eta(x, x') = e^{-(x-x')^\top \operatorname{diag}(\eta)(x-x')}$, for $x, x' \in \mathbb{R}^d$ be the Gaussian kernel with precision η. The function k_η is a translation invariant kernel, since $k_\eta(x, x') = v(x - x')$ with $v(z) = e^{-\|D^{1/2}z\|^2}, D = \operatorname{diag}(\eta)$ and $\mathcal{F}[v](\omega) = c_\eta e^{-\pi^2\|D^{-1/2}\omega\|^2}$, $c_\eta = \pi^{d/2}\det(D)^{-1/2}$, for $\omega \in \mathbb{R}^d$ is in $L^1(\mathbb{R}^d)$ and satisfies $\mathcal{F}[v](\omega) \geq 0$ for all $\omega \in \mathbb{R}^d$. The associated reproducing kernel Hilbert space $\mathcal{H}_\eta$ is defined according to Eq. (6.26), with norm*

$$\|f\|^2_{\mathcal{H}_\eta} = \frac{1}{c_\eta}\int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 \, e^{\pi^2\|D^{-1/2}\omega\|^2} \, d\omega, \qquad \forall f \in L^2(\mathbb{R}^d). \tag{6.27}$$

The inner product and the feature map ϕ_η are defined as in the discussion above.

6.A .1 Sobolev spaces

Let $\beta \in \mathbb{N}, p \in [1, \infty]$ and let $\Omega \subseteq \mathbb{R}^d$ be an open set. The set $L^p(\Omega)$ denotes the set of p-integrable functions on Ω for $p \in [1, \infty)$ and that of the essentially bounded on Ω when $p = \infty$. The set $W^\beta_p(\Omega)$ denotes the Sobolev space, i.e., the set of measurable functions with their distributional derivatives up to β-th order belonging to $L^p(\Omega)$,

$$W^\beta_p(\Omega) = \{f \in L^p(\Omega) \mid \|f\|_{W^\beta_p(\Omega)} < \infty\}, \quad \|f\|^p_{W^\beta_p(\Omega)} = \sum_{|\alpha|\leq\beta} \|D^\alpha f\|^p_{L^p(\Omega)}, \tag{6.28}$$

where D^α denotes the distributional derivative. In the case of $p = \infty$,

$$\|f\|_{W^\beta_\infty(\Omega)} = \max_{|\alpha|\leq\beta} \|D^\alpha f\|_{L^\infty(\Omega)}$$

We now recall some basic results about Sobolev spaces that are useful for the proofs in this paper. First we start by recalling the restriction properties of Sobolev spaces. Let $\Omega \subseteq \Omega' \subseteq \mathbb{R}^d$ be two open sets. Let $\beta \in \mathbb{N}$ and $p \in [1, \infty]$. By definition of the Sobolev norm above we have

$$\|g|_\Omega\|_{W^s_p(\Omega)} \leq \|g\|_{W^s_p(\Omega')},$$

and so $g|_\Omega \in W^s_p(\Omega)$ for any $g \in W^s_p(\Omega')$. Now we recall the extension properties of Sobolev spaces, which will allow us to consider the case

The formal definition of a set with Lipschitz boundary is provided by Adams and Fournier (2003). Note that if $\mathcal{X} = (-1, 1)^d$, as will be the case later on for simplicity, then $\mathcal{X}$ is bounded and has Lipschitz boundary.

The following result shows that being in an intersection space allows to extend the function to the whole of $\mathbb{R}^d$. This will be useful in order to use the properties of translation invariant kernels in order to approximate functions which are a priori defined only on $\mathcal{X}$ but which we extend using this result.

Proposition 6.4 (Rudi and Ciliberto (2021, Corollary A.3))**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a non-empty open set with Lipschitz boundary. Let $\beta \in \mathbb{N}, p \in [1, \infty]$. Then for any function $f \in W^\beta_p(\mathcal{X}) \cap L^\infty(\mathcal{X})$ there exists an extension $\tilde{f}$ on $\mathbb{R}^d$, i.e. a function $\tilde{f} \in W^\beta_p(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ such that*

$$f = \tilde{f}|_\mathcal{X} \text{ a.e. on } \mathcal{X}, \quad \|\tilde{f}\|_{L^\infty(\mathbb{R}^d)} \leq C\|f\|_{L^\infty(\mathcal{X})}, \quad \|\tilde{f}\|_{W^\beta_p(\mathbb{R}^d)} \leq C'\|f\|_{W^\beta_p(\mathcal{X})}.$$

The constant C depends only on $\mathcal{X}, d$, and the constant C' only on $\mathcal{X}, \beta, d, p$

The following proposition gives an idea of what these intersection spaces contain.

Proposition 6.5 (Rudi and Ciliberto (2021, Proposition A.4)). *Let $\mathcal{X}$ be an open bounded set with Lipschitz boundary. Let f be a function that is m times differentiable on the closure of $\mathcal{X}$. Then there exists a function $\tilde{f} \in W_p^m(\mathcal{X}) \cap L^\infty(\mathcal{X})$ for any $p \in [1, \infty]$, such that $\tilde{f} = f$ on $\mathcal{X}$.*

The following proposition provides a useful characterization of the space $W_2^\beta(\mathbb{R}^d)$ in terms of Fourier transform; this will be particularly useful when approximating functions in $W_2^\beta(\mathbb{R}^d)$ by functions in a Gaussian RKHS $\mathcal{H}_\eta$ using the characterization of the norm in terms of Fourier transform for those kernels in Eq. (6.26).

Proposition 6.6 (Characterization of the Sobolev space $W_2^k(\mathbb{R}^d)$, (Wendland, 2004), (Rudi and Ciliberto, 2021, Proposition A.4)). *Let $k \in \mathbb{N}$. The norm of the Sobolev space $\|\cdot\|_{W_2^k(\mathbb{R}^d)}$ is equivalent to the following norm*

$$\|f\|'^2_{W_2^k(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |\mathcal{F}[f](\omega)|^2 \, (1 + \|\omega\|^2)^k \, d\omega, \quad \forall f \in L^2(\mathbb{R}^d)$$

and satisfies

$$\tfrac{1}{(2\pi)^{2k}} \|f\|^2_{W_2^k(\mathbb{R}^d)} \leq \|f\|'_{W_2^k(\mathbb{R}^d)} \leq 2^{2k} \|f\|^2_{W_2^k(\mathbb{R}^d)}, \quad \forall f \in L^2(\mathbb{R}^d) \tag{6.29}$$

Moreover, when $k > d/2$, then $W_2^k(\mathbb{R}^d)$ is a reproducing kernel Hilbert space.

6.A .2 Measuring distances between probability densities

In this work, since our aim is to approximate a probability distribution, we will often compare probability distributions, with different distances.

To simplify definitions, we will only consider distances between probability densities p_1, p_2 defined on a Borel subset $\mathcal{X}$ of $\mathbb{R}^d$ with respect to the Lebesgue measure. Note that while the total variation distance, the Hellinger distance and the Wasserstein distance do not actually depend on the choice of such a base measure and can be defined intrinsically, the L^2 distance cannot; that is why it is less appropriate from a statistical point of view. We consider it here because it is the natural distance in which we are able to solve Eq. (6.15).

The total variation (TV) or L^1 distance :

$$d_{TV}(p_1, p_2) := \|p_1 - p_2\|_{L^1(\mathcal{X})} = \int_{\mathcal{X}} |p_1(x) - p_2(x)| dx. \tag{6.30}$$

This distance can also be expressed using a dual formulation (Lucien Le Cam, 1990, chapter 3.2).

$$d_{TV}(p_1, p_2) = \sup_{|f| \leq 1} \left| \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x)) \, dx \right| \tag{6.31}$$

The Hellinger distance : (this distance is particularly suitable in the case of exponential models; see the works by Lucien Le Cam (1990) and in particular Chapter 3).

$$H(p_1, p_2) := \|\sqrt{p_1} - \sqrt{p_2}\|_{L^2(\mathcal{X})} = \left(\int_{\mathcal{X}} |\sqrt{p_1}(x) - \sqrt{p_2}(x)|^2 \, dx \right)^{1/2} \tag{6.32}$$

The Wasserstein distance In the case where $\mathcal{X}$ is bounded (for simplicity), the p Wasserstein distance for $p \geq 1$ (see chapter 5 by Santambrogio (2015)):

$$\mathbb{W}_p^p(p_1, p_2) = \inf_{\gamma \in \Pi(p_1, p_2)} \int_{\mathcal{X} \times \mathcal{X}} |x - y|^p \, d\gamma(x, y), \tag{6.33}$$

where $\Pi(p_1, p_2)$ is the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals p_1 and p_2. Note that one has the following easier dual formulation when $p = 1$ (see the chapter on Kantorovich duality by Santambrogio (2015)):

$$\mathbb{W}_1(p_1, p_2) = \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x))dx, \tag{6.34}$$

where $\mathrm{Lip}_1(\mathcal{X})$ is the set of 1-Lipschitz functions on $\mathcal{X}$. Wasserstein distances capture the moving of mass; they are quite weak but are well-adapted to capture the behavior of our sampling algorithm which approximates probability densities on each hyper-rectangle.

The L^2 distance :

$$\|p_1 - p_2\|_{L^2(\mathcal{X})} = \left(\int_{\mathcal{X}} (p_1(x) - p_2(x))^2 \, dx \right)^{1/2} \tag{6.35}$$

Relating these difference distances . The following well known bounds exist between distances.

$$H^2(p_1, p_2) \leq d_{TV}(p_1, p_2) \leq \sqrt{2} H(p_1, p_2). \tag{6.36}$$

Moreover, if $\mathcal{X}$ is bounded, we have for any $p \geq 1$, using the Holder inequality:

$$\mathbb{W}_p(p_1, p_2) \leq \mathrm{diam}(\mathcal{X})^{(p-1)/p} \mathbb{W}_1(p_1, p_2)^{1/p}, \tag{6.37}$$

$$\mathbb{W}_1(p_1, p_2) \leq \mathrm{diam}(\mathcal{X}) d_{TV}(p_1, p_2), \tag{6.38}$$

$$d_{TV}(p_1, p_2) \leq |\mathcal{X}|^{1/2} \|p_1 - p_2\|_{L^2(\mathcal{X})}, \tag{6.39}$$

where $\mathrm{diam}(\mathcal{X})$ denotes the diameter of the set $\mathcal{X}$.

6.A .3 General PSD models

In this section, we recall the definition of a PSD model more generally as introduced by Rudi and Ciliberto (2021).

Following Marteau-Ferey, Bach, and Rudi (2020); Rudi and Ciliberto (2021), we consider the family of positive semi-definite (PSD) models, namely non-negative functions parametrized by a feature map $\phi : \mathcal{X} \to \mathcal{H}$ from an input space $\mathcal{X}$ to a suitable feature space $\mathcal{H}$ (a separable Hilbert space e.g. $\mathbb{R}^q$) and a linear operator $M \in \mathbb{S}_+(\mathcal{H})$, of the form

$$f(x;\ M, \phi) = \phi(x)^\top M \ \phi(x). \tag{6.40}$$

PSD models offer a general way to parametrize non-negative functions (since M is positive semidefinite, $f(x;\ M, \phi) \geq 0$ for any $x \in \mathcal{X}$) and enjoy several additional appealing properties discussed in the following. In this work. we focus on a special family of models i.e. Gaussian PSD models defined in Sec. 6.2 and Eq. (6.1). These models parametrize probability densities

over $\mathcal{X} \subset \mathbb{R}^d$. It is a special case of Eq. (6.40) where *i)* $\phi = \phi_\eta : \mathbb{R}^d \to \mathcal{H}_\eta$ is a feature map associated to the Gaussian kernel defined in Example 6.1, or by Scholkopf and Smola (2001) and, *ii)* the operator M lives in the span of $\phi(x_1), \dots, \phi(x_n)$ for a given set of points $(x_i)_{i=1}^n$, namely there exists $A \in \mathbb{S}_+(\mathbb{R}^n)$ such that $M = \sum_{ij} A_{ij} \phi_\eta(x_i) \phi_\eta(x_j)^\top$.

Thus, given the triplet (A, X, η) characterizing the Gaussian PSD model in Eq. (6.1), we have

$$\sum_{1 \leq i,j \leq n} A_{ij} k_\eta(x, x_i) k_\eta(x, x_j) = f(x; A, X, \eta) \qquad = f(x; M, \phi_\eta)$$
$$M = \sum_{1 \leq i,j \leq n} A_{ij} \phi_\eta(x_i) \otimes \phi_\eta(x_j),$$

where $(u \otimes v)w = uv^\top w = \langle v, w \rangle u$.

6.B Properties of the Gaussian RKHS

In this section, we introduce notations and results associated to the Gaussian RKHS (see Example 6.1) $\mathcal{H}_\eta$ for a given $\eta \in \mathbb{R}^d_{++}$ (η will sometimes be taken in the form $\tau \mathbf{1}_d$). Recall that the Gaussian embedding is written $\phi_\eta : \mathbb{R}^d \to \mathcal{H}_\eta$ and that the Gaussian kernel is denoted with k_η.

6.B .1 Properties of the Gaussian kernel k_η

The following lemma has an immediate proof.

Lemma 6.1 (product of Gaussian kernels)**.** *Let $K \in \mathbb{N}$, let $\eta_1, ..., \eta_K \in \mathbb{R}^d_{++}$ and let $y_1, ..., y_K \in \mathbb{R}^d$. The following equality holds:*

$$\forall x \in \mathbb{R}^d, \ \prod_{k=1}^K k_{\eta_k}(x, y_k) = k_{\overline{\eta}}(x, \overline{y}) \prod_{k=1}^K k_{\eta_k}(y_k, \overline{y})$$

where $\overline{\eta} = \sum_{k=1}^K \eta_k$ and $\overline{y} = \sum_k \eta_k y_k / \overline{\eta}$

Let us now state an useful corollary.

Corollary 6.1. *Let $\eta \in \mathbb{R}^d_{++}$, $y_1, y_2 \in \mathbb{R}^d$. Then*

$$\forall x \in \mathbb{R}^d, \ k_\eta(x, y_1) k_\eta(x, y_2) = k_{2\eta}(x, (y_1 + y_2)/2) k_{\eta/2}(y_1, y_2). \tag{6.41}$$

Lemma 6.2 (Gaussian embedding derivative)**.** *Let $\eta \in \mathbb{R}^d_{++}$, $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{N}^d$. The derivative $\partial_\alpha \phi_\eta(x)$ is well defined in $\mathcal{H}_\eta$, and $\|\partial_\alpha \phi_\eta(x)\|_{\mathcal{H}_\eta} = 2^{|\alpha|/2} \eta^{\alpha/2}$. Moreover, if $g \in \mathcal{H}_\eta$, then $\sup_{x \in \mathbb{R}^d} |(\partial_\alpha g)(x)| \leq 2^{|\alpha|/2} \eta^{\alpha/2} \|g\|_{\mathcal{H}_\eta}$.*

Proof. Let $\alpha \in \mathbb{N}^d$ and let $v_\eta(z) = k_\eta(z, 0) = \exp(-z^\top \operatorname{diag}(\eta) z)$. If the function $\frac{\partial^\alpha}{\partial x^\alpha} k_\eta(x, y)$ belongs to $\mathcal{H}_\eta$, then $\partial_\alpha \phi_\eta(x)$ is in $\mathcal{H}_\eta$ and is equal to that function by the reproducing property.

First, note that

$$\forall x, y \in \mathbb{R}^d, \ \tfrac{\partial^\alpha}{\partial x^\alpha} k_\eta(x, y) = (-1)^{|\alpha|} \partial_\alpha \tau_x[v_\eta](y),$$

where $\tau_x : f \mapsto f(\cdot - x)$, commutes with the differential operator ∂_α, and satisfies the following relation wrt to the Fourier transform : $\mathcal{F}[\tau_x g](\xi) = e^{-2i\pi x\xi}\mathcal{F}[f](\xi)$. Hence, using (e) of proposition 6.3, we get the following fourier transform wrt y:

$$\mathcal{F}_y[\tfrac{\partial^\alpha}{\partial x^\alpha}k_\eta(x,y)](\xi) = (-2\pi i)^{|\alpha|}\xi^\alpha e^{-2i\pi\xi x}\mathcal{F}[v_\eta](\xi).$$

Hence, we have using Eq. (6.26):

$$\begin{aligned}\|\tfrac{\partial^\alpha}{\partial x^\alpha}k_\eta(x,\cdot)\|^2_{\mathcal{H}_\eta} &= \int_{\mathbb{R}^d}(2\pi)^{2|\alpha|}\xi^{2\alpha}\mathcal{F}[v_\eta](\xi)\ d\xi \\ &= (-1)^{|\alpha|}\int_{\mathbb{R}^d}(2i\pi)^{2|\alpha|}\xi^{2\alpha}\mathcal{F}[v_\eta](\xi)\ d\xi \\ &= (-1)^{|\alpha|}\int_{\mathbb{R}^d}\mathcal{F}[\partial_{2\alpha}v_\eta](\xi)\ d\xi = (-1)^{|\alpha|}\partial_{2\alpha}v_\eta(0),\end{aligned}$$

where the last equality comes from the inverse Fourier transform. A simple recursion then shows that $(-1)^{|\alpha|}\partial_{2\alpha}v_\eta(0) = 2^{|\alpha|}\eta^\alpha$, hence the result. The last point of the lemma is simply a consequence of the fact that $\partial_\alpha g(x) = \langle g, \partial_\alpha\phi_\eta(x)\rangle_{\mathcal{H}_\eta}$.

□

6.B .2 Useful Matrices and Linear Operators on the Gaussian RKHS

Recall that we denote with ϕ_η the embedding associated to the RKHS $\mathcal{H}_\eta$ of the Gaussian kernel k_η defined in Example 6.1. In this section, we define operators which will be useful throughout the rest of this section and which we will use in Appendixes 6.E and 6.F . In order to make the dependence in η appear (indeed, η will be a parameter to choose in the the next sections), we will keep it as an index for all of these operators. Recall that for any two vectors u, v in a Hilbert space $\mathcal{H}$, we can define their tensor product $u \otimes v$ which is a linear rank one operator on $\mathcal{H}$ defined by $(u \otimes v)w = \langle v, w\rangle_{\mathcal{H}}\, u$. For the sake of simplicity, we will often write $u \otimes v$ as $uv^\top$, so that the formula $(u \otimes v)w = uv^\top w$ is formally true.

Kernel matrices. We start off by setting the notations for kernel matrices as done by Rudi and Ciliberto (2021). Let $X \in \mathbb{R}^{n\times d}$ and $X' \in \mathbb{R}^{n'\times d}$ be two matrices corresponding to points $x_1, ..., x_n \in \mathbb{R}^d$ and $x'_1, ..., x'_{n'} \in \mathbb{R}^d$. We denote with $K_{X,X',\eta}$ the matrix in $\mathbb{R}^{n\times n'}$ such that

$$\forall 1 \le i \le n,\ \forall 1 \le j \le n',\ [K_{X,X',\eta}]_{ij} = k_\eta(x_i, x'_j). \tag{6.42}$$

If $X = X'$, then we just write $K_{X,\eta}$ and it is positive semi-definite, i.e. $K_{X,\eta} \in \mathbb{S}_+(\mathbb{R}^n)$.

Integration matrices. In this work, we also define, for a given hyper-rectangle $Q = \prod_{k=1}^d [a_k, b_k]$, the following integration matrix $G_{X,X',\eta,Q} \in \mathbb{R}^{n\times n'}$:

$$\begin{aligned}&\forall 1 \le i \le n,\ \forall 1 \le j \le n',\ [G_{X,X',\eta,Q}]_{ij} = \int_Q k_\eta(x - (x_i + x_j)/2)dx \\ &= \prod_{k=1}^d \sqrt{\tfrac{\pi}{4\eta_k}}\left(\operatorname{erf}(\sqrt{\eta_k}(b_k + (x_{ik} + x'_{jk})/2)) - \operatorname{erf}(\sqrt{\eta_k}(a_k + (x_{ik} + x'_{jk})/2))\right),\end{aligned} \tag{6.43}$$

where the erf function is defined in the notations section. Similarly, if $X = X'$, we simply write $G_{X,\eta,Q}$.

This matrix is defined in order to satisfy the following property, which is a direct application of Eq. (6.41): for any $X \in \mathbb{R}^{n\times d}$, any $A \in \mathbb{S}_+^n$ and $\eta \in \mathbb{R}^d_{++}$, the following holds.

$$\int_Q f(x;\ A, X, \eta)\ dx = \sum_{1\leq i,j\leq n} [A \circ K_{X,\eta/2} \circ G_{X,2\eta,Q}]_{ij} = \mathrm{vec}(A \circ K_{X,\eta/2} \circ G_{X,2\eta,Q})^\top \mathbf{1}_{n^2} \quad (6.44)$$

Co-variance operator. Let $\mathcal{X} \subset \mathbb{R}^d$ be a measurable set of $\mathbb{R}^d$ with finite Lebesgue measure $|\mathcal{X}|$. Define the associated co-variance operator:

$$C_\eta \in \mathbb{S}_+(\mathcal{H}_\eta), \qquad C_\eta = \tfrac{1}{|\mathcal{X}|}\int_{\mathcal{X}} \phi_\eta(x) \otimes \phi_\eta(x)\ dx, \qquad C_{\eta,\lambda} = C_\eta + \lambda I. \quad (6.45)$$

Note that C_η is a trace class operator with and that $\mathrm{Tr}\,(C_\eta) = 1$ by linearity of the trace and since $\mathrm{Tr}\,(\phi_\eta(x) \otimes \phi_\eta(x)) = \|\phi_\eta(x)\|^2 = k_\eta(x,x) = 1$. Moreover, since $C_{\eta,\lambda} \succeq \lambda I$, $C_{\eta,\lambda}$ is inverible for any $\lambda > 0$.

Note that we do not make the set $\mathcal{X}$ appear in the notation of the co-variance operator (which can actually be defined with respect to any probability distribution on $\mathbb{R}^d$ and not just $\frac{\mathbf{1}_{\mathcal{X}}\ dx}{|\mathcal{X}|}$). This is because the set $\mathcal{X}$ will usually explicit in the next sections, and in particular equal to the unit hyper-cube $\mathcal{X} = (-1,1)^d$.

Sampling operators. Let $n \in \mathbb{N}$ $(x_1, .., x_n) \in (\mathbb{R}^d)^n$ be points of $\mathbb{R}^d$ which should be seen as samples from a certain distribution. We define the following sampling operators.

$$\widehat{C}_\eta \in \mathbb{S}_+(\mathcal{H}_\eta), \qquad \widehat{C}_\eta = \tfrac{1}{n}\sum_{i=1}^n \phi_\eta(x_i) \otimes \phi_\eta(x_i), \qquad \widehat{C}_{\eta,\lambda} = \widehat{C}_\eta + \lambda I \quad (6.46)$$

$$\widehat{S}_\eta : \mathcal{H}_\eta \to \mathbb{R}^n, \qquad \widehat{S}_\eta(g) = \tfrac{1}{\sqrt{n}}(g(x_i))_{1\leq i\leq n} \quad (6.47)$$

$$\widehat{S}^*_\eta : \mathbb{R}^n \to \mathcal{H}_\eta, \qquad \widehat{S}^*_\eta(a) = \tfrac{1}{\sqrt{n}}\sum_{i=1}^n a_i \phi_\eta(x_i) \quad (6.48)$$

where $\widehat{S}^*_\eta$ and $\widehat{S}_\eta$ are adjoint operators. We will usually use the $\widehat{\bullet}$ notation to denote sampling operators, and imply the underlying $(x_1, ..., x_n)$. These operators will be used in later sections in order to quantify the difference between objects resulting from the sampling of distributions and the "ideal" objects (typically the difference between an empirical risk minimizer and the true expected risk minimizer). For instance, it is clear the $\widehat{C}_\eta$ is an empirical version of C_η, if the x_i are i.i.d. samples from the uniform distribution on $\mathcal{X}$.

Compression operators. Following the notations used by Rudi and Rosasco (2017); Rudi, Camoriano, and Rosasco (2015); Rudi and Ciliberto (2021), a compression operator of size m is an operator $\widetilde{Z}_{\eta,m} : \mathcal{H}_\eta \to \mathbb{R}^m$. We call it a *compression operator* since we use it to project every element of $\mathcal{H}_\eta$ onto the range of the adjoint operator $\widetilde{Z}^*_{\eta,m} : \mathbb{R}^m \to \mathcal{H}_\eta$. This range, which we denote with $\widetilde{\mathcal{H}}_{\eta,m} \subset \mathcal{H}_\eta$, is a subset of dimension at most m. We also denote with $\widetilde{P}_{\eta,m} : \mathcal{H}_\eta \to \mathcal{H}_\eta$

the orthogonal projection onto $\widetilde{\mathcal{H}}_{\eta,m}$, which can also be written $\widetilde{P}_{\eta,m} = \widetilde{Z}^*_{\eta,m}(\widetilde{Z}_{\eta,m}\widetilde{Z}^*_{\eta,m})^\dagger \widetilde{Z}_{\eta,m}$, where $\dagger$ denotes the Moore-Penrose pseudo-inverse.

In this work, we will always use the notation $\widetilde{\bullet}_m$ to denote a compression operator, and the index m to make the size of the compression explicit.

In this work, we take a specific form of compression operator as in appendix C by Rudi and Ciliberto (2021). Indeed, let $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ be a data point matrix representing vectors $\widetilde{x}_1, ..., \widetilde{x}_m \in \mathbb{R}^d$. The compression operator associated to $\widetilde{X}_m$ is the following :

$$\widetilde{Z}_{\eta,m} : \mathcal{H}_\eta \to \mathbb{R}^m, \qquad \widetilde{Z}_{\eta,m}(g) = (g(\widetilde{x}_j))_{1\le j\le m} = (g^\top \phi_\eta(\widetilde{x}_j))_{1\le j\le m}. \tag{6.49}$$

Note that $\widetilde{Z}_{\eta,m}\widetilde{Z}^*_{\eta,m} = K_{\widetilde{X}_m,\eta}$ and hence the projection operator can be written $\widetilde{P}_{\eta,m} = \widetilde{Z}^*_{\eta,m}K^\dagger_{\widetilde{X}_m,\eta}\widetilde{Z}_{\eta,m}$ and that it is simply the projection onto span $\phi_\eta(\widetilde{x}_i)_{1\le i\le m}$. This compression is also chosen to satisfy the two following properties :

- if $h \in \mathcal{H}_\eta$, then $\widetilde{P}_{\eta,m}h$ represents a function of the form $g(\bullet;\ a, \widetilde{X}_m, \eta)$ where $a = K^\dagger_{\widetilde{X}_m,\eta}\widetilde{Z}_{\eta,m}h$ (see Eq. (6.2) for the definition of the Gaussian linear model $g(x;\ a, \widetilde{X}_m, \eta)$);
- if $M \in \mathbb{S}_+(\mathcal{H}_\eta)$, then for any $x \in \mathbb{R}^d$, it holds

$$f(x;\ \widetilde{P}_{\eta,m}M\widetilde{P}_{\eta,m}, \phi_\eta) = f(x;\ A, \widetilde{X}_m, \eta), \qquad A = K^\dagger_{\widetilde{X}_m,\eta}\widetilde{Z}_{\eta,m}M\widetilde{Z}^*_{\eta,m}K^\dagger_{\widetilde{X}_m,\eta}, \tag{6.50}$$

meaning that compressed linear (resp. PSD) models can be compressed as a sum of m (resp. m^2) Gaussian kernel functions. We quantify the effect of this compression in Lemma 6.3 and Theorem 6.5.

6.B .3 Approximation properties of the Gaussian kernel

This section aims in quantifying the approximation power of the Gaussian RKHS. We start in proposition 6.7 by quantifying the approximation power of the Gaussian RKHS by finding an ε approximation of a regular function with controlled norm. We then quantify the "size" of a compression for the Gaussian RKHS in Lemma 6.3, which essentially bounds the possible variations of a function in $\mathcal{H}_\eta$ if it is equal to zero on the compression points $\widetilde{X}_m$.

Approximation of a Sobolev function. This paragraph remolds results in the proof of Theorem D.4 by Rudi and Ciliberto (2021) whose goal is to approximate any function $g \in W_2^\beta(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ by a function in $\mathcal{H}_\eta$.

Proposition 6.7 (Approximation of $W_2^\beta(\mathbb{R}^d)\cap L^\infty(\mathbb{R}^d)$ in $\mathcal{H}_\eta$)**.** *Let g be a function in $W_2^\beta(\mathbb{R}^d)\cap L^\infty(\mathbb{R}^d)$ and $\eta \in \mathbb{R}^d_{++}$. Denote with $|\eta|$ the product $|\eta| := \prod_{i=1}^d \eta_i$ and $\eta_0 = \min_{1\le i\le d}\eta_i$. For any $\varepsilon \in (0,1]$, there exists $\theta \in \mathcal{H}_\eta$ such that*

$$\begin{cases} \|\theta - g\|_{L^2(\mathbb{R}^d)} \le \varepsilon \|g\|_{W_2^\beta(\mathbb{R}^d)} \\ \|\theta - g\|_{L^\infty(\mathbb{R}^d)} \le C_1\ \varepsilon^{1-\nu}\|g\|_\bullet \end{cases}, \quad \|\theta\|_{\mathcal{H}_\eta} \le C_2\ \|g\|_{W_2^\beta(\mathbb{R}^d)}\ |\eta|^{1/4}\left(1 + \varepsilon \exp\left(\frac{50}{\eta_0 \varepsilon^{2/\beta}}\right)\right), \tag{6.51}$$

where $\|g\|_\bullet = \|g\|_{L^\infty(\mathbb{R}^d)}$ if $\beta \le d/2$ and $\|g\|_\bullet = \|g\|_{W_2^\beta(\mathbb{R}^d)}$ if $\beta > d/2$, $\nu = \min(1, d/(2\beta))$ and C_1, C_2 are constants which depend only on d, β.

Proof. Recalling the notations from the proof of Theorem D.4. by Rudi and Ciliberto (2021), let $g_t := t^{-d} g_1(x/t)$ where g_1 is defined as g in equation (D.2) by Rudi and Ciliberto (2021). The following result hold.

- By step 1 of the proof of Theorem D.4, $\|g - g \star g_t\|_{L^2(\mathbb{R}^d)} \leq (2t)^\beta \ \|g\|_{W_2^\beta(\mathbb{R}^d)}$.
- By step 2 and the beginning of step 3 of the proof of Theorem D.4,
$$\|g \star g_t\|_{\mathcal{H}_\eta} \leq 2^\beta \pi^{-d/4} |\eta|^{1/4} (1 + (t/3)^\beta \exp(\tfrac{50}{\eta_0 t^2})) \|g\|_{W_2^\beta(\mathbb{R}^d)}.$$
- As in step 5 of the proof of Theorem D.4 and in particular the Young convolution inequality combined with the fact that $\|g_1\|_{L^1(\mathbb{R}^d)}$ is finite, $\|g \star g_t\|_{L^\infty(\mathbb{R}^d)} \leq \|g_1\|_{L^1(\mathbb{R}^d)} \ \|g\|_{L^\infty(\mathbb{R}^d)}$ which in turn implies $\|g - g \star g_t\| \leq (1 + \|g_1\|_{L^1(\mathbb{R}^d)}) \ \|g\|_{L^\infty(\mathbb{R}^d)}$.

Replacing t by $\varepsilon^{1/\beta}/2$, we get all the bounds except the bound for the L^∞ norm in the case where $\beta > d/2$. In that case, we proceed in the following way. Recycling results and notations from the proof of Theorem D.4 by Rudi and Ciliberto (2021), denoting with $\mathcal{F}$ the Fourier transform defined in proposition 6.3, it holds

$$\begin{aligned}
\|f - f \star g_t\|_{L^\infty(\mathbb{R}^d)} &\leq \|\mathcal{F}(f - f \star g_t)\|_{L^1(\mathbb{R}^d)} \text{ proposition 6.3} \\
&= \|\mathcal{F}(f)(1 - \mathcal{F}(g_t))\|_{L^1(\mathbb{R}^d)} \\
&\leq \|(1 + \|\omega\|^2)^{\beta/2} \mathcal{F}(f)\|_{L^2(\mathbb{R}^d)} \ \|(1 + \|\omega\|^2)^{-\beta/2} \ \mathcal{F}(1 - g_t)\|_{L^2(\mathbb{R}^d)} \\
&\leq 2^\beta \left(\int_{\|\omega\| > 1/t} (1 + \|\omega\|^2)^{-\beta} \ d\omega \right)^{1/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \text{ Eq. (6.29)} \\
&= 2^\beta \left(S_d \int_{r > 1/t} r^{d-1} \ (1 + r^2)^{-\beta} \ dr \right)^{1/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \text{ (spherical coord.)} \\
&\leq 5^{\beta/2} S_d^{1/2} \left(\int_{r > 1/t} r^{d-1-2\beta} \ dr \right)^{1/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \ (t < 1/2) \\
&= 5^{\beta/2} \tfrac{1}{\sqrt{2\beta - d}} S_d^{1/2} t^{\beta - d/2} \|f\|_{W_2^\beta(\mathbb{R}^d)} \\
&= 5^{\beta/2} 2^{d/2 - \beta} S_d^{1/2} \tfrac{1}{\sqrt{2\beta - d}} \varepsilon^{1 - d/(2\beta)} \|f\|_{W_2^\beta(\mathbb{R}^d)},
\end{aligned}$$

where S_d is the surface area of the $d-1$ dimensional hyper-sphere. □

A bound on the performance of compression when using uniform samples from $\mathcal{X} = (-1, 1)^d$. In this paragraph, we study the effect of performing compression with a compression operator of the form $\widetilde{Z}_{\eta,m}$ (see Eq. (6.49)) where the associated $\widetilde{X}_m$ are i.i.d. samples from the uniform measure on the unit hyper-cube $\mathcal{X} = (-1, 1)^d$.

Lemma 6.3. *Let $m \in \mathbb{N}$, $\delta \in (0, 1]$, $\tau \geq 1$ and $\rho \in (0, 1]$. Let $\eta = \tau \mathbf{1}_d \in \mathbb{R}^d_{++}$. Let $\widetilde{X}_m \in \mathbb{R}^{m \times d}$ be a data matrix corresponding to vectors $\widetilde{x}_1, ..., \widetilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X} = (-1, 1)^d$ and let $\widetilde{P}_{\eta,m}$ be the associated projection operator in $\mathcal{H}_\eta$. With probability at least $1 - \delta$, if $m \geq C_1 \tau^{d/2} (\log \frac{C_2}{\rho})^d \left(\log \frac{C_3}{\delta} + \log \tau + \log \log \frac{C_2}{\rho} \right)$, then it holds :*

$$\sup_{x \in \mathcal{X}} \|(I - \widetilde{P}_{\eta,m}) \phi_\eta(x)\| \leq \rho, \tag{6.52}$$

where C_1, C_2, C_3 are constants which depend only on the dimension d and not on τ, m, δ, ρ.

Proof. Let h denote the fill distance with respect to $\widetilde{X}_m$, i.e.

$$h = \max_{x\in[-1,1]^d} \min_{1\leq j\leq m} \|x - \widetilde{x}_j\| \tag{6.53}$$

Using Lemma 12 p.19 by Vacher, Muzellec, Rudi, Bach, and Vialard (2021), we there exists two constants C_1, C_2 depending only on d such that $h \leq \left(C_1 m^{-1}(\log(C_2 m/\delta)\right)^{1/d}$.

Applying Theorem C.3 by Rudi and Ciliberto (2021) in the case where $\mathcal{X} = (-1,1)^d$, $\eta = \tau \mathbf{1}_d$, there exists constants C_3, C_4, C_5 depending only on the dimension d such that when $h \leq \tau^{-1/2} C_3^{-1}$, the following holds :

$$\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leq C_4 e^{-\frac{C_5}{\tau^{1/2}h} \log \frac{C_5}{\tau^{1/2}h}} \tag{6.54}$$

Now note that taking $C_6 = \max(C_3^{-1}, eC_5)$ and $C_7 = \max(e, C_4)$, as soon as $h \leq C_6 \tau^{-1/2}/\log \frac{C_7}{\rho}$, it holds a) $h \leq \tau^{-1/2} C_3^{-1}$, b) $\frac{C_5}{\tau^{1/2}h} \geq e$ and thus $\log \frac{C_5}{\tau^{1/2}h} \geq 1$, and hence c) $\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leq \rho$ using Eq. (6.54). Using the bound on h, this is satisfied as soon as

$$m \geq C_8 \tau^{d/2} \left(\log \frac{C_7}{\rho}\right)^d \log(C_2 m/\delta),$$

where $C_8 = \max(C_1/C_6^d, e)$. Using the fact that $C_2, C_8 \geq e$, and using the reasoning in the proof of Theorem C.5 by Rudi and Ciliberto (2021), in equation (C.44), a sufficient condition is the following :

$$m \geq 2C_8 \tau^{d/2} \left(\log \frac{C_7}{\rho}\right)^d \left(\log(2C_2C_8/\delta) + \frac{d}{2}\log\tau + d\log\log\frac{C_7}{\rho}\right). \tag{6.55}$$

The result in the theorem is obtained by taking $C_1 \leftarrow 2C_8 d,\ C_2 \leftarrow C_7,\ C_3 \leftarrow 2C_2C_8$. □

6.C Properties of Gaussian PSD models

In this section, we detail some of the properties specific to Gaussian PSD models.

6.C .1 Bounds on the support and the derivatives

In this section, we present results which can be used to bound the tail and derivatives of a Gaussian PSD model. These bounds can be used both for theoretical purposes (see Appendixes 6.E and 6.F) and to perform adaptive bounds in an algorithm (see Sec. 6.3)

Lemma 6.4 (tail bound)**.** *Let $\delta = (\delta_k) \in \mathbb{R}^d$, $\eta \in \mathbb{R}^d_{++}$, $X \in \mathbb{R}^{n\times d}$ and $A \in \mathbb{S}_+(\mathbb{R}^n)$. Let $f(x; A, X, \eta)$ be the associated PSD model. Define $\overline{x}$, $\underline{x}$:*

$$\forall 1 \leq k \leq d,\ \overline{x}_k = \max_{1\leq i\leq n} X_{ik},\ \underline{x}_k = \min_{1\leq i\leq n} X_{ik}.$$

Let $Q_\delta = Q(\underline{x} - \delta, \overline{x} + \delta)$. Then the following bound holds:

$$\int_{\mathbb{R}^d\setminus Q_\delta} |f(x; A, X, \eta)|dx \leq \left(2\pi^{d/2} \det(\mathrm{diag}(2\eta))^{-1/2} \sum_{k=1}^d e^{-2\eta_k\delta_k^2}\right) \sum_{i,j} [A \circ K_{X,\eta/2}]_{ij} \tag{6.56}$$

Proof. Start by recalling the following simple Chernoff bound:

$$\forall x > 0,\ \int_x^{+\infty} e^{-t^2}\, dt \leq \sqrt{\pi} e^{-x^2} \tag{6.57}$$

Indeed, take $\lambda > 0$. Since $e^{-2\lambda x}\, e^{2\lambda t} \leq \mathbf{1}_{t>x}$, it holds

$$\int_x^{+\infty} e^{-t^2}\, dt \leq e^{-2\lambda x} e^{\lambda^2} \int_{-\infty}^{+\infty} e^{-(t-\lambda)^2}\, dt \leq \sqrt{\pi} e^{-x^2} e^{(\lambda - x)^2}.$$

Hence, taking $\lambda = x$, we get the bound. Then we perform the following bound.

$$\begin{aligned}
\int_{\mathbb{R}^d \setminus Q(-\delta,\delta)} k_\eta(x,0)dx &= \frac{1}{\prod_{k=1}^d \eta_k^{1/2}} \int_{\mathbb{R}^d \setminus Q(-\delta\sqrt{\eta},\delta\sqrt{\eta})} k_1(x,0)dx \\
&\leq \frac{1}{\prod_{k=1}^d \eta_k^{1/2}} \sum_{k=1}^d \left(\pi^{(d-1)/2} 2 \int_{\delta_k\sqrt{\eta_k}}^{\infty} e^{-t^2}\, dt \right) \\
&\leq 2\pi^{d/2} \det(\mathrm{diag}(\eta))^{-1/2} \sum_{k=1}^d e^{-\delta_k^2 \eta_k},
\end{aligned}$$

where we go from the first to the second line by noting that

$$\mathbb{R}^d \setminus Q(-\delta,\delta) \subset \cup_{k=1}^d \mathbb{R} \times .... \times \mathbb{R} \setminus [-\delta_k, \delta_k] \times ... \times \mathbb{R},$$

and the last inequality comes from a Eq. (6.57).

The result immediately follows from Eq. (6.41) as well as the fact that Q_δ contains $(x_i + x_j)/2 + Q(-\delta,\delta)$ for all $1 \leq i,j \leq n$.

□

Lemma 6.5 (derivative bound for general PSD model)**.** *Let $\eta \in \mathbb{R}^d_{++}$, $M \in \mathbb{S}_+(\mathcal{H}_\eta)$ $X \in \mathbb{R}^{n\times d}$ and $A \in \mathbb{S}^n_+$. The following bounds hold :*

$$\sup_{x\in\mathbb{R}^d} |\partial_\alpha f(x;\ M, \phi_\eta)| \leq 2^{3|\alpha|/2}\ \eta^{\alpha/2}\ \|M\| \tag{6.58}$$

$$\sup_{x\in\mathbb{R}^d} |\partial_\alpha f(x;\ A, X, \eta)| \leq 2^{3|\alpha|/2}\ \eta^{\alpha/2}\ \|K_{X,\eta}^{1/2} A K_{X,\eta}^{1/2}\| \tag{6.59}$$

Proof. By derivation of a bi-linear form, we get

$$\partial_\alpha f(x;\ M, \phi_\eta) = \sum_{\beta\leq\alpha} \binom{\alpha}{\beta} \langle \partial_\beta \phi_\eta(x), M \partial_{\alpha-\beta}\phi_\eta(x) \rangle_{\mathcal{H}_\eta}$$

Hence, using Lemma 6.2, we get, for any $x \in \mathbb{R}^d$,

$$|\partial_\alpha f(x;\ M, \phi_\eta)| \leq \|M\| \sum_{\beta\leq\alpha} \binom{\alpha}{\beta} 2^{|\beta|/2} \eta^{\beta/2} 2^{|\alpha-\beta|/2} \eta^{(\alpha-\beta)/2} = 2^{3|\alpha|/2} \eta^{\alpha/2} \|M\|. \tag{6.60}$$

In particular, since $f(x;\ A, X, \eta) = f(x;\ M_A, \phi_\eta)$ with $M_A = Z^* A Z$ for $Z : h \in \mathcal{H}_\eta \mapsto h(x_i)_{1\leq i\leq n}$, and since $ZZ^* = K_{X,\eta}$, it holds

$$\|M_A\| = \|Z^* A Z\| = \|A^{1/2} Z Z^* A^2\| = \|A^{1/2} K_{X,\eta} A^{1/2}\| = \|K_{X,\eta}^{1/2} A K_{X,\eta}^{1/2}\|,$$

and hence the second equation of the lemma. □

6.C .2 Compression as a Gaussian PSD model

In this section, we restate Theorem C.4 by Rudi and Ciliberto (2021) on the compression of a PSD model of the form $f(x;\ M, \phi_\eta)$ into a Gaussian PSD model.

Let $\eta \in \mathbb{R}^d_{++}$, $M \in \mathbb{S}_+(\mathcal{H}_\eta)$. Given a matrix $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ representing vectors $\widetilde{x}_1, ..., \widetilde{x}_m \in \mathbb{R}^d$, and the associated projection operator $\widetilde{P}_{\eta,m}$ (for more details, see Appendix 6.B .2), one can compress the PSD model $f(\bullet;\ M, \phi_\eta)$ into $f(\bullet;\ \widetilde{P}_{\eta,m} M \widetilde{P}_{\eta,m})$ which is also a Gaussian PSD model of the form $f(\bullet;\ A, \widetilde{X}_m, \eta)$ (A is defined in Eq. (6.50)). The quality of the compression is given by the following theorem.

Theorem 6.5 ((Rudi and Ciliberto, 2021, Theorem C.4))**.** *Using the previous notations, the compressed model associated to $\widetilde{P}_{\eta,m} M \widetilde{P}_{\eta,m}$ of M onto $\widetilde{X}_m$ has a distance to the original PSD model associated to M bounded, for any $x \in \mathcal{X}$, by*

$$|f(x;\ M, \phi_\eta) - f(x;\ \widetilde{P}_{\eta,m} M \widetilde{P}_{\eta,m}, \phi_\eta)| \leq \sqrt{f(x;\ M, \phi_\eta)}\ \|M\|^{1/2} \sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| + \|M\| \sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\|^2. \tag{6.61}$$

We therefore see that the quality of the compression depends mainly on the quantity

$$\sup_{x\in\mathcal{X}} \|(I - \widetilde{P})\phi_\eta(x)\|,$$

which can be bounded using Eq. (6.52) in Lemma 6.3.

6.C .3 Approximation properties of Gaussian PSD model

Define, for any measurable $\Omega \subset \mathbb{R}^d$, and any $f : \Omega \to \mathbb{R}$, the following function (set to $+\infty$ if the set is empty).

$$\|f\|_{\mathtt{sos},\Omega,\beta} = \inf \left\{ \sum_{i=1}^{Q} \max(\|f_j\|_{L^\infty(\Omega)}, \|f_j\|_{W_2^\beta(\Omega)})^2 \mid f = \sum_{j=1}^{Q} f_j^2,\ Q \in [0, +\infty] \right\} \tag{6.62}$$

Here, we recall Theorem D.4 by Rudi and Ciliberto (2021), refined in a small way to have more control over the dependence in the f_j.

Theorem 6.6 ((Rudi and Ciliberto, 2021, Theorem D.4))**.** *Let $\tau \geq 1$ and $\varepsilon \in (0, 1]$ and f such that $\|f\|_{sos,\mathbb{R}^d,\beta} < \infty$. Let $\eta = \tau \mathbf{1}_d$. There exists $M_{\tau,\epsilon} \in \mathbb{S}_+(\mathcal{H}_\eta)$ such that $f_{\tau,\epsilon} := f(\bullet;\ M_{\tau,\epsilon}, \phi_\eta)$ is ε close to p in L^2 norm and has controlled trace norm:*

$$\begin{aligned} &\|f_{\tau,\epsilon} - f\|_{L^2(\mathbb{R}^d)} \leq C_1\ \ \|f\|_{sos,\mathbb{R}^d,\beta}\ \ \varepsilon, \\ &\mathrm{Tr}\,(M_{\tau,\epsilon}) \leq C_2\ \ \|f\|_{sos,\mathbb{R}^d,\beta}\ \ \tau^{d/2}(1 + \varepsilon^2 \exp(C_3\ \varepsilon^{-2/\beta}/\tau)), \end{aligned} \tag{6.63}$$

where the constants C_1, C_2, C_3 depend only on β, d.

Proof. Let $\delta > 0$ and take $Q_\delta \in [0, +\infty]$ as well as $f_{\delta,j}$ such that $f = \sum_{j=1}^{Q_\delta} f_{\delta,j}^2$ point-wise and

$$\sum_{i=1}^{Q} \|f_{\delta,j}\|_{W_2^\beta(\mathbb{R}^d)} \max(\|f_{\delta,j}\|_{L^\infty(\mathbb{R}^d)}, \|f_{\delta,j}\|_{W_2^\beta(\mathbb{R}^d)}) \leq \|f\|_{\mathtt{sos},\beta}\,.$$

Now using exactly the same reasoning than in the proof of Theorem D.4 by Rudi and Ciliberto (2021) but setting simply $t = \varepsilon^{1/\beta}$, it holds the existence of $M_{\delta,\tau,\varepsilon}$ and C_1, C_2, C_3 depending only on β, d such that

$$\|f_{\delta,\tau,\varepsilon} - f\|_{L^2(\mathbb{R}^d)} \leq C_1 \left(\|f\|_{\mathtt{sos},\mathbb{R}^d,\beta} + \delta\right) \varepsilon,$$
$$\mathrm{Tr}\left(M_{\delta,\tau,\varepsilon}\right) \leq C_2 \left(\|f\|_{\mathtt{sos},\mathbb{R}^d,\beta} + \delta\right) \tau^{d/2}(1 + \varepsilon^2 \exp(C_3\, \varepsilon^{-2/\beta}/\tau)).$$

Note that in the proof, $M_{\delta,\tau,\varepsilon}$ is well defined since its trace norm is bounded (normal convergence). Now if $\|f\|_{\mathtt{sos},\mathbb{R}^d,\beta} = 0$, then $f = 0$ and there is nothing to prove. If not, then taking $\delta = \|f\|_{\mathtt{sos},\mathbb{R}^d,\beta}$, the theorem holds. □

6.D The sampling algorithm

In this section, we formally prove that algorithm 2 converges, as in Theorem 6.1, as well as the different results of Sec. 6.3 . We start by introducing some notations around dyadic decomposition of hyper-rectangles. We then introduce a well founded order relation, which we will then use to both construct the random variables we study, justify the convergence of the algorithm and prove its correctness.

Recall we are given a density (up to a scaling factor) $f(x)$ and that we denote with $I(Q)$ the quantity $\int_Q f(x)dx$ on any hyper-rectangle Q.

6.D .1 Dyadic decompositions and convergence of algorithm 2

Dyadic sub-rectangles Let $Q = \prod_{k=1}^d [a_k, b_k[$ be a hyper-rectangle where $a \leq b$ and let $\delta = b - a$. Let $q \in \mathbb{N}^d$. We define $\mathcal{D}_{Q,q}$ to be the set of dyadic sub-rectangles of Q whose k-th size is cut in half q_k times, i.e.

$$\mathcal{D}_{Q,q} = \left\{ \prod_{k=1}^d [a_k + \delta_k \tfrac{s}{2^{q_k}}, a_k + \delta_k \tfrac{s+1}{2^{q_k}}[\;:\; s \in \prod_{k=1}^d [\![0, 2^{q_k} - 1]\!] \right\}.$$

We denote with $\mathcal{D}_Q$ the set of dyadic sub-rectangles of Q, i.e. the union $\bigcup_{q\in\mathbb{N}^d} \mathcal{D}_{Q,q}$.

Moreover, if $q_k^\rho = \max(0, \lceil \log_2 \frac{\delta_k}{\rho} \rceil)$, we also define $\mathcal{D}_{Q,\varepsilon} := \mathcal{D}_{Q,q^\rho}$ to be the set of dyadic sub-rectangles whose size is just below ρ.

Well founded order relation on hyper-rectangles For all $\rho > 0$, we define the following strict order relation. We say that $Q \prec_\rho Q'$ if the following conditions hold :

1. $Q \in \mathcal{D}_{Q'}$;
2. There exists $k \in [\![1, d]\!]$ such that $\delta'_k > \rho$ and $\delta_k < \delta'_k$.

This relation is obviously transitive. Moreover, if $s(Q) := \sum_{k=1}^d \delta_k(Q)$, it is easy to show that $Q \prec_\rho Q'$ implies $s(Q) \leq s(Q') - \rho/2$. Since $s \geq 0$, this in turn shows that any strictly decreasing sequence for $\prec_\rho$ is finite, and that $Q \prec_\rho Q'$ and $Q' \prec_\rho Q$ are incompatible.

We are now ready to define the random variable $\boldsymbol{Y}_{\rho,Q,n}$ by structural induction on Q for any $n \in \mathbb{N}$. Recall that for $\Omega \subset \mathbb{R}^d$, we denote with $\mathcal{U}_\Omega$ the uniform law on Ω.

Definition of the random variable $\boldsymbol{Y}_{\rho,Q,n}$ and relation to the algorithm We now define a random variable from whose distribution we sample when SamplerRec in 2 is applied.

- If $\delta(Q) \leq \rho$, then for any $n \in \mathbb{N}$, $\boldsymbol{Y}_{\rho,Q,n} \sim \mathcal{U}_Q^{\otimes n}$
- Else, let $n \in \mathbb{N}$ and $k_Q = \min \arg\max_{1 \leq k \leq d} \delta_k(Q)$ be the smallest index amongst the largest sides of Q. Define Q_1 and Q_2 to be the two hyper-rectangles obtained by cutting Q in half along the direction k_Q. Since $\delta_{k_Q} > \rho$ and Q_1, Q_2 are dyadic sub-rectangles of Q, we have $Q_1, Q_2 \prec_\rho Q$.

  By structural induction, we give ourselves a probability space on which we take we take the following random variables to be independent : $\boldsymbol{Y}_{1,m} \sim \boldsymbol{Y}_{\rho,Q_1,m}$, $\boldsymbol{Y}_{2,m} \sim \boldsymbol{Y}_{\rho,Q_2,m}$ for $0 \leq m \leq n$ and $M \sim \mathcal{B}(n, I(Q_1)/I(Q))$ and define

$$\boldsymbol{Y}_{\rho,Q,n} = (\boldsymbol{Y}_{\rho,Q_1,M}, \boldsymbol{Y}_{\rho,Q_2,n-M}) := \sum_{m=0}^{n} \mathbf{1}_{M=m}(\boldsymbol{Y}_{1,m}, \boldsymbol{Y}_{2,n-m}). \tag{6.64}$$

Lemma 6.6 (Termination of the algorithm and first result)**.** *For any inputs $\rho > 0$, hyper-rectangle Q and $n \in \mathbb{N}$,* SamplerRec *in algorithm 2 terminates and returns a sample $(y_1, ..., y_n)$ from $\boldsymbol{Y}_{\rho,Q,n}$.*

Proof. This is a simple application of structural induction on the well-founded order $\prec_\rho$ for the termination and then again for the fact that a sample $(y_1, ..., y_n)$ from $\boldsymbol{Y}_{\rho,Q,n}$, using the definition of $\boldsymbol{Y}$ above. □

6.D .2 Proof of Theorem 6.1

In this section, we prove Theorem 6.1. To do so, we define a random variable $X_{\rho,Q}$, compute its density with respect to the Lebesgue measure on the hyper-rectangle Q (and show it is our target density), and show that $\boldsymbol{Y} = X$ up to some random shuffling.

Definition of the variable $X_{\rho,Q}$ Recall the definition of $\mathcal{D}_{Q,\rho}$ from Appendix 6.D .1. We define a random variable $R_{\rho,Q}$ on $\mathcal{D}_{Q,\rho}$ whose law is defined $P(R_{\rho,Q} = r) = I(r)/I(Q)$. Recall that for any $r \subset \mathbb{R}^d$, we denote with $\mathcal{U}_r$ the uniform law on r. We give ourselves a measure space on which there exists a family of random variables $U_r \sim \mathcal{U}_r$ for $r \in \mathcal{D}_{Q,\rho}$ and $R \sim R_{\rho,Q}$ which are all independent and define

$$X_{\rho,Q} = U_R := \sum_{r \in \mathcal{D}_{Q,\rho}} \mathbf{1}_{R=r} U_r \tag{6.65}$$

Lemma 6.7 (density of $X_{\rho,Q}$)**.** *The density of $X_{\rho,Q}$ with respect to the Lebesgue measure is given by Eq. (6.7), i.e.*

$$\forall x \in Q,\ p_{X_{\rho,Q}}(x) = \sum_{r \in \mathcal{D}_{Q,\rho}} \frac{I(r)}{I(Q)} \frac{\mathbf{1}_r(x)}{|r|}. \tag{6.66}$$

Proof. For any measurable function f, it holds

$$\begin{aligned}\mathbb{E}[f(X_{\rho,Q})] &= \sum_{r\in\mathcal{D}_{Q,\rho}} \mathbb{E}[\mathbf{1}_{R=r} f(U_r)] \\ &= \sum_{r\in\mathcal{D}_{Q,\rho}} P(R=r)\mathbb{E}[f(U_r)] \\ &= \sum_{r\in\mathcal{D}_{Q,\rho}} \frac{I(r)}{I(Q)} \int_{\mathbb{R}^d} f(x) \frac{\mathbf{1}_r(x)}{|r|} dx \\ &= \int_{\mathbb{R}^d} f(x) \left(\sum_{r\in\mathcal{D}_{Q,\rho}} \frac{I(r)}{I(Q)} \frac{\mathbf{1}_r(x)}{|r|} \right) dx\end{aligned}$$

□

Action of a permutation and decomposition Let $n \in \mathbb{N}$. For any permutation $\tau \in \mathfrak{S}_n$ and vector $v \in \mathbb{R}^n$, denote with $\tau \star v$ the permuted vector $(v_{\tau^{-1}(i)})_{1\leq i\leq n}$.

We now define a decomposition of a permutation of n variables as i) a permutation of the first m variables and a permutation of the last $n-m$ variables ii) followed by a rearrangement of these variables.

Given $I \subset [\![1,n]\!]$ of size m, define τ_I as the unique permutation satisfying $I = \{\tau_I(1),...,\tau_I(m)\}$, $I^c = \{\tau_I(m+1),...,\tau_I(n)\}$ and $\tau_I(1) < ... < \tau_I(m)$ and $\tau_I(m+1) < ... < \tau_I(n)$. For any $m \in [\![0,n]\!]$, if $\mathcal{P}_m(n)$ denotes the set of subsets of $\{1,...,n\}$ of size m, the map from $\mathcal{P}_m(n) \times \mathfrak{S}_m \times \mathfrak{S}_{n-m}$ to $\mathfrak{S}_n$ defined as

$$(I, \sigma_m, \sigma_{n-m} \mapsto \left(i \mapsto \begin{cases} \tau_I(\sigma_m(i)) & \text{if } i \leq m \\ \tau_I(m + \sigma_{n-m}(i-m)) & \text{otherwise} \end{cases} \right) \tag{6.67}$$

is a bijection.

Lemma 6.8. *Let $\rho > 0$. Let $n \in \mathbb{N}$, Q be a hyper-rectangle of $\mathbb{R}^d$. Let σ be a random permutation independent of $\boldsymbol{Y}_{\rho,Q,n}$. Then $(\boldsymbol{Y}^{\sigma(i)}_{\rho,Q,n})_{1\leq i\leq n} \sim X^{\otimes n}_{\rho,Q}$.*

Proof. Once again, we prove this by structural induction. Fix $\rho > 0$. We will prove the following property by structural induction on the set of hyper-rectangles Q equipped with the strict order relation $\prec_\rho$:

For any $n \in \mathbb{N}$, if σ is a random permutation (i.e. distributed uniformly amongst all permutations in $\mathfrak{S}_n$), $\boldsymbol{Y}_{Q,n} \sim \boldsymbol{Y}_{\rho,Q,n}$ and both random variables are independent, then $(\boldsymbol{Y}^{\sigma(i)}_{Q,n})_{1\leq i\leq n} \sim X^{\otimes n}_{Q,\rho}$.

1) If $\delta(Q) \leq \rho$.

On the one hand, by definition of $\boldsymbol{Y}_{\rho,Q,n}$, it holds that for any $n \in \mathbb{N}$, $\boldsymbol{Y}_{\rho,Q,n} \sim \mathcal{U}_Q^{\otimes n}$ and hence $\boldsymbol{Y}_{Q,n} \sim \mathcal{U}_Q^{\otimes n}$. By invariance of the product measure by permutation, it also holds that $(\boldsymbol{Y}^{\sigma(i)}_{Q,n})_{1\leq i\leq n} \sim \mathcal{U}_Q^{\otimes n}$.

On the other hand, since $\delta(Q) \leq \rho$, it is easy to see that $q^\rho = 0$ and hence $\mathcal{D}_{Q,\rho} = \{Q\}$. Hence, by definition of $X_{\rho,Q}$ in Eq. (6.65), R is deterministic and hence $X_{\rho,Q} = U_Q \sim \mathcal{U}_Q$.

Putting things together, this yields $(\boldsymbol{Y}^{\sigma(i)}_{Q,n})_{1\leq i\leq n} \sim X^{\otimes n}_{\rho,Q}$.

2) Assume $\delta(Q) > \rho$ and take $n \in \mathbb{N}$. By definition of $\boldsymbol{Y}_{\rho,Q,n}$ in Eq. (6.64), and since our property only concerns a convergence in law, we can assume that $\boldsymbol{Y}_{Q,n}$ is of the form

$$\boldsymbol{Y}_{Q,n} = \sum_{m=0}^{n} \mathbf{1}_{M=m}(\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}),$$

where $\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,m}$ and M are independent and independent of σ, $\boldsymbol{Y}_{Q_1,m} \sim \boldsymbol{Y}_{\rho,Q_1,m}$, $\boldsymbol{Y}_{Q_2,m} \sim \boldsymbol{Y}_{\rho,Q_2,m}$ for $0 \leq m \leq n$ and $M \sim \mathcal{B}(n, I(Q_1)/I(Q))$, and Q_1, Q_2 are defined just before Eq. (6.64). It is easy to see that since $Q_1 \sqcup Q_2 = Q$ and $Q_1, Q_2 \prec_\rho Q$, it holds $\mathcal{D}_{Q,\rho} = \mathcal{D}_{Q_1,\rho} \sqcup \mathcal{D}_{Q_2,\rho}$ where $\sqcup$ symbolises a disjoint union.

Fix a measurable function f. Using the independence of M from the other variables and the fact that it is discrete, it holds

$$\mathbb{E}[f(\sigma \star \boldsymbol{Y}_{Q,n})] = \sum_{m=0}^{n} P(M = m)\mathbb{E}[f(\sigma \star (\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}))].$$

Now note that using our bijection Eq. (6.67), it holds

$$\begin{aligned}
&\mathbb{E}_{\sigma,\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\sigma \star (\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}))\right] \\
&= \frac{1}{n!}\sum_{\tau \in \mathfrak{S}_n} \mathbb{E}_{\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\tau \star (\boldsymbol{Y}_{Q_1,m}, \boldsymbol{Y}_{Q_2,n-m}))\right] \\
&= \frac{1}{n!}\sum_{\substack{I \subset [\![1,n]\!] \\ |I|=m}} \sum_{\sigma_1 \in \mathfrak{S}_m} \sum_{\sigma_2 \in \mathfrak{S}_{n-m}} \mathbb{E}_{\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\tau_I \star (\sigma_1 \star \boldsymbol{Y}_{Q_1,m}, \sigma_2 \star \boldsymbol{Y}_{Q_2,n-m}))\right] \\
&= \frac{1}{\binom{n}{m}}\sum_{\substack{I \subset [\![1,n]\!] \\ |I|=m}} \mathbb{E}_{\sigma_1,\sigma_2,\mathbf{Y}_{Q_1,m},\mathbf{Y}_{Q_2,m}}\left[f(\tau_I \star (\sigma_1 \star \boldsymbol{Y}_{Q_1,m}, \sigma_2 \star \boldsymbol{Y}_{Q_2,n-m}))\right]
\end{aligned}$$

Now note that by induction, $\sigma_1 \star \boldsymbol{Y}_{Q_1,m} \sim X_{\rho,Q_1}^{\otimes m}$ and $\sigma_2 \star \boldsymbol{Y}_{Q_2,n-m} \sim X_{\rho,Q_2}^{\otimes(n-m)}$.

Let $X_1^1, ..., X_1^n \sim X_{\rho,Q_1}$ and $X_1^1, ..., X_1^n \sim X_{\rho,Q_2}$ be $2n$ i.i.d. random variables; the previous statement shows that , $\tau_I \star (\sigma_1 \star \boldsymbol{Y}_{Q_1,m}, \sigma_2 \star \boldsymbol{Y}_{Q_2,n-m}) \sim (X_1^i \mathbf{1}_{i\in I} + (\mathbf{1} - \mathbf{1}_{i\in I})X_2^i)_{1\leq i\leq n}$ (here, I is fixed). Moreover, note that $P(M = m) = \binom{n}{m} q^m (1-q)^{n-m}$ where $q = I(Q_1)/I(Q)$. Hence

$$\mathbb{E}[f(\sigma \star \boldsymbol{Y}_{Q,n})] = \sum_{I \subset [\![1,n]\!]} q^{|I|}(1-q)^{n-|I|}\mathbb{E}_{X_1^i,X_2^i}(X_1^i \mathbf{1}_{i\in I} + (\mathbf{1} - \mathbf{1}_{i\in I})X_2^i)_{1\leq i\leq n}$$

Now let $B_1, ..., B_n$ be n i.i.d. Bernoulli variables of parameter q independent of the X_1, X_2. Note that from the previous equation,

$$\mathbb{E}[f(\sigma \star \boldsymbol{Y}_{Q,n})] = \mathbb{E}[f((X_1^i B_i + X_2^i(1 - B_i))_{1\leq i\leq n})]$$

It is easy to see that $(X_1^i B_i + X_2^i(1-B_i))_{1\leq i\leq n}$ are i.i.d. and distributed as $X_{\rho,Q}$, which concludes the proof.

□

Proof of Theorem 6.1. Theorem 6.1 is now a simple consequence of Lemmas 6.6 to 6.8.The bound on the number of integral computations can be easily obtained by noting that for any sample, at most $sum_{k=1}^{d} q_k^\rho$ hyper-rectangles are visited (we do not count the first since this computation is done once and for all in any case). Since $q_k^\rho = \lceil \log_2(\delta_k/\rho) \rceil \leq \log_2(2\delta_k/\rho)$, this yields a bound of $\log_2(2^d|Q|/\rho^d) = \log_2(|Q|) + d\log_2(2/\rho)$ per sample, hence the result. □

6.D .3 Evaluating the error of the sampling algorithm : proof of Theorem 6.2

Theorem 6.2 is a specific case of the following theorem. For a given function g defined on a hyper-rectangle Q, define its Lipschitz constant with respect to the infinity norm :

$$\forall x \in Q,\ \|x\|_\infty = \sup_{1\leq k\leq d} |x_i|, \qquad \mathrm{Lip}_\infty(g) = \sup_{\substack{x,y\in Q \\ x\neq y}} \frac{|g(x)-g(y)|}{\|x-y\|_\infty}. \tag{6.68}$$

Theorem 6.7 (Variation bounds)**.** *Let Q be a hyper-rectangle, $\rho > 0$, $p_Q = f/I(Q)$ and $p_{Q,\rho}$ defined in Eq. (6.7). Recall the definition of $\mathrm{Lip}_\infty(f), \mathrm{Lip}_\infty(\sqrt{f})$ from Eq. (6.68). The following bounds hold.*

$$d_{TV}(p_Q, p_{Q,\rho}) \leq \tfrac{|Q|}{I(Q)}\ \mathrm{Lip}_\infty(f)\ \rho \tag{6.69}$$

$$H(p_Q, p_{Q,\rho}) \leq \sqrt{\tfrac{|Q|}{I(Q)}}\ \mathrm{Lip}_\infty(\sqrt{f})\ \rho \tag{6.70}$$

$$\mathbb{W}_p(p_Q, p_{Q,\rho}) \leq \sqrt{d}\rho, \qquad p \geq 1. \tag{6.71}$$

Proof. Recall that $p_Q = f\mathbf{1}_Q/I(Q)$ and hence

$$\forall x \in Q,\ p_Q(x) = \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} f(x)\mathbf{1}_{Q_\rho}(x)$$

Combining the previous equation with Eq. (6.7), it holds :

$$\forall x \in Q,\ p_Q(x) - p_{Q,\rho}(x) = \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} (f(x) - \tfrac{I(Q_\rho)}{|Q_\rho|})\mathbf{1}_{Q_\rho}(x) \tag{6.72}$$

1. Distance between f and its mean on a small cube. Let $Q_\rho \in \mathcal{D}_{Q,\rho}$ and $x \in Q_\rho$, it holds

$$|f(x) - \tfrac{I(Q_\rho)}{|Q_\rho|}| \leq \mathrm{Lip}_\infty(f)\ \rho. \tag{6.73}$$

Indeed, expanding the mean, we get $f(x) - \frac{I(Q_\rho)}{|Q_\rho|} = \frac{1}{|Q_\rho|}\int_{Q_\rho} (f(x) - f(y))\ dy$. Moreover, $|f(x) - f(y)| \leq \mathrm{Lip}_\infty(f)\ \|x-y\|_\infty$. Plugging that back in the previous equation and using the fact that $\|x-y\|_\infty \leq \rho$ on Q_ρ, we get Eq. (6.73)

2. Bounds on the total variation and L^2 distances. Using Eqs. (6.72) and (6.73), we immediately get

$$\begin{aligned}\int_Q |p_Q(x) - p_{Q,\rho}(x)|dx &= \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} |f(x) - \tfrac{I(Q_\rho)}{|Q_\rho|}|dx \\ &\leq \tfrac{|Q|\,\mathrm{Lip}_\infty(f)\ \rho}{I(Q)}.\end{aligned}$$

3. Bound on the Wasserstein norm $\mathbb{W}_p$. Consider the following density on $Q \times Q$:

$$\gamma(x,y) = \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} f(x)\mathbf{1}_{Q_\rho}(x)\tfrac{1}{|Q_\rho|}\mathbf{1}_{Q_\rho}(y). \tag{6.74}$$

A simple computation shows that $\gamma \in \Pi(p_Q, p_{Q,\rho})$ (see the work by Santambrogio (2015) and Eq. (6.33)), i.e. that its marginals are p_Q and $p_{Q,\rho}$. Hence, by definition Eq. (6.33), we have

$$\mathbb{W}_p^p(p_Q, p_{Q,\rho}) \leq \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho \times Q_\rho} |x-y|^p \tfrac{f(x)}{|Q_\rho|}\, dxdy.$$

Now using the fact that if $x, y \in Q_\rho$, we have $\|x-y\| \leq \sqrt{d}\rho$ as Q_ρ is a hyper-rectangle with all sides of length less than or equal to ρ, we finally get : $\mathbb{W}_p(p_Q, p_{Q,\rho}) \leq \sqrt{d}\rho$

4. Hellinger distance bound. Note that we could get a looser bound using Eq. (6.36) which only relies on the Lipschitz constant of f and not on that of $\sqrt{f}$. Here, we concentrate on that case.

Let $Q_\rho \in \mathcal{D}_{Q,\rho}$. By the intermediate value theorem, there exists $z \in Q_\rho$ such that $f(z) = \frac{I(Q_\rho)}{|Q_\rho|}$ and hence for any $x \in Q_\rho$, it holds

$$\left|\sqrt{f}(x) - \sqrt{\tfrac{I(Q_\rho)}{|Q_\rho|}}\right| = \left|\sqrt{f}(x) - \sqrt{f}(z)\right| \leq \mathrm{Lip}_\infty(\sqrt{f})\ \|x-z\|_\infty \leq \mathrm{Lip}_\infty(\sqrt{f})\ \rho.$$

Bounding the distance between $p_{Q,\rho}$ and p_Q by decomposing on dyadic hyper-rectangles using the previous expression, it holds

$$\begin{aligned} H(p_Q, p_{Q,\rho})^2 &= \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} \left|\sqrt{\tfrac{f(x)}{I(Q)}} - \sqrt{\tfrac{I(Q_\rho)}{|Q_\rho| I(Q)}}\right|^2 dx \\ &= \tfrac{1}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} \left|\sqrt{f(x)} - \sqrt{\tfrac{I(Q_\rho)}{|Q_\rho|}}\right|^2 dx \\ &\leq \tfrac{(\mathrm{Lip}_\infty(\sqrt{f})\ \rho)^2}{I(Q)} \sum_{Q_\rho \in \mathcal{D}_{Q,\rho}} \int_{Q_\rho} 1\ dx = \left(\sqrt{\tfrac{|Q|}{I(Q)}}\,\mathrm{Lip}_\infty(\sqrt{f})\ \rho\right)^2. \end{aligned}$$

□

6.D .4 Time complexity

In the Theorem 6.1, we measure the cost of the algorithm in terms of evaluation of integrals of the PSD model and in particular in the number of calls to the erf function (or subtractions) in the computation of such integrals. The fact that this is the true bottleneck of the algorithm can be seen in Sec. 6.D .4, as integrals take 95% of the CPU time.

6.E A general method of approximation and sampling

In this section, we prove proposition 6.1 and Theorem 6.3 using mainly results by Rudi and Ciliberto (2021). We introduce those results sequentially, showing the how each one is a building block towards the final result.

Table 6.1: Main computing times (% of the CPU time)

PART	MAIN OPERATION	TIME
Integration	Eqs. (6.3) to (6.5)	
Computing $K_{X,\eta/2}$		71%
Computing $\overline{X}$		6%
Computing $G_{X,2\eta,Q}$	Computing $\mathcal{A}, \mathcal{B}$	8%
	Calls to erf	6%
	Other	8%
Other		1%
Sampling	algorithm 2	
Computing $I(Q)$	Calls to erf	34%
	Computing $\mathcal{A}, \mathcal{B}$	26%
	Mulitplications $\sqrt{\eta}$	11%
	Other	24%
Other		5%

For this section, fix a probability distribution p on the set $\mathcal{X} = (-1, 1)^d$ (this is for the sake of simplicity; any hyper-rectangle could do), and assume that Assumption 6.1 holds for a certain $\beta \in \mathbb{N},\ \beta > 0$, i.e. there exists $J \in \mathbb{N}$ and $q_1, ..., q_J \in W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ such that $p = \sum_j q_j^2$. In this section, this probability distribution p is only known through a function f_p proportional to its density. Denote with $Z_p > 0$ this proportionality constant, i.e. $f_p/Z_p = p$, and with f_j the renormalized $q_j : q_j/\sqrt{Z_p} = f_j$ s.t. $f_p = \sum_j f_j^2$. Our goal is to be able to generate i.i.d. samples from a distribution as close as possible to p.

To do so, we first approximate f_p by a Gaussian PSD model $\widehat{f}_{\tau,m,\lambda} = f(\cdot; \widehat{A}_{\tau,m,\lambda}, \widetilde{X}_m, \eta)$ where $\eta = \tau \mathbf{1}_d$ and $\tau > 0$, $\widetilde{X}_m \in \mathbb{R}^{m \times d}$ is obtained as $(\widetilde{x}_1, ..., \widetilde{x}_m)^\top$ from m i.i.d. uniform samples from $\mathcal{X}$, and $\widehat{A}_{\tau,m,\lambda}$ is obtained by solving the problem Eq. (6.15) which we rewrite here for a given $\lambda > 0$:

$$\min_{A \in \mathbb{S}_+(\mathbb{R}^m)} \int_{\mathcal{X}} f(x; A, X, \eta)^2 dx - 2 \sum_{i=1}^{n} f_p(x_i) f(x_i; A, X, \eta) + \lambda \|K^{1/2} A K^{1/2}\|_F, \tag{6.15}$$

where $K = K_{\widetilde{X}_m,\eta}$ and the $(x_i)_{1 \le i \le n}$ represented by $X \in \mathbb{R}^{n \times d}$ are n i.i.d. samples from the uniform distribution on $\mathcal{X}$.

The parameters τ, m, n, λ are selected in order to have an ε approximation of the probability p.

Using the fact that we can easily compute integrals of Gaussian PSD models, we can easily have access to $\widehat{p}_{\tau,m,\lambda} = \widehat{f}_{\tau,m,\lambda}/\widehat{Z}_{\tau,m,\lambda}$ where $\widehat{Z}_{\tau,m,\lambda} = \|\widehat{f}_{\tau,m,\lambda}\|_{L^1(\mathcal{X})} = \int_{\mathcal{X}} \widehat{f}_{\tau,m,\lambda}(x)\ dx$.

We then apply algorithm 2 to $\widehat{p}_{\tau,m,\lambda}$, the hyper-rectangle $\mathcal{X}$, the desired number of samples N and a certain ρ controlling the size of the dyadic decomposition of $\mathcal{X}$ in order to sample from a distribution whose total variation distance to p is less than a constant times ε.

Existence of a compressed ε-close Gaussian PSD model. We start by invoking Theorem 6.6 in order to obtain an ε-approximation of f_p in the form of a general PSD $f_{\tau,\epsilon}$ with associated operator $M_{\tau,\epsilon} \in \mathbb{S}_+(\mathcal{H}_\eta)$. This PSD model can then be compressed using a compression operator as described in Appendix 6.C .2. This is the object of the following proposition.

Proposition 6.8 (Compression of $M_{\tau,\epsilon}$)**.** *Let $\varepsilon \in (0,1]$, $\tau \geq \varepsilon^{-2/\beta}$ and define $\eta = \tau \mathbf{1}_d \in \mathbb{R}^d$. Let $M_{\tau,\epsilon}$ be given by Theorem 6.6 applied to f_p and satisfying Eq. (6.63) and $f_{\tau,\epsilon}$ the corresponding PSD model.*

Let $m \in \mathbb{N}$, $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ be a data matrix corresponding to vectors $\widetilde{x}_1, ..., \widetilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X}$, and $\widetilde{P}_{\eta,m}$ be the associated orthogonal projection in $\mathcal{H}_\eta$. Let $\widetilde{M}_{\tau,m,\epsilon} := \widetilde{P}_{\eta,m} M_{\tau,\epsilon} \widetilde{P}_{\eta,m}$ be the operator associated to the compressed PSD model $\widetilde{f}_{\tau,m,\epsilon}$ of $f_{\tau,\epsilon}$ onto $\widetilde{X}_m$ (see Eq. (6.49) and Eq. (6.50) for the definitions).

Let $\delta \in (0,1]$. If one of the two following are true

$$m \geq C_1' \tau^{d/2} \left(\log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{2}\log\tau\right)^d \left(\log \tfrac{C_3'}{\delta} + \tfrac{d}{2}\log\tau + \log\log \tfrac{C_2'}{\varepsilon}\right); \tag{6.75}$$

$$m \geq C_1'' \varepsilon^{-d/\beta} \left(\log \tfrac{C_2'}{\varepsilon}\right)^d \left(\log \tfrac{C_2'}{\varepsilon} + \log \tfrac{C_3'}{\delta}\right), \qquad \tau = \varepsilon^{-2/\beta} \tag{6.76}$$

then with probability at least $1-\delta$, it holds

$$\begin{aligned} \|f_{\tau,\epsilon} - \widetilde{f}_{\tau,m,\epsilon}\|_{L^2(\mathcal{X})} \leq 2^d \ \|f_{\tau,\epsilon} - \widetilde{f}_{\tau,m,\epsilon}\|_{L^\infty(\mathcal{X})} \leq 2C \ \|f_p\|_{\mathit{sos},\mathbb{R}^d,\beta}\, \varepsilon \\ \mathrm{Tr}\,(\widetilde{M}_{\tau,m,\epsilon}) \leq \mathrm{Tr}\,(M_{\tau,\epsilon}) \leq C \ \|f_p\|_{\mathit{sos},\mathbb{R}^d,\beta} \ \tau^{d/2} \end{aligned} \tag{6.77}$$

The constants $C, C_1', C_2', C_3', C_1''$ depends only on d, β, and not on $\tau, \varepsilon, m, \delta$.

Proof. Using Eq. (6.63) in Theorem 6.6 applied to f_p, we see that if $\varepsilon \leq 1$ and $\tau \geq \varepsilon^{-2/\beta}$, there exists constants C_4, C_5 depending only on d ,β, and not on τ, ε such that $\|f(\cdot; M_{\tau,\epsilon}, \phi_\eta) - f_p\|_{L^2(\mathcal{X})} \leq C_4 \ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta} \ \varepsilon$ and $\mathrm{Tr}\,(M_{\tau,\epsilon}) \leq C_5 \ \|f_p\|_{\mathsf{sos},\mathbb{R}^d,\beta} \ \tau^{d/2}$ (we set $C_5 = C_2(1+e^{C_3})$ where C_2, C_3 are introduced in Theorem 6.6). Now setting $\rho = \frac{\varepsilon}{2^d \ \tau^{d/2}}$ which is less than 1 since $\varepsilon \leq 1$ and $\tau \geq \varepsilon^{-2/\beta} \geq 1$, we can apply Lemma 6.3 and hence, with probability at least $1-\delta$, if

$$m \geq C_1 \tau^{d/2} (\log \tfrac{C_2 \ \tau^{d/2}}{\varepsilon})^d \left(\log \tfrac{C_3}{\delta} + \log\tau + \log\log \tfrac{C_2 \ \tau^{d/2}}{\varepsilon}\right), \tag{6.78}$$

with $C_1 \leftarrow C_1$ from Lemma 6.3, $C_2 \leftarrow \max(e, C_2 2^d)$ where C_2 is given by Lemma 6.3 and $C_3 \leftarrow C_3$ from Lemma 6.3, it holds $\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leq \rho$ (hence C_1, C_2, C_3 depend only on d).

1. Let us now show that Eq. (6.78) is implied by Eq. (6.75). Let us bound :

$$\begin{aligned} \log\log \tfrac{C_2 \ \tau^{d/2}}{\varepsilon} &= \log\left(\log \tfrac{C_2}{\varepsilon}\left(1 + \frac{d/2\log\tau}{\log \frac{C_2}{\varepsilon}}\right)\right) \\ &= \log\log \tfrac{C_2}{\varepsilon} + \log\left(1 + \frac{d/2\log\tau}{\log \frac{C_2}{\varepsilon}}\right) \\ &\leq \log\log \tfrac{C_2}{\varepsilon} + \tfrac{d}{2}\log\tau, \end{aligned}$$

where the last inequality is obtained since $\log(1+t) \leq t$ and $C_2/\varepsilon \geq C_2 \geq e$ by definition of C_2 and since $\varepsilon \leq 1$. Setting $C_1' = 3C_1$, $C_2' = C_2$ and $C_3' = C_3$, it is therefore clear that Eq. (6.75) implies Eq. (6.78).

2. Moreover, Eq. (6.75) is in turn implied by Eq. (6.76). Indeed, in the case where $\tau = \varepsilon^{-2/\beta}$, we have the bound

$$\log\log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{2}\log\tau \leq \log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{2}\log\tau = \log \tfrac{C_2'}{\varepsilon} + \tfrac{d}{\beta}\log\tfrac{1}{\varepsilon} \leq (1 + d/\beta)\log \tfrac{C_2'}{\varepsilon}$$

since $C_2' \geq e \geq 1$. Thus, taking $C_1'' = C_1'(1+d/\beta)^{d+1}$, Eq. (6.76) implies Eq. (6.75).

3. If Eq. (6.78) holds, then Eq. (6.77) holds with probability at least $1-\delta$. Indeed, for the first part, since Eq. (6.78) holds, with probability at least $1-\delta$, $\sup_{x\in\mathcal{X}} \|(I-\widetilde{P}_{\eta,m})\phi_\eta(x)\| \leq \rho = \frac{\varepsilon}{2^d\tau^{d/2}}$

Moreover, using Eq. (6.61) combined with the fact that for any $x \in \mathcal{X}$, $|f(x; M_{\tau,\epsilon}, \phi_\eta)| = |\langle \phi_\eta(x), M_{\tau,\epsilon}\phi_\eta(x)\rangle| \leq \|\phi_\eta(x)\|^2_{\mathcal{H}_\eta}\|M_{\tau,\epsilon}\| = \|M_{\tau,\epsilon}\|$ since $\|\phi_\eta(x)\|^2 = k_\eta(x,x) = 1$, it holds

$$\|f(\cdot; M_{\tau,\epsilon}, \phi_\eta) - f(\cdot; \widetilde{M}_{\tau,m,\epsilon}, \phi_\eta)\|_{L^\infty(\mathcal{X})} \leq \|M_{\tau,\epsilon}\|(\rho^2+\rho) \leq 2\|M_{\tau,\epsilon}\|\rho.$$

We conclude using the fact that for any operator M, and any orthogonal projection P, $\|M\| \leq \mathrm{Tr}\,(M)$ and $\mathrm{Tr}\,(PMP) \leq \mathrm{Tr}\,(M)$. We then conclude the proof by using the definition of ρ and the fact that $\int_{\mathcal{X}} 1\, dx = 2^d$, and setting $C \leftarrow C_5$.

□

Combining Eq. (6.63) and Eq. (6.77), we see that if m is large enough, one can find a Gaussian PSD model of the form $\widetilde{f}_{\tau,m,\epsilon} = f(\,\cdot\,; \widetilde{A}_{\tau,m,\epsilon}, \widetilde{X}_m, \tau\mathbf{1}_d)$ (where $\widetilde{A}_{\tau,m,\epsilon}$ is defined through Eq. (6.50) from $\widetilde{M}_{\tau,m,\epsilon}$) which is $C\,\|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\,\varepsilon$ close to f_p and whose trace is controlled. It now remains to compare the performance of $\widetilde{f}_{\tau,m,\epsilon}$ with the Gaussian PSD model learned from evaluations of f_p, $\widehat{f}_{\tau,m,\lambda}$, which is the solution of Eq. (6.15) which we can compute.

Controlling the L^2 distance between $\widehat{f}_{\tau,m,\lambda}$ and f_p. This theorem is a rewriting of Theorem 7 by Rudi and Ciliberto (2021), but with the point of view of ε instead of n.

Proposition 6.9 (Performance of $\widehat{f}_{\tau,m,\lambda}$)**.** *Let $n \in \mathbb{N}$ and let $(x_1, ..., x_n)$ be n i.i.d. samples from p. Let $\delta \in (0,1]$ and $\varepsilon \leq \frac{1}{e}$. Assume n satisfies*

$$n \geq \varepsilon^{-(d+2\beta)/\beta} \log^d\left(\tfrac{1}{\varepsilon}\right) \log\left(\tfrac{2}{\delta}\right), \tag{6.79}$$

Let $m \in \mathbb{N}$ and assume m satisfies Eq. (6.76) and let $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ be a data matrix corresponding to vectors $\widetilde{x}_1, ..., \widetilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X}$. Let $\lambda = \varepsilon^{2(\beta+d)/\beta}$, $\tau = \varepsilon^{-2/\beta}$ and $\widehat{f}_{\tau,m,\lambda}$ be the Gaussian PSD model associated to the solution $\widehat{A}_{\tau,m,\lambda}$ of Eq. (6.15) with $\widetilde{X}_m, \lambda, \tau$. With probability at least $1-2\delta$, the following holds

$$\left(\|\widehat{f}_{\tau,m,\lambda} - f_p\|^2_{L^2(\mathcal{X})} + \lambda\|\widehat{M}_{\tau,m,\lambda}\|^2_F\right)^{1/2} \leq C\ \ \|f_p\|_{\mathtt{sos},\mathcal{X},\beta}\ \ \varepsilon, \tag{6.80}$$

where C is a constant depending only on d, β, and not on $\varepsilon, \delta, \lambda, m, \tau, f_p$.

Proof. We start by applying the same reasoning as in the proof of Theorem 7 by Rudi and Ciliberto (2021).

Note that since $\tau = \varepsilon^{-2/\beta}$ and Eq. (6.76) is satisfied, with probability at least $1-\delta$, it holds $\|\widetilde{f}_{\tau,m,\epsilon} - \widehat{f}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})} \leq 2C_1\ \ \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\ \ \varepsilon$ (where $C_1 \leftarrow C$ from Eq. (6.77)) and hence $\|f_p - \widehat{f}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})} \leq (C_0 + 2C_1)\ \ \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\ \ \varepsilon$, (where $C_0 \leftarrow C_1$ from Theorem 6.6). C_0, C_1 are both constants depending only on d, β. Moreover, since the Frobenius norm is bounded by the trace norm, by definition of τ, we also have $\|\widehat{M}_{\tau,m,\lambda}\|_F \leq \mathrm{Tr}\,(\widehat{M}_{\tau,m,\lambda}) \leq C_1\ \ \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\ \ \tau^{d/2} \leq C_1\ \ \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\ \ \varepsilon^{-d/\beta}$.

We can modify Theorem E.2 by Rudi and Ciliberto (2021) by taking $\widehat{v} = \frac{1}{n}\sum_{i=1}^n f_p(x_i)\psi_\eta(x_i)$ and $v = \int_{\mathcal{X}} f_p(x)\psi_\eta(x)\, dx$; all the formulas then remain true and adapt to our problem Eq. (6.15). Applying Theorem E.2 by Rudi and Ciliberto (2021) to $\widetilde{A}_{\tau,m,\epsilon}$ and using Lemma E.3 by Rudi and Ciliberto (2021) to simplify notation, as well as the bound on the term $\|Q_\lambda^{-1/2}(\widehat{v} - v)\|$

combining Lemma E.4 (with $\zeta = Q_\lambda^{-1/2} f_p(x)\psi_\eta(x)$) using $s = d$ and Lemma E.5 (again, for more details, see part 2 of the proof of Theorem 7 by Rudi and Ciliberto (2021)) and using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b},\ a, b \geq 0$, with probability at least $1-\delta$, it holds :

$$\begin{aligned}\left(\|\widehat{f}_{\tau,m,\lambda} - f_p\|^2_{L^2(\mathcal{X})} + \lambda\|\widehat{M}_{\tau,m,\lambda}\|^2_F\right)^{1/2} &\leq \|\widetilde{f}_{\tau,m,\epsilon} - f_p\|_{L^2(\mathcal{X})} \\ &+ \sqrt{\lambda}\|\widetilde{M}_{\tau,m,\epsilon}\|_F + C_2\ \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\ \frac{\log\frac{2}{\delta}}{n\lambda^{1/4}} \\ &+ C_3\ \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\ \frac{\tau^{d/4}\left(\log\frac{1}{\lambda}\right)^{d/2}\left(\log\frac{2}{\delta}\right)^{1/2}}{n^{1/2}}, \qquad (6.81)\end{aligned}$$

where C_2 and C_3 are constants which depend only on d.

Note that in the proof of Lemma E.4 by Rudi and Ciliberto (2021), $\|\zeta\|$ is bounded in essential supremum and standard deviation by $\|f_p\|_{L^\infty(\mathcal{X})}\times$ a quantity independent of f_p which is then bounded, hence the previous concentration bound since $\|f_p\|_{L^\infty(\mathcal{X})} \leq \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}$.

Now combining both events in a union bound, and plugging in the fact that $\lambda = \varepsilon^{\frac{2\beta+2d}{\beta}}$ and $\tau = \varepsilon^{-2/\beta}$, we see that with probability at least $1-2\delta$, the left hand term is bounded by the following quantity:

$$\begin{aligned}&\varepsilon\ \|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}\ (C_0 + 3C_1 + T), \qquad (6.82)\\ &T = C_2\frac{\varepsilon^{-\frac{3\beta+d}{2\beta}}\log\frac{2}{\delta}}{n} + C_3\frac{\varepsilon^{-(d+2\beta)/2\beta}\left(\frac{2\beta+2d}{\beta}\log\frac{1}{\varepsilon}\right)^{d/2}\left(\log\frac{2}{\delta}\right)^{1/2}}{n^{1/2}}.\end{aligned}$$

Now the goal is to bound the term T. Note that as soon as $\varepsilon \leq e^{-1}$ and $\delta \leq 2$, if $Y = \frac{\varepsilon^{-(d+2\beta)/\beta}\log^d\left(\frac{1}{\varepsilon}\right)\log\left(\frac{2}{\delta}\right)}{n}$, then it holds $T \leq \frac{C_2}{\log^2 2}Y + C_3\sqrt{Y}$. Now note that $Y \leq 1$ iif $n \geq \varepsilon^{-(d+2\beta)/\beta}\log^d\left(\frac{1}{\varepsilon}\right)\log\left(\frac{2}{\delta}\right)$. The theorem therefore holds with $C \leftarrow 1 + 3C_1 + C_2/\log^2 2 + C_3$.

Finally, the fact that all bounds involving $\|f_p\|_{\mathtt{sos},\mathbb{R}^d,\beta}$ can be replaced, up to constants depending only on β, d, by the norm $\|f_p\|_{\mathtt{sos},\mathcal{X},\beta}$, is simply a consequence of proposition 6.4. □

We now come to the final part of our section detailing the proof of proposition 6.10 and Theorem 6.8, which consists in approximately sampling from the learnt model $\widehat{f}_{\tau,m,\lambda}$ using algorithm 2 with well chosen parameters.

Performance of the re-normalized probability measure $\widehat{p}_{\tau,m,\lambda}$. We start off with a technical lemma.

Lemma 6.9 (Technical lemma)**.** *Let $\|\cdot\|$ be a norm on a vector space E, and let $x, y \in E \setminus \{0\}$. Then it holds:*

$$\left\|\frac{a}{\|a\|} - \frac{b}{\|b\|}\right\| \leq \frac{2\|a-b\|}{\|a\|}. \qquad (6.83)$$

Moreover, if $\|a-b\| \leq \|a\|/2$, it holds

$$\frac{\|a\|}{\|b\|} \leq 2. \qquad (6.84)$$

Proof. Introduce the quantity $\frac{b}{\|a\|}$ in order to get

$$\left\|\frac{a}{\|a\|} - \frac{b}{\|b\|}\right\| \leq \left\|\frac{a}{\|a\|} - \frac{b}{\|a\|}\right\| + \left\|\frac{b}{\|a\|} - \frac{b}{\|b\|}\right\| = \frac{\|a-b\|}{\|a\|} + \|b\|\left|\frac{1}{\|a\|} - \frac{1}{\|b\|}\right|.$$

One concludes by writing

$$\left|\frac{1}{\|a\|}-\frac{1}{\|b\|}\right| = \frac{|\|b\|-\|a\||}{\|a\|\ \|b\|} \leq \frac{\|b-a\|}{\|a\|\ \|b\|},$$

where the last inequality is simply the triangle inequality. This concludes the proof of Eq. (6.83). The proof of Eq. (6.84) is simply the result of applying the bound $\frac{1}{\|b\|} \leq \frac{1}{\|a\|-\|b-a\|} \leq \frac{2}{\|a\|}$. □

Proposition 6.10 (Performance of $\widehat{p}_{\tau,m,\lambda}$)**.** *Let p be a probability density w.r.t. the Lebesgue measure on $\mathcal{X}=(-1,1)^d$ satisfying Assumption 6.1 for a certain β. There exists $\varepsilon_0>0$ depending only on d,β, and $\|p\|_{sos,\mathcal{X},\beta}$ and C_1,C_2,C_1',C_2',C_3' depending only on d,β such that the following holds.*

Let $n\in\mathbb{N}$ and let $(x_1,...,x_n)$ be n i.i.d. samples selected uniformly at random from $\mathcal{X}$. Let $\delta\in(0,1]$ and $\varepsilon\leq\varepsilon_0$, $\lambda=\varepsilon^{2(\beta+d)/\beta}$ and $\tau=\varepsilon^{-2/\beta}$. Assume n satisfies Eq. (6.79), i.e.

$$n \geq \varepsilon^{-(d+2\beta)/\beta}\log^d\left(\tfrac{1}{\varepsilon}\right)\log\left(\tfrac{2}{\delta}\right). \tag{6.79}$$

Let $m\in\mathbb{N}$ and assume m satisfies Eq. (6.76), i.e.

$$m \geq C_1'\varepsilon^{-d/\beta}\left(\log\tfrac{C_2'}{\varepsilon}\right)^d\left(\log\tfrac{C_2'}{\varepsilon}+\log\tfrac{C_3'}{\delta}\right), \tag{6.76}$$

and let $\widetilde{X}_m\in\mathbb{R}^{m\times d}$ be a data matrix corresponding to vectors $\widetilde{x}_1,...,\widetilde{x}_m$ which are sampled independently and uniformly from $\mathcal{X}$.

Let $\widehat{f}_{\tau,m,\lambda}$ be the Gaussian PSD model associated to the solution $\widehat{A}_{\tau,m,\lambda}$ of Eq. (6.15) with $\widetilde{X}_m,\lambda,\tau$ and let $\widehat{p}_{\tau,m,\lambda}$ be the associated probability density on $\mathcal{X}$ (i.e. the re-normalization of $\widehat{f}_{\tau,m,\lambda}$). Let $\widehat{R}_{\tau,m,\lambda}$ be PSD operator on $\mathcal{H}_\eta$ associated to $\widehat{p}_{\tau,m,\lambda}$. With probability at least $1-2\delta$, it holds

$$d_{TV}(\widehat{p}_{\tau,m,\lambda},p)\leq C_1\ \|p\|_{sos,\mathcal{X},\beta}\ \varepsilon, \qquad \|\widehat{R}_{\tau,m,\lambda}\|_F\leq C_2\ \|p\|_{sos,\mathcal{X},\beta}\ \varepsilon^{-d/\beta}. \tag{6.85}$$

Proof. Since the assumptions of proposition 6.9 are satisfied, we have by Eq. (6.80) the existence of a constant C depending only on d, β, and not on $\varepsilon,\delta,\lambda,m,\tau,f_p$, such that

$$\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^2(\mathcal{X})}\leq C\ \|f_p\|_{\mathtt{sos},\mathcal{X},\beta}\ \varepsilon, \qquad \|\widehat{M}_{\tau,m,\lambda}\|_F\leq C\ \|f_p\|_{\mathtt{sos},\mathcal{X},\beta}\ \varepsilon^{-d/\beta}, \tag{6.86}$$

where we have used the fact that $\lambda=\varepsilon^{2+2d/\beta}$.

Now using the fact that $\|\bullet\|_{L^1(\mathcal{X})}\leq 2^{d/2}\|\bullet\|_{L^2(\mathcal{X})}$ (by Cauchy-Schwarz inequality), Eq. (6.86) shows in particular that $\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^1(\mathcal{X})}\leq 2^{d/2}C\ \|f_p\|_{\mathtt{sos},\mathcal{X},\beta}\ \varepsilon$. Now applying Eq. (6.83) of Lemma 6.9, using the fact that $\widehat{p}_{\tau,m,\lambda}=\widehat{f}_{\tau,m,\lambda}/\|\widehat{f}_{\tau,m,\lambda}\|_{L^1(\mathcal{X})}$ and $p=f_p/\|f_p\|_{L^1(\mathcal{X})}$, it holds

$$\begin{aligned} d_{TV}(\widehat{p}_{\tau,m,\lambda},p)=\|\widehat{p}_{\tau,m,\lambda}-p\|_{L^1(\mathcal{X})} &\leq 2\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^1(\mathcal{X})}/\|f_p\|_{L^1(\mathcal{X})} \\ &\leq 2^{d/2+1}C\ \|f_p\|_{\mathtt{sos},\mathcal{X},\beta}/\|f_p\|_{L^1(\mathcal{X})}\ \varepsilon. \end{aligned} \tag{6.87}$$

Since $p=f_p/\|f_p\|_{L^1(\mathcal{X})}$, we have $\|f_p\|_{\mathtt{sos},\mathcal{X},\beta}/\|f_p\|_{L^1(\mathcal{X})}=\|p\|_{\mathtt{sos},\mathcal{X},\beta}$. This shows

$$d_{TV}(\widehat{p}_{\tau,m,\lambda},p)\leq 2^{d/2+1}C\ \|p\|_{\mathtt{sos},\mathcal{X},\beta}\,\varepsilon.$$

Now set $\varepsilon_0=\min(e^{-1},2^{-d/2-1}C^{-1}\|p\|^{-1}_{\mathtt{sos},\mathcal{X},\beta})$. If $\varepsilon\leq\varepsilon_0$, we have $2^{d/2}C\,\|f_p\|_{\mathtt{sos},\mathcal{X},\beta}\,\varepsilon\leq\|f_p\|_{L^1(\mathcal{X})}/2$ and hence $\|\widehat{f}_{\tau,m,\lambda}-f_p\|_{L^1(\mathcal{X})}\leq\|f_p\|_{L^1(\mathcal{X})}/2$. By Eq. (6.84) of Lemma 6.9, we

therefore have $\|f_p\|_{L^1(\mathcal{X})}/\|\widehat{f}_{\tau,m,\lambda}\|_{L^1(\mathcal{X})} = Z_p/\widehat{Z}_{\tau,m,\lambda} \leq 2$. Now since $\widehat{R}_{\tau,m,\lambda} = \widehat{M}_{\tau,m,\lambda}/\widehat{Z}_{\tau,m,\lambda}$, using Eq. (6.86), it holds $\|\widehat{R}_{\tau,m,\lambda}\| \leq C_2 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon^{-d/\beta}$ where $C_2 = 2C$, which depends only on β, d.

□

Theorem 6.8 (Performance of $p_{\mathtt{sample}}$)**.** *Under the assumptions and notations of the previous theorem (proposition 6.1), there exists a constant C_3 depending only on d, β, such that the following holds.*

Let $\widehat{p}_{\tau,m,\lambda}$ be given by the previous proposition. Let $p_{\mathtt{sample}}$ be the dyadic approximation of $\widehat{p}_{\tau,m,\lambda}$ on $Q = \mathcal{X} = (-1,1)^d$ and of width ρ (see Eq. (6.7)). Recall from Theorem 6.1 that algorithm 2 applied to $Q = (-1,1)^d, N, \rho$ returns N i.i.d. samples from $p_{\mathtt{sample}}$.

If on the one hand ρ is set to $\varepsilon^{1+(d+1)/\beta}$, then with probability at least $1-2\delta$,

$$d_{TV}(\widehat{p}_{\tau,m,\lambda}, p_{\mathtt{sample}}) \leq C_3 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon, \qquad d_{TV}(p, p_{\mathtt{sample}}) \leq (C_1 + C_3) \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon. \tag{6.88}$$

If on the other ρ is set adaptively to guarantee $d_{TV}(p_{\mathtt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq \varepsilon$ as in Remark 23 then with probability at least $1-2\delta$, $\rho \geq \varepsilon^{1+(d+1)/\beta}/(C_3 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta})$, and hence

$$d_{TV}(\widehat{p}_{\tau,m,\lambda}, p_{\mathtt{sample}}) \leq \varepsilon, \qquad d_{TV}(p, p_{\mathtt{sample}}) \leq C_1 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon + \varepsilon. \tag{6.89}$$

In any case, this guarantees that the complexity in terms of erf *computations is bounded by*

$$O(Nm^2 \log \tfrac{1}{\rho}) = O\left(N \ \varepsilon^{-2d/\beta} \log^{2d+1}\left(\tfrac{1}{\varepsilon}\right)\left(\log\left(\tfrac{1}{\varepsilon}\right) + \log\left(\tfrac{1}{\delta}\right)\right)\right), \tag{6.90}$$

where the O notations is taken with constants depending on d, β, $\|p\|_{\mathtt{sos},\mathcal{X},\beta}$.

Proof. Let us bound $\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda})$. Note that

$$\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda}) \leq \sup_{x \in \mathcal{X}} \sum_{k=1}^{d} \partial_k \widehat{p}_{\tau,m,\lambda}(x).$$

Using Lemma 6.5, we get $\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda}) \leq d 2^{3/2}\sqrt{\tau}\|\widehat{R}_{\tau,m,\lambda}\|$. Using the fact that $\tau = \varepsilon^{-2/\beta}$ and that by Eq. (6.85), $\|\widehat{R}_{\tau,m,\lambda}\| \leq \|\widehat{R}_{\tau,m,\lambda}\|_F \leq C_2 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon^{-d/\beta}$, we therefore have $\mathrm{Lip}_\infty(\widehat{p}_{\tau,m,\lambda}) \leq 2^{3/2} d C_2 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon^{-(d+1)/\beta}$. Hence, applying Theorem 6.7 to $\widehat{p}_{\tau,m,\lambda}$, we get

$$d_{TV}(p_{\mathtt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq 2^{3/2} \ 2^d d C_2 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon^{-(d+1)/\beta} \ \rho. \tag{6.91}$$

On the one hand, if we use algorithm 2 with $\rho = \varepsilon^{1+\frac{(d+1)}{\beta}}$, by the previous equation, we get $d_{TV}(p_{\mathtt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq 2^{3/2} d \ 2^d C_2 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon$.

If on the other hand we find ρ adaptively by computing a bound

$$\widetilde{\mathrm{Lip}}(A) = 2^{3/2}\tau^{1/2} d \|K^{1/2} A K^{1/2}\| = 2^{3/2}\tau^{1/2} d \|\widehat{R}_{\tau,m,\lambda}\|_F$$

from $\widehat{p}_{\tau,m,\lambda}$ as in Remark 23, and finding ρ such that $2^d \widetilde{\mathrm{Lip}}(A) \ \rho = \frac{|Q|}{I(Q)} \widetilde{\mathrm{Lip}}(A) \ \rho = \varepsilon$, since the adaptive bound will have computed

$$\widetilde{\mathrm{Lip}}(A) \ \leq 2^{3/2} d C_2 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta} \ \varepsilon^{-(d+1)/\beta},$$

we will get $\rho \geq \frac{\varepsilon^{1+(d+1)/\beta}}{2^{d+3/2} \ d \ C_2 \ \|p\|_{\mathtt{sos},\mathcal{X},\beta}}$ and hence $d_{TV}(p_{\mathtt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq \varepsilon$. The last point is just a consequence of Theorem 6.1 and the bound on m in Eq. (6.76). □

6.F Approximation and sampling using a rank one PSD model

In this section, we prove the results in Sec. 6.4 .2, i.e. proposition 6.2 and Theorem 6.4.

For this section, fix a probability which has density p with respect to the Lebesgue measure dx on $\mathcal{X} = (-1,1)^d$, (this is for the sake of simplicity; any hyper-rectangle could do), and assume that Assumption 6.2 holds for a certain $\beta \in \mathbb{N},\ \beta > 0$, i.e. there exists $q \in W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ such that $p = q^2$. This is the case, for instance, when $p \propto e^{-V(x)}$ where V is β times differentiable.

One of the main advantages of our method will be to deal with probability measures which are known up to a constant; therefore, in this section, we take f_p such that $p = f_p/Z(f_p)$ where $Z(f_p) = \int_{\mathcal{X}} f_p(x)dx$. Assuming Assumption 6.2 holds, we take $g_p \in W_2^\beta(\mathcal{X}) \cap L^\infty(\mathcal{X})$ such that $g_p^2 = f_p$ as and assume that p is only known through function evaluations of g_p, i.e. we can evaluate the function $g_p(x)$ for any $x \in \mathcal{X}$.

Once again, our goal is to be able to generate N i.i.d. samples from a distribution which is ε-close to p, in a sense which we will define. To do so, we first approximate g_p by a Gaussian linear model $\widehat{g}_{\tau,m,\lambda} = g(\bullet;\ \widehat{a}_{\tau,m,\lambda}, \widetilde{X}_m, \eta)$ (see Eq. (6.2) for a definition) where $\eta = \tau \mathbf{1}_d$ for some $\tau > 0$, $\widetilde{X}_m \in \mathbb{R}^{m\times d}$ is obtained as $(\widetilde{x}_1, ..., \widetilde{x}_m)^\top$ from m i.i.d. uniform samples from $\mathcal{X}$, and $\widehat{a}_{\tau,m,\lambda}$ is obtained by solving the problem Eq. (6.20) which we rewrite here for a given $\lambda > 0$ and for n i.i.d. samples $(x_1, .., x_n)$ sampled uniformly from $\mathcal{X}$:

$$\widehat{a}_{\tau,m,\lambda} = \operatorname*{arg\,min}_{a\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \left(g(x_i;\ a, \widetilde{x}_m, \tau\mathbf{1}_d) - g_p(x_i)\right)^2 + \lambda a^\top K_{\widetilde{X}_m,\eta} a. \tag{6.20}$$

This yields a Gaussian linear model $\widehat{g}_{\tau,m,\lambda} \in \mathcal{H}_\eta$ of g_p. Since $\widehat{g}^2_{\tau,m,\lambda} = \widehat{f}_{\tau,m,\lambda}$ is a PSD model (indeed $\widehat{f}_{\tau,m,\lambda} = f(\bullet;\ \widehat{A}_{\tau,m,\lambda}, \widetilde{X}_m, \tau\mathbf{1}_d)$ with $\widehat{A}_{\tau,m,\lambda} = \widehat{a}_{\tau,m,\lambda}\widehat{a}^\top_{\tau,m,\lambda}$), we can see $\widehat{f}_{\tau,m,\lambda}$ as a Gaussian PSD model of f_p, and hence its renormalized version $\widehat{p}_{\tau,m,\lambda}$ as a PSD model of p.

The parameters τ, m, λ, n are selected in order to have an ε approximation of the probability p.

Furthermore, note that the first term in the optimized quantity in Eq. (6.20) is an empirical version of the quantity

$$\frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} \left|\sqrt{\widehat{f}_{\tau,m,\lambda}(x)} - \sqrt{f_p}(x)\right|^2 dx \leq \frac{1}{|\mathcal{X}|}\int_{\mathcal{X}} |\widehat{g}_{\tau,m,\lambda}(x) - g_p(x)|^2\ dx.$$

This quantity is related to Hellinger distance $H(p, \widehat{p}_{\tau,m,\lambda})$ defined in Eq. (6.32).

This will therefore be the natural measure in which to express the quality of the approximation $\widehat{p}_{\tau,m,\lambda}$ of p in this section.

The bound obtained on the performance of $\widehat{p}_{\tau,m,\lambda}$ can be decomposed into two steps.

- We start by bounding the distance between any $g \in \mathcal{H}_\eta$ and $\widehat{g}_{\tau,m,\lambda}$ in Theorem 6.9.
- We then select a $g_{\tau,\varepsilon}$ which is ε-close to g_p, and use it as a reference point in order to bound the distance between g_p and $\widehat{g}_{\tau,m,\lambda}$. To do so, we need to apply different concentration inequalities to obtain a final bound in terms of performance for both $\widehat{f}_{\tau,m,\lambda}$ with respect to f_p and $\widehat{p}_{\tau,m,\lambda}$ with respect to p in Hellinger distance in proposition 6.2.

Bound on the performance of $\widehat{g}_{\tau,m,\lambda}$ compared to an arbitrary function g. Here, we adapt Theorem 2. by Rudi, Camoriano, and Rosasco (2015).

Theorem 6.9 (Bounding the error (Rudi et al., 2015))**.** *Let $\eta \in \mathbb{R}^d_{++}$ and $g \in \mathcal{H}_\eta$.*

$$\begin{aligned}\|C_{\eta,\lambda}^{1/2}(g - \widehat{g}_{\tau,m,\lambda})\| \leq \theta_1^2\theta_2 \ \|\widehat{g}_p - \widehat{S}_\eta g\|_{\mathbb{R}^n} \\ + \|g\|_{\mathcal{H}_\eta}(1 + \theta_1\theta_2 + \theta_1^2) \ \left(\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| + \lambda^{1/2}\right),\end{aligned} \tag{6.92}$$

where $\theta_1 = \|\widehat{C}_{\eta,\lambda}^{-1/2}C_{\eta,\lambda}^{1/2}\|$, $\theta_2 = \|\widehat{C}_{\eta,\lambda}^{1/2}C_{\eta,\lambda}^{-1/2}\|$ and $\widehat{g}_p = (g_p(x_i)/\sqrt{n})_{1\leq i\leq n} \in \mathbb{R}^n$.

Proof. Let $g \in \mathcal{H}_\eta$. We can apply a modification of Theorem 2 by Rudi, Camoriano, and Rosasco (2015). Indeed, consider in the notations by Rudi, Camoriano, and Rosasco (2015) the loss $\mathcal{E}(f) = \|C_\eta^{1/2}(f - g)\|_{\mathcal{H}_\eta}$, and note that the assumptions are satisfied with $\nu = 0$ and $R = \|g\|_{\mathcal{H}_\eta}$, since g minimizes $\mathcal{E}$ and $\|C^{-0}g\|_{\mathcal{H}_\eta} = \|g\|_{\mathcal{H}_\eta}$. Moreover, note that in the proof of that theorem, one can replace C_η by $C_{\eta,\lambda}$ without changing the result (indeed, in the proof, one always bounds $\|C_\eta^{1/2} \star\| \leq \|C_\eta^{1/2}C_{\eta,\lambda}^{-1/2}\| \ \|C_{\eta,\lambda}^{1/2} \star\| \leq \|C_{\eta,\lambda}^{1/2} \star\|$). Thus, in that setting, without combining the "constant" terms in the bounds and looking into the proof of Theorem 2 by Rudi, Camoriano, and Rosasco (2015), it holds

$$\|C_{\eta,\lambda}^{1/2}(\widehat{g}_{\tau,m,\lambda} - g)\| \leq \theta_1^2 \ \|C_{\eta,\lambda}^{-1/2}\widehat{S}_\eta^*(\widehat{g}_p - \widehat{S}_\eta g)\| + R(1 + \theta_1\theta_2)\|(I - \widetilde{P}_{\eta,m})C_{\eta,\lambda}^{1/2}\| + R\theta_1^2\lambda^{1/2}, \tag{6.93}$$

where $\theta_1 = \|\widehat{C}_{\eta,\lambda}^{-1/2}C_{\eta,\lambda}^{1/2}\|$ and $\theta_2 = \|\widehat{C}_{\eta,\lambda}^{1/2}C_{\eta,\lambda}^{-1/2}\|$.

Note that $\|C_{\eta,\lambda}^{-1/2}\widehat{S}_\eta^*(\widehat{g}_p - \widehat{S}_\eta g)\| \leq \|C_{\eta,\lambda}^{-1/2}\widehat{S}_\eta^*\| \ \|\widehat{g}_p - \widehat{S}_\eta g\|_{\mathbb{R}^n} \leq \theta_2 \ \|\widehat{g}_p - \widehat{S}_\eta g\|_{\mathbb{R}^n}$ since $\|C_{\eta,\lambda}^{-1/2}\widehat{S}_\eta^*\|^2 = \|C_{\eta,\lambda}^{-1/2}\widehat{C}_\eta C_{\eta,\lambda}^{-1/2}\| \leq \|C_{\eta,\lambda}^{-1/2}\widehat{C}_{\eta,\lambda}C_{\eta,\lambda}^{-1/2}\| = \theta_2^2$.

Moreover, using the definition of C_η, it holds

$$\begin{aligned}\|(I - \widetilde{P}_{\eta,m})C_{\eta,\lambda}^{1/2}\|^2 &= \|(I - \widetilde{P}_{\eta,m})C_\eta(I - \widetilde{P}_{\eta,m}) + \lambda(I - \widetilde{P}_{\eta,m})\| \\ &\leq \tfrac{1}{|X|}\left\|\int_{\mathcal{X}} (I - \widetilde{P}_{\eta,m})\phi_\eta(x) \otimes \phi_\eta(x)(I - \widetilde{P}_{\eta,m}) \ dx\right\| + \lambda\|(I - \widetilde{P}_{\eta,m})\| \\ &\leq \sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\|^2 + \lambda.\end{aligned}$$

Combining these results and using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, we get the bound. □

Performance of $\widehat{p}_{\tau,m,\lambda}$. We can now state the main results of this section, i.e. the bound on the performance of $\widehat{p}_{\tau,m,\lambda}$.

Proposition 6.11 (Performance of $\widehat{p}_{\tau,m,\lambda}$)**.** *Let p be a probability density on $\mathcal{X} = (-1, 1)^d$, and assume $p = q^2$ and $q \in L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})$ for some $\beta \geq 0$. Let $\widetilde{\nu} > \min(1, d/(2\beta))$. There exists a constant ε_0 depending only on $\|q\|_{L^\infty(\mathcal{X})}, \|q\|_{W_2^\beta(\mathcal{X})}, \beta, d$, constants C_1, C_2, C_3, C_4, C_5 depending only on β, d and a constant C_1' depending only on $\beta, d, \widetilde{\nu}$ such that the following holds.*

Let $\delta \in (0, 1]$ and $\varepsilon \leq \varepsilon_0$, and assume $(x_1, ..., x_n)$ and $(\widetilde{x}_1, ..., \widetilde{x}_m)$ are respectively n and m uniform i.i.d. samples on $\mathcal{X}$, satisfying

$$m \geq C_1\varepsilon^{-d/\beta} \log^d \tfrac{C_2}{\varepsilon} \log \tfrac{C_3}{\delta\varepsilon} \tag{6.94}$$

$$n \geq C_1'\varepsilon^{-2\widetilde{\nu}} \log \tfrac{8}{\delta} \tag{6.95}$$

Let $\tau = \varepsilon^{-2/\beta}$, $\eta = \tau \mathbf{1}_d$ *and* $\lambda = \varepsilon^{2+d/\beta}$. *Let* $\widehat{a}_{\tau,m,\lambda} \in \mathbb{R}^n$ *be the vector obtained by solving Eq. (6.20) and* $\widehat{g}_{\tau,m,\lambda} \in \mathcal{H}_\eta$ *the associated Gaussian linear model (see Eq. (6.2)). Let* $\widehat{f}_{\tau,m,\lambda} = \widehat{g}^2_{\tau,m,\lambda}$ *be the associated Gaussian PSD model,* $\widehat{Z}_{\tau,m,\lambda} = \int_{\mathcal{X}} \widehat{f}_{\tau,m,\lambda}(x)\, dx$ *be the normalizing constant, and* $\widehat{p}_{\tau,m,\lambda} = \widehat{f}_{\tau,m,\lambda}/\widehat{Z}_{\tau,m,\lambda}$ *be the renormalized PSD model, which is a probability density. Let* $\widehat{R}_{\tau,m,\lambda}$ *be PSD operator in* $\mathbb{S}_+(\mathcal{H}_\eta)$ *associated to* $\widehat{p}_{\tau,m,\lambda}$.

With probability at least $1 - 3\delta$, *it holds*

$$\begin{aligned} H(\widehat{p}_{\tau,m,\lambda}, p) &\leq C_4 \|q\|_{L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})} \varepsilon \\ \operatorname{Tr}(\widehat{R}_{\tau,m,\lambda}) &= \left\| \frac{\widehat{g}_{\tau,m,\lambda}}{\sqrt{\widehat{Z}_{\tau,m,\lambda}}} \right\|^2_{\mathcal{H}_\eta} \leq C_5 \|q\|^2_{L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})} \varepsilon^{-d/\beta}, \end{aligned} \tag{6.96}$$

where $\| \bullet \|_{L^\infty(\mathcal{X}) \cap W_2^\beta(\mathcal{X})} = \max(\| \bullet \|_{W_2^\beta(\mathcal{X})}, \| \bullet \|_{L^\infty(\mathcal{X})})$.

Proof. Let $\tau > 0$, and define $\eta = \tau \mathbf{1}_d$. By proposition 6.4, we can extend g_p to the whole of $\mathbb{R}^d$ and there exists an constant C such that $\|g_p\|_{W_2^\beta(\mathbb{R}^d)} \leq \|g_p\|_{W_2^\beta(\mathcal{X})}$ and $\|g_p\|_{L^\infty(\mathbb{R}^d)} \leq C \|g_p\|_{L^\infty(\mathcal{X})}$. We still denote with g_p such an extension. Let $g_{\tau,\varepsilon}$ be given by proposition 6.7 when approximating g_p.

Setting $\tau = \varepsilon^{-2/\beta}$ and $\lambda = \varepsilon^{\frac{2\beta+d}{\beta}}$, since we assume $\varepsilon \leq 1$, Eq. (6.51) gives us two constants C_1, C_2 depending only on β, d such that

$$\begin{cases} \|g_{\tau,\varepsilon} - g_p\|_{L^2(\mathbb{R}^d)} \leq \varepsilon \|g_p\|_{W_2^\beta(\mathbb{R}^d)} \\ \|g_{\tau,\varepsilon} - g_p\|_{L^\infty(\mathbb{R}^d)} \leq C_1\, \varepsilon^{1-\nu}\, \|g_p\|_\bullet \end{cases} \|g_{\tau,\varepsilon}\|_{\mathcal{H}_\eta} \leq C_2\, \|g_p\|_{W_2^\beta(\mathbb{R}^d)}\, \tau^{d/4} = C_2\, \|g_p\|_{W_2^\beta(\mathbb{R}^d)} \varepsilon^{-\frac{d}{2\beta}}.$$

1. Bounding $\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|_{\mathbb{R}^n}$ Apply Theorem 3 by Boucheron, Lugosi, and Massart (2013), reformulated in Proposition 10 by Rudi, Camoriano, and Rosasco (2015). Consider the random variable $\zeta = (g_{\tau,\varepsilon} - g_p)(X)^2 - \frac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})}$ where X follows the uniform law on $\mathcal{X}$. Then $|\zeta| \leq \|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})}$ almost surely, and $\mathbb{E}[\zeta^2] \leq \|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})} \frac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})}$. Applying the concentration bound yields that with probability at least $1 - \delta$, it holds

$$\begin{aligned} \|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|^2_{\mathbb{R}^n} - \tfrac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})} \leq{}& \frac{2\|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})} \log\frac{1}{\delta}}{3n} \\ &+ \sqrt{\frac{2\|g_{\tau,\varepsilon} - g_p\|^2_{L^\infty(\mathcal{X})} \frac{1}{|\mathcal{X}|}\|g_{\tau,\varepsilon} - g_p\|^2_{L^2(\mathcal{X})} \log\frac{1}{\delta}}{n}}, \end{aligned}$$

and thus

$$\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|^2_{\mathbb{R}^n} \leq \left(\tfrac{1}{\sqrt{|\mathcal{X}|}}\|g_{\tau,\varepsilon} - g_p\|_{L^2(\mathcal{X})} + \|g_{\tau,\varepsilon} - g_p\|_{L^\infty(\mathcal{X})} \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \right)^2.$$

Hence, by Eq. (6.51), and because $|\mathcal{X}| = 2^d$, there exists two constants C_3 and C_4 depending only on d and β such that with probability at least $1 - \delta$, it holds

$$\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|_{\mathbb{R}^n} \leq C_3 \varepsilon\, \|g_p\|_{W_2^\beta(\mathbb{R}^d)} + C_4 \varepsilon \frac{\|g_p\|_\bullet \log\frac{1}{\delta}}{\varepsilon^\nu \sqrt{n}}. \tag{6.97}$$

2. Guaranteeing $\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leq \lambda^{1/2} = \varepsilon^{1+d/(2\beta)}$ Using Lemma 6.3 and proceeding in the same way as in point 2 of the proof of proposition 6.8, we see that there exists constants C_5, C_6, C_7 depending only on d and β such that as soon as

$$m \geq C_5 \varepsilon^{-d/\beta} \left(\log \tfrac{C_6}{\varepsilon}\right)^d \log \tfrac{C_7}{\delta\varepsilon}, \tag{6.98}$$

it holds $\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leq \lambda^{1/2}$ with probability at least $1-\delta$.

3. Finding a lower bound for $\|C_\eta\|$ This will be necessary in the next bound. Let $v(z) = k_\eta(0,z) = e^{-\tau\|z\|^2}$. Then $\|v\|_{\mathcal{H}_\eta} = 1$ and

$$\begin{aligned}
\|C_\eta^{1/2} v\|_{\mathcal{H}}^2 &= \tfrac{1}{|\mathcal{X}|} \int_{\mathcal{X}} |v(x)|^2 \, dx \\
&= \tfrac{1}{|\mathcal{X}|} \left(\int_{-1}^{1} e^{-2\tau t^2} \, dt \right)^d \\
&\geq \tfrac{1}{2^d} \left(\int_{-1}^{1} e^{-2t^2} \, dt \right)^d \tau^{-d/2} = C_8 \, \tau^{-d/2},
\end{aligned}$$

where the last inequality comes from the fact that $\tau \geq 1$ since $\varepsilon \leq 1$. Hence, $\|C_\eta\| \geq C_8 \tau^{-d/2}$ where C_8 is a constant depending only on d. Hence, as soon as $\lambda \leq C_8 \tau^{-d/2}$ which rewrites $\varepsilon \leq \sqrt{C_8}$, it holds $\lambda \leq \|C_\eta\|$.

4. Bounding θ_1, θ_2**.** Using the same reasoning as that of Proposition 2. by Rudi, Camoriano, and Rosasco (2015), if $b = \|C_{\eta,\lambda}^{-1/2}(\widehat{C}_\eta - C_\eta)C_{\eta,\lambda}^{-1/2}\|$, then $\theta_1 \leq 1/(1-b)$ and $\theta_2^2 \leq 1+b$. Bounding b can be done using Proposition 8 by Rudi, Camoriano, and Rosasco (2015): if $\lambda \leq \|C_\eta\|$, and $\delta \in (0,1]$ it holds, with probability at least $1-\delta$:

$$\|C_{\eta,\lambda}^{-1/2}(\widehat{C}_\eta - C_\eta)C_{\eta,\lambda}^{-1/2}\| \leq \frac{2(1+\mathcal{N}_\infty(\lambda))\log\frac{8}{\lambda\delta}}{3n} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log\frac{8}{\lambda\delta}}{n}}, \tag{6.99}$$

where we have used the fact that $\mathrm{Tr}\,(C_\eta) \leq 1$.

Note that $\mathcal{N}_\infty(\lambda) = \sup_{x\in\mathcal{X}} \|C_{\eta,\lambda}^{-1/2}\phi_\eta(x)\|^2 \leq C_9 \tau^{(s-d)d/(2s)} \lambda^{-d/(2s)}$ for any $s > d/2$ where C_9 depends only on s, d by a proof completely analogous as that of Step 2 of Lemma E.4 by Rudi and Ciliberto (2021). Replacing the values of τ, λ yields : $\mathcal{N}_\infty(\lambda) \leq C_9 \varepsilon^{-\frac{2d(\beta+s)-d^2}{2s\beta}}$.

Note that the function $\gamma : s \in]d/2, +\infty[\mapsto \frac{2d\beta+2ds-d^2}{2s\beta}$ is a homography and therefore reaches all the values $\widetilde{\nu}$ strictly between 2 and d/β.

Therefore, for any $\widetilde{\nu} > \nu$, there exists a constant C_{10} depending only on d and $\widetilde{\nu}$ such that $(1+\mathcal{N}_\infty(\lambda))\log\frac{1}{\lambda} \leq C_{10}\varepsilon^{-2\widetilde{\nu}}$.

Hence, there exists a constant depending only on $d, \beta, \widetilde{\nu}$ such that if $n \geq C_{11}\varepsilon^{-2\widetilde{\nu}} \log\frac{8}{\delta}$, and if $\varepsilon \leq \min(1/2, \sqrt{C_8})$ then $b \leq 1/3$ (here we have bounded $\log\frac{8}{\delta\lambda}$ by a constant times $\log\frac{1}{\lambda}\log\frac{8}{\delta}$ provided $\varepsilon \leq 1/2$ and hence $\lambda \leq 1/4$. Moreover, note that C_{11} can be taken large enough, by Eq. (6.97), to guarantee the following, also with probability $1-\delta$:

$$\|\widehat{g}_p - \widehat{S}_\eta g_{\tau,\varepsilon}\|_{\mathbb{R}^n} \leq C_3\varepsilon \, \|g_p\|_{W_2^\beta(\mathbb{R}^d)} + C_4\varepsilon\|g_p\|_\bullet. \tag{6.100}$$

5. Applying Theorem 6.9 to $g_{\tau,\varepsilon}$. Combining all the previous equations, we get that if $n \geq C_{11}\varepsilon^{-2\widetilde{\nu}} \log \frac{8}{\delta}$, $\varepsilon \leq \min(1/2, \sqrt{C_8})$ and $m \geq C_5 \varepsilon^{-d/\beta} \left(\log \frac{C_6}{\varepsilon}\right)^d \log \frac{C_7}{\delta\varepsilon}$, it holds Eq. (6.100) and $b \leq 1/3$ as well as $\sup_{x\in\mathcal{X}} \|(I - \widetilde{P}_{\eta,m})\phi_\eta(x)\| \leq \lambda^{1/2}$ and hence, using the bound on $g_{\tau,\varepsilon}$, there exists a constant C_{12} depending only on d, β such that

$$\|C_{\eta,\lambda}^{1/2}(g_{\tau,\varepsilon} - \widehat{g}_{\tau,m,\lambda})\| \leq C_{12} \max(\|g_p\|_{W_2^\beta(\mathbb{R}^d)}, \|g_p\|_\bullet)\varepsilon.$$

Thus, using the bound on $\|g_{\tau,\varepsilon} - g_p\|_{L^2(\mathbb{R}^d)}$, and the fact that $gC_\eta g = \frac{1}{|\mathcal{X}|}\|g\|^2_{L^2(\mathcal{X})}$ we get

$$\begin{aligned} \|g_p - \widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})} &\leq C_{13} \max(\|g_p\|_{W_2^\beta(\mathbb{R}^d)}, \|g_p\|_\bullet)\varepsilon, \\ \|\widehat{g}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta} &\leq C_{14} \max(\|g_p\|_{W_2^\beta(\mathbb{R}^d)}, \|g_p\|_\bullet)\ \varepsilon^{-d/2\beta}. \end{aligned} \tag{6.101}$$

6. Bounding the performance of $\widehat{p}_{\tau,m,\lambda}$. Note that $q = \frac{g_p}{\|g_p\|_{L^2(\mathcal{X})}}$ and $\sqrt{\widehat{p}_{\tau,m,\lambda}} = \frac{|\widehat{g}_{\tau,m,\lambda}|}{\|\widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})}}$. Thus, using Eq. (6.83), it holds

$$\begin{aligned} H(\widehat{p}_{\tau,m,\lambda}, p) &= \left\| \frac{g_p}{\|g_p\|_{L^2(\mathcal{X})}} - \frac{|\widehat{g}_{\tau,m,\lambda}|}{\|\widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})}} \right\|_{L^2(\mathcal{X})} \\ &\leq 2\frac{\|\widehat{g}_{\tau,m,\lambda} - g_p\|_{L^2(\mathcal{X})}}{\|g_p\|_{L^2(\mathcal{X})}}. \end{aligned}$$

Hence, since $q = g_p/\|g_p\|_{L^2(\mathcal{X})}$, we have by Eq. (6.101) :

$$H(p, \widehat{p}_{\tau,m,\lambda}) \leq 2C_{13} \max(\|q\|_{W_2^\beta(\mathbb{R}^d)}, \|q\|_\bullet)\varepsilon.$$

Moreover, by Eq. (6.84), if $2C_{13} \max(\|q\|_{W_2^\beta(\mathbb{R}^d)}, \|q\|_\bullet)\varepsilon \leq 1$, then $\frac{\|g_{\tau,\varepsilon}\|_{L^2(\mathcal{X})}}{\|\widehat{g}_{\tau,m,\lambda}\|_{L^2(\mathcal{X})}} \leq 2$ and hence again by Eq. (6.101), $\|\widehat{p}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta} \leq 2C_{14} \max(\|q\|_{W_2^\beta(\mathbb{R}^d)}, \|q\|_\bullet)\ \varepsilon^{-d/2\beta}$. Setting $\varepsilon_0 = \min(1/2, \sqrt{C_8}, (2C_{13} \max($ we therefore have all the desired properties.

7. Replacing norms on $\mathbb{R}^d$ with norm on $\mathcal{X}$. To do so, we just use proposition 6.4, which does not change anything up to multiplicative constants depending only on d, β.

□

Theorem 6.10 (Performance of $p_{\texttt{sample}}$)**.** *Under the assumptions and notations of the previous theorem (proposition 6.11), there exists a constant C_6 depending only on d, β, such that the following holds. Let $\widehat{p}_{\tau,m,\lambda}$ be given by the previous proposition. Let $p_{\texttt{sample}}$ be the dyadic approximation of $\widehat{p}_{\tau,m,\lambda}$ on $Q = \mathcal{X} = (-1,1)^d$ and of width ρ (see Eq. (6.7)). Recall from Theorem 6.1 that algorithm 2 applied to $Q = (-1,1)^d, N, \rho$ returns N i.i.d. samples from $p_{\texttt{sample}}$.*

If on the one hand ρ is set to $\varepsilon^{1+(d+2)/(2\beta)}$, then with probability at least $1 - 3\delta$,

$$\begin{aligned} H(\widehat{p}_{\tau,m,\lambda}, p_{\texttt{sample}}) &\leq C_6\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\ \varepsilon, \\ H(p, p_{\texttt{sample}}) &\leq (C_4 + C_6)\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\ \varepsilon. \end{aligned} \tag{6.102}$$

If on the other ρ is set adaptively to guarantee $H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq \varepsilon$ as in Remark 23, then with probability at least $1-3\delta$,

$$\begin{aligned}&\rho \geq \varepsilon^{1+(d+2)/\beta}/(C_6\ \|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}),\\ &H(\widehat{p}_{\tau,m,\lambda}, p_{\texttt{sample}}) \leq \varepsilon, H(p, p_{\texttt{sample}}) \leq (C_1+1)\varepsilon.\end{aligned} \tag{6.103}$$

In any case, this guarantees that the complexity in terms of erf *computations is bounded by*

$$O(Nm^2 \log \tfrac{1}{\rho}) = O\left(N\ \varepsilon^{-2d/\beta} \log^{2d+1}\left(\tfrac{1}{\varepsilon}\right)\left(\log\left(\tfrac{1}{\varepsilon}\right) + \log\left(\tfrac{1}{\delta}\right)\right)\right), \tag{6.104}$$

where the O notations is taken with constants depending on d, β, $\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}$.

Proof. Let us bound $\mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}})$. Note that since for any $x, y \in \mathcal{X}$, it holds

$$\begin{aligned}\left|\sqrt{\widehat{p}_{\tau,m,\lambda}}(x) - \sqrt{\widehat{p}_{\tau,m,\lambda}}(y)\right| &= \left||\widehat{g}_{\tau,m,\lambda}(x)| - |\widehat{g}_{\tau,m,\lambda}(y)|\right|/\sqrt{\widehat{Z}_{\tau,m,\lambda}}\\ &\leq |\widehat{g}_{\tau,m,\lambda}(x) - \widehat{g}_{\tau,m,\lambda}(y)|/\sqrt{\widehat{Z}_{\tau,m,\lambda}},\end{aligned}$$

we have $\mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}}) \leq \mathrm{Lip}_\infty(\widehat{g}_{\tau,m,\lambda})/\widehat{Z}_{\tau,m,\lambda}$. Now

$$\mathrm{Lip}_\infty(\widehat{g}_{\tau,m,\lambda}) \leq \sup_{x\in\mathcal{X}} \sum_{k=1}^{d} \partial_k \widehat{g}_{\tau,m,\lambda}(x).$$

Using Lemma 6.2, we get $\mathrm{Lip}_\infty(\widehat{g}_{\tau,m,\lambda}) \leq d\sqrt{2\tau}\|\widehat{g}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta}$. Using the fact that $\tau = \varepsilon^{-2/\beta}$ and that by Eq. (6.96), $\|\widehat{g}_{\tau,m,\lambda}\|_{\mathcal{H}_\eta}/\sqrt{\widehat{Z}_{\tau,m,\lambda}} \leq \sqrt{C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\varepsilon^{-d/(2\beta)}$, we therefore have $\mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}}) \leq d\sqrt{2C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\varepsilon^{-(d+2)/(2\beta)}$. Hence, applying Theorem 6.7 to $\widehat{p}_{\tau,m,\lambda}$, we get

$$H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq 2^{d/2} d\sqrt{2C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}\varepsilon^{-(d+2)/(2\beta)}\ \rho. \tag{6.105}$$

On the one hand, if we use algorithm 2 with $\rho = \varepsilon^{1+\frac{(d+2)}{2\beta}}$, by the previous equation, we get $H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq 2^{d/2} d\sqrt{2C_5}\ \varepsilon$.

If on the other hand we find ρ adaptively by computing an upper bound $\widetilde{\mathrm{Lip}}(a)$ defined in s.t. $\widetilde{\mathrm{Lip}}(a) = \sqrt{2\tau}d\|K^{1/2}a\| = \sqrt{2\tau}d\|\widehat{g}_{\tau,m,\lambda}\|/\sqrt{\widehat{Z}_{\tau,m,\lambda}} \geq \mathrm{Lip}_\infty(\sqrt{\widehat{p}_{\tau,m,\lambda}})$ from $\widehat{p}_{\tau,m,\lambda}$ and finding ρ such that $2^{d/2}\widetilde{\mathrm{Lip}}(a)\ \rho = \varepsilon$, we will get $\rho \geq \frac{\varepsilon^{1+(d+2)/(2\beta)}}{2^{d/2}\ d\ \sqrt{2C_5}\|q\|_{L^\infty(\mathcal{X})\cap W_2^\beta(\mathcal{X})}}$ and hence $H(p_{\texttt{sample}}, \widehat{p}_{\tau,m,\lambda}) \leq \varepsilon$. The last point is just a consequence of Theorem 6.1 and the bound on m in Eq. (6.94).

□

6.G Additional experimental details

As mentioned in Sec. 6.5 , we report in Fig. 6.4 an experiment in which we learn the density of the indicator function of $[-1, 1]$ using algorithm 5.

Note that this is out of the setting of Theorem 6.4, as these bounds rely on the regularity of the target density which is not at all the case here.

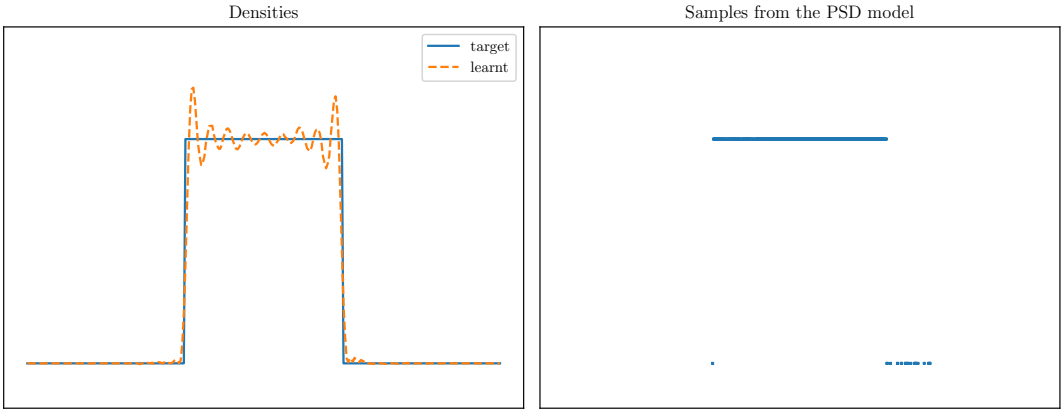

Figure 6.4: Trying to learn a non-continuous function using a rank one PSD model. *(left)* Plot of the target and learnt distributions using algorithm 5. *(right)* 1000 samples generated from the learnt distribution $p_{\texttt{sample}}$.

However, in order to sample approximately from p as a rough approximation, algorithm 5 could be relevant : it shows that we must develop tools which analyse these algorithms beyond notions of regularity, with rougher objectives.

Part III

Sum of squares of functions

Table of Contents

7 A parallel with moment-SOS hierarchies 321

8 Finding global minima via kernel approximations 347

9 SOS decompositions of smooth functions 419

Chapter 7

A parallel between kernel sums of squares and polynomial moment-SOS hierarchies

Contents

7.1 Polynomial optimization 322
7.2 Global optimization through kernel sums of squares 332
7.3 Similarities and differences between the two approaches 342

In this part, we present the works by Rudi, Marteau-Ferey, and Bach (2020); Marteau-Ferey, Bach, and Rudi (2022b). As explained in Sec. 1.3.3, the general goal is to minimize a function f defined on a set $\mathcal{X}$, *i.e.*, solve the minimization problem

$$\min_{x \in \mathcal{X}} f(x). \tag{7.1}$$

As explained in Secs. 1.1.4 and 1.3.3, this method is based on the reformulation of the minimization as a convex optimization problem with a potentially infinite number of constraint. This problem simply maximizes a lower bound on the function :

$$\begin{aligned} &\sup\ c \\ &\text{subject to } c \in \mathbb{R},\ g = f - c,\ g \geq 0. \end{aligned} \tag{1.25}$$

In the work by Rudi, Marteau-Ferey, and Bach (2020), we assume to have access to n function values of f at points $x_1, ..., x_n$ (these points will be sampled randomly). We will leverage smoothness in order to design an algorithm which approximate and solves Eq. (1.25). This method is related, in spirit, to the method of optimizing polynomials (Lasserre, 2001, 2010), which is based on the same convex reformulation.

In this introduction to the works by Rudi, Marteau-Ferey, and Bach (2020); Marteau-Ferey, Bach, and Rudi (2022b), we will therefore make a parallel with the works on polynomial optimization. Parallel is not a randomly chosen word : while there are structural similarities between the

problems (already highlighted in Sec. 1.1.4), as well as similar applications, the two methods are different on many aspects and have different behaviors and guarantees.

In the rest of this paragraph, we wil therefore define the general setting of optimization over measures, which is a general framework emcompassing both polynomial optimization and our contributions. We will then present a generic method to solve such problems for polynomials in Sec. 7.1 , developed by Lasserre (2010). In Sec. 7.2 , we will follow the same main steps as for the introduction of polynomial methods, to describe the method developed by Rudi, Marteau-Ferey, and Bach (2020). We will end Sec. 7.2 by presenting the work by Marteau-Ferey, Bach, and Rudi (2022b), which gives the theoretical basis to extend our method to different problems. The idea is to highlight the high level ingredients of these two methods, in terms of formulation, guarantees and algorithm, to provide a basis for comparison in Sec. 7.3 . The goal is to make a bridge between the works of the polynomial optimization community and our own, as our method offers the potential to go beyond polynomials, while also facing challenges which appear when the structure is not as rigid as that of polynomials. This comparison will have important repercussion in chapter 10 as it opens up the problems handled by polynomial optimization to other classes of functions and domains.

Optimization over measures

Let us present informally Eqs. (1.25) and (7.1) as optimization problems over measures, and their dual counterpart. Note that here, we do not strictly establish duality. Assume that $\mathcal{X}$ is equipped with a topology. In this chapter only, denote with $\mathscr{M}(\mathcal{X})$ the set of signed borel measures on $\mathcal{X}$ and with $\mathscr{M}_+(\mathcal{X})$ the set of non-negative signed measures (i.e. the set of finite measures). The "measure optimization" equivalent of Eq. (7.1) is :

$$\begin{aligned} &\inf \int_{\mathcal{X}} f(x)\mu(dx) \\ &\text{subject to } \mu \in \mathscr{M}_+(\mathcal{X}),\ \int_{\mathcal{X}} 1\ \mu(dx) = 1 \end{aligned} \tag{7.2}$$

This problem seems, of course, much harder than the original problem. However, it is convex and encompasses important information in the measure : if f is continuous, and has a non empty set Γ of minimizers, then the support of μ is concentrated on Γ, and in fact the set of measures minimizing Eq. (7.2) is exactly the set of probability distributions with support included in Γ. This fact will be crucial in Sec. 7.1 in order to extract the minimizer. The dual of Eq. (7.2) is simply Eq. (1.25) (see 1.2 by Lasserre (2010)) :

$$\begin{aligned} &\sup c \\ &\text{subject to } c \in \mathbb{R},\ g(x) = f(x) - c,\ g(x) \geq 0,\ \forall x \in \mathcal{X}. \end{aligned} \tag{1.25}$$

7.1 Polynomial optimization

In this section, we will present so-called moment-SOS (for moment sum of squares) hierarchies, to solve the global optimization problem for polynomials. This section follows the works and result by Lasserre (2010, 2001); Henrion, Korda, and Lasserre (2020), to name a few. Note that Parrilo (2003) also studied these kinds of hierarchies.

Notations. In this section, we denote with $\mathbb{R}[x_1, ..., x_d] = \mathbb{R}[\boldsymbol{x}]$, $\boldsymbol{x} = (x_1, ..., x_d)$ the set of polynomial functions on $\mathbb{R}^d$ with real coefficients, and with $\mathbb{R}_r[\boldsymbol{x}]$ the set of polynomial functions on $\mathbb{R}^d$ with degree at most r. Given a polynomial function f, we will denote with f_α the coefficient in front of the monomial x^α, and the sequence $(f_\alpha)_{\alpha \in \mathbb{N}^d}$, which has finite support, is identified with the polynomial.

We also define the set $\mathbb{N}^d_r$ wich is the set of $\alpha \in \mathbb{N}^d$ so that $|\alpha| = \sum_{i=1}^d \alpha_i \leq r$. Note that $\mathbb{N}^d_r$ is of size $s(r) = \binom{r+d}{d}$, and that coefficients of polynomials in $\mathbb{R}_r[x_1, ..., x_d]$ are naturally indexed by $\mathbb{N}^d_r$. We will often identify the set $\mathbb{R}^{\mathbb{N}^d_r}$ of families $(h_\alpha)_{|\alpha| \leq r}$ with the set of vectors in $\mathbb{R}^{s(r)}$. With a slight abuse of notations, we will therefore use the $\alpha \in \mathbb{N}^d_r$ to index such vectors, as well as matrices defined on $\mathbb{R}^{s(r)} \approx \mathbb{R}^{\mathbb{N}^d_r}$. Given $x \in \mathbb{R}^d$, we will denote with $v_r(x)$ the vector $(x^\alpha) \in \mathbb{R}^{s(r)}$, so that for any $q \in \mathbb{R}^{s(r)}$, it holds $q^\top v_r(x) = \sum_{|\alpha| \leq r} q_\alpha x^\alpha$.

As we will see in Sec. 7.1 .2, it will be interesting to study global optimization of polynomials over subsets $\mathbb{K}$ of $\mathbb{R}^d$. In particular, we will usually focus on semi-algebraic sets $\mathbb{K}$, *i.e.*, sets of the form $\mathbb{K} = \{x \in \mathbb{R}^d \ : \ f_j(x) \geq 0, \ 1 \leq j \leq m\}$ for polynomials f_j of degree $2r_j$ or $2r_j - 1$. In the semi-algebraic case, in order to lighten notations, we will not write the dependence of semi-algebraic sets in the f_j (even though the quantities and sets we define might depend on the f_j). If $\mathbb{K} = \mathbb{R}^d$, we consider that the canonical choice of f_j is the empty set of polynomial constraints.

In this section, great emphasis will be put on certain sets of non-negative polynomials. We will denote with $\mathcal{P}_+(\mathbb{K})$ the set of non-negative polynomials on a set $\mathbb{K} \subset \mathbb{R}^d$. We will also use the following notations:

- $\mathrm{SOS}[\boldsymbol{x}]$ will denote the set of sum of squares of polynomials of and $\mathrm{SOS}_r[\boldsymbol{x}]$ will denote the set of sum of squares of polynomials of degree at most r, *i.e.*,

$$\mathrm{SOS}_r[\boldsymbol{x}] := \left\{ f \in \mathbb{R}_{2r}[\boldsymbol{x}] \ : \ \exists N \in \mathbb{N}, \ \exists h_1, ..., h_N \in \mathbb{R}_r[\boldsymbol{x}], \ f = \sum_{i=1}^N h_i^2 \right\}; \tag{7.3}$$

- following the notations by Slot (2021), given $\mathbb{K}$ defined by inequalities $f_j \geq 0$ we denote with $\mathcal{Q}_r(\mathbb{K})$ the elements of degree less than $2r$ of the associated "quadratic module", *i.e.*, the set of polynomial functions of the form

$$\mathcal{Q}_r(\mathbb{K}) = \left\{ f \in \mathbb{R}_{2r}[\boldsymbol{x}] \ : \ \exists \sigma_0 \in \mathrm{SOS}_r[\boldsymbol{x}], \ \exists \sigma_j \in \mathrm{SOS}_{r-r_j}[\boldsymbol{x}], \ f = \sigma_0 + \sum_{i=1}^m f_j \sigma_j \right\}; \tag{7.4}$$

- finally, still following the notations by Slot (2021), we define $f_J = \prod_{j \in J} f_j$ for any subset $J \subset \{1, ..., m\}$ (note that the inequalities $f_J \geq 0$ still define $\mathbb{K}$) and with $\mathcal{T}_r(\mathcal{X})$ the set of elements of degree less than $2r$ of the associated "preordering", *i.e.*, the set of polynomial functions of the form

$$\mathcal{T}_r(\mathbb{K}) = \left\{ f \in \mathbb{R}_{2r}[\boldsymbol{x}] \ : \ \exists \sigma_0 \in \mathrm{SOS}_r[\boldsymbol{x}], \ \exists \sigma_J \in \mathrm{SOS}_{r_J}[\boldsymbol{x}], \ f = \sigma_0 + \sum_{J \subset \{1,...,m\}} f_J \sigma_J \right\}. \tag{7.5}$$

Note that we have the following inclusion, for any $r \in \mathbb{N}$:

$$\mathrm{SOS}_r[\boldsymbol{x}] \subset \mathcal{Q}_r(\mathbb{K}) \subset \mathcal{T}_r(\mathbb{K}) \subset \mathcal{P}_+(\mathbb{K}) \tag{7.6}$$

Outline. We will start by introducing the moment-SOS hierarchies for lower bounds in Sec. 7.1 .1, which can be seen either as a way of approximating the set of measures through its moments, or the set of non-negative polynomials by sums of squares. We will use this tool to to define surrogate problems for both the unconstrained and constrained polynomial minimizations problems in Sec. 7.1 .2, as well as give the guarantees for the solution of these surrogate problems. In Sec. 7.1 .3, we will briefly present another hierarchy presented by Lasserre (2011), whose aim is to provide upper bounds for the minimization problem, and which comes with theoretical guarantees. We conclude by a small summary in Sec. 7.1 .4.

7.1 .1 Moment relaxation, SOS strengthening

From measure to moments

When dealing with polynomials, a measure μ is only seen through moments, i.e. values of the form $\int_{\mathbb{R}^d} x^\alpha \mu(dx)$. As an example, if we take Eq. (7.2), one can write $\int_{\mathbb{R}^d} f(x)\mu(dx) = \sum_\alpha f_\alpha y_\alpha$, where $y_\alpha = \int_{\mathbb{R}^d} x^\alpha \mu(dx)$. The moment point of view on moment-SOS hierarchies focuses precisely on defining and approximating the set of moments sequences associated to measures, and not the measures themselves. An element $\boldsymbol{y} = (y_\alpha) \in \mathbb{R}^{\mathbb{N}^d}$ is the full moment sequence associated to μ if $y_\alpha = \int_{\mathbb{R}^d} x^\alpha \mu(dx),\ \alpha \in \mathbb{N}^d$. We denote with $M(\mathbb{K})$ the set of full moment sequences of measures μ with support in $\mathbb{K}$, *i.e.*,

$$M(\mathbb{K}) = \left\{ \boldsymbol{y} = (y_\alpha)_{\alpha \in \mathbb{N}^n} \ : \ \exists \mu \in \mathscr{M}(\mathbb{R}^d),\ \forall \alpha \in \mathbb{N}^d,\ y_\alpha = \int_{\mathbb{R}^d} x^\alpha \mu(dx),\ \mathrm{supp}(\mu) \subset \mathbb{K} \right\} \tag{7.7}$$

For a sequence $\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}^d}$, we define the linear form $L_{\boldsymbol{y}}$ on $\mathbb{R}[x]$ as $L_{\boldsymbol{y}}(f) = \sum_\alpha y_\alpha f_\alpha$, which is linear both in f and $\boldsymbol{y}$. The global optimization on problem in Eq. (7.2) for a polynomial f on a set $\mathbb{K}$ can simply be rewritten as

$$\begin{aligned} &\inf\ L_{\boldsymbol{y}}(f) \\ &\text{subject to } \boldsymbol{y} \in M(\mathbb{K}),\ \boldsymbol{y}_0 = 1, \end{aligned} \tag{7.8}$$

Which is a finite dimensional linear convex program (indeed, if f is of degree r, we can just optimize on the projection of $M(\mathbb{K})$ on its $|\alpha| \leq r$ first coefficients), but on a cone which does not seem to have any kind of finite dimensional characterization. This is in a sense exactly the same situation as in the dual problem

$$\begin{aligned} &\sup\ c \\ &\text{subject to } f - c \in \mathcal{P}_+(\mathbb{K}), \end{aligned} \tag{7.9}$$

where the cone $\mathcal{P}_+(\mathbb{K})$ of non-negative polynomials over $\mathbb{K}$ does not admit any nice finite-dimensional respresentation even when restricted ro $\mathbb{R}_r[\boldsymbol{x}]$.

Note that the cones $M(\mathbb{K})$ and $\mathcal{P}_+(\mathbb{K})$ are dual cones of each other with respect to the bilinear form $L : \boldsymbol{y}, f \mapsto L_{\boldsymbol{y}}(f)$. The idea of moment relaxation and SOS strengthening (which is the

dual operation) is to approximate these cones with cones which are representable as sections of PSD cones, and hence are representable in finite dimension, leading to a solvable semidefinite program.

Unconstrained measures and polynomials

Let us start with the simple case $\mathbb{K} = \mathbb{R}^d$. If $\boldsymbol{y} \in M(\mathbb{K})$, then necessarily, for any non-negative polynomial h, $L_{\boldsymbol{y}}(h) \geq 0$. In particular, this is true for all polynomial functions h wich are squares or sums of squares of polynomials, *i.e.*,

$$\forall r \in \mathbb{N},\ \forall h \in \mathbb{R}_r[\boldsymbol{x}],\ L_{\boldsymbol{y}}(h^2) \geq 0 \tag{7.10}$$

For $r \in \mathbb{N}$, define the moment matrix $\mathbf{M}_r(\boldsymbol{y}) = (y_{\alpha+\beta})_{\alpha,\beta\in\mathbb{N}^d_r} \in \mathbb{R}^{s(r)\times s(r)}$ to be the matrix associated to the bilinear form $h_1, h_2 \in \mathbb{R}_r[\boldsymbol{x}] \mapsto L_{\boldsymbol{y}}(h_1 h_2)$ in the canonical basis $(x^\alpha)_{\alpha\in\mathbb{N}^d_r}$. The previous remark shows exactly that if $\boldsymbol{y} \in M(\mathbb{R}^d)$, then $\mathbf{M}_r(\boldsymbol{y}) \succeq 0$ for all $r \geq 0$.

For $r \in \mathbb{N}$, define $M_r(\mathbb{R}^d) = \{\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}^d} \ :\ \mathbf{M}_r(\boldsymbol{y}) \succeq 0\}$. This set is PSD representable (it can be represented as a linear subset of a PSD cone), and satisfies $M_r(\mathbb{R}^d) \supset M(\mathbb{R}^d)$: it is therefore a *PSD relaxation* of the cone $M(\mathbb{R}^d)$. Moreover, these relaxations for an outer approximation sequence of the moment cone, *i.e.*,

$$M_0(\mathbb{R}^d) \supset ..... \supset M_r(\mathbb{R}^d) \supset ... \supset M(\mathbb{R}^d). \tag{7.11}$$

A dual perspective provides the so-called SOS strengthenings. Indeed, the dual of $M_r(\mathbb{R}^d)$ with respect to L can be shown to be the set of sum of squares of polynomials of degree less than r, defined in Eq. (7.3) and which we denote with $\mathrm{SOS}_r[\boldsymbol{x}]$.Rewriting the definition of this cone as

$$\mathrm{SOS}_r[\boldsymbol{x}] = \{f \in \mathbb{R}_{2r}[\boldsymbol{x}],\ \exists A \in \mathbb{S}_+(\mathbb{R}^{s(r)}),\ f(\cdot) = v_r(\cdot)^\top A v_r(\cdot)\}, \tag{7.12}$$

it is clear that it is a PSD representable cone. Moreover, the cones $\mathrm{SOS}_r[\boldsymbol{x}] \subset \mathcal{P}_+(\mathbb{R}^d)$, as sums of squares are always non-negative, and forms an inner approximation of $\mathcal{P}_+(\mathbb{R}^d)$, *i.e.*,

$$\mathrm{SOS}_0[\boldsymbol{x}] \subset ..... \subset \mathrm{SOS}_r[\boldsymbol{x}] \subset ... \subset \mathcal{P}_+(\mathbb{R}^d). \tag{7.13}$$

As we will see in Sec. 7.1 .2, these unconstrained approximation sequences are usually too weak, in the sense that they do not approximate $M(\mathbb{R}^d)$ and $\mathcal{P}_+(\mathbb{R}^d)$ well enough at infinity. Indeed, if $\mathrm{SOS}[\boldsymbol{x}] = \bigcup_{r\geq 0} \mathrm{SOS}_r[\boldsymbol{x}]$ is the set of SOS polynomials, $\mathrm{SOS}[\boldsymbol{x}]$ is not "dense" in $\mathcal{P}_+(\mathbb{R}^d)$ for certain metrics of interest, such as the uniform convergence, needed in global optimization (Lasserre, 2010). Note that in the one dimensional case, the uncontrained SOS approximation sequence and hence both approximation sequences are tight : every polynomial which is non-negative can be written as a sum of squares of polynomials.

Constrained measures and polynomials

In the case of constrained measures, which is the most common in practice, the necessary condition Eq. (7.10) is not strong enough as it does not incorporate the algebraic inequalities f_j which define the set $\mathbb{K}$. Let $\boldsymbol{y} \in M(\mathbb{K})$. It is clear that for any $1 \leq j \leq m$, if h is non-negative on $\{x \in \mathbb{R}^d \ :\ f_j(x) \geq 0\}$, then $L_{\boldsymbol{y}}(h) \geq 0$. In particular, since for any polynomial h, the polynomial $f_j h^2$ is non-negative on $\{x \in \mathbb{R}^d \ :\ f_j(x) \geq 0\}$, the following sufficient conditions hold

$$\forall 1 \leq j \leq m,\ \forall r \in \mathbb{N},\ \mathbf{M}_r(f_j \boldsymbol{y}) \succeq 0. \tag{7.14}$$

where $\mathbf{M}_r(f_j\boldsymbol{y})$ is called the *localizing matrix* associated to f_j, and is associate to the bilinear form $h_1, h_2 \in \mathbb{R}_r[\boldsymbol{x}] \mapsto L_{\boldsymbol{y}}(f_j h_1 h_2)$. Note that in that case, $\mathbf{M}_r(f_j\boldsymbol{y})_{\alpha,\beta} = \sum_\gamma f_\gamma y_{\alpha+\beta+\gamma}$

Following Lasserre (2010), for $r \geq r_0 = \max_{1\leq j\leq m}(r_j)$, define

$$M_r(\mathbb{K}) := \{\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}^d} \ : \ \mathbf{M}_r(\boldsymbol{y}) \succeq 0, \ \mathbf{M}_{r-r_j}(f_j\boldsymbol{y}) \succeq 0, \ 1 \leq j \leq m\}. \tag{7.15}$$

The set $M_r(\mathbb{K})$ is still PSD representable (see the expression of the localization matrices), and are the sets of moments such that the necessary condition Eqs. (7.10) and (7.14) hold with certain degree. They are still relaxations of $M(\mathbb{K})$, and we have the following outer approximation of the moment cone:

$$M_0(\mathbb{K}) \supset ..... \supset M_r(\mathbb{K}) \supset ... \supset M(\mathbb{K}). \tag{7.16}$$

A dual perspective provides the constrained SOS strengthening and approximation sequence. In this case, the dual of $M_r(\mathbb{R}^d)$ with respect to L can be shown to be the set $\mathcal{Q}_r(\mathbb{K})$ defined in Eq. (7.4). These cones are still PSD representable and are strenghtenings of the cone $\mathcal{P}_+(\mathbb{K})$. They form an inner approximation sequence of $\mathcal{P}_+(\mathbb{K})$, *i.e.*,

$$\mathcal{Q}_0(\mathbb{K}) \subset ..... \subset \mathcal{Q}_r(\mathbb{K}) \subset ... \subset \mathcal{P}_+(\mathbb{K}). \tag{7.17}$$

A note on preorderings

Since $\mathrm{SOS}_r[\boldsymbol{x}] \subset \mathcal{Q}_r(\mathbb{K})$, it is clear that we will better approximate non-negative polynomials on $\mathbb{K}$ using the inner approximation sequence given by the $\mathcal{Q}_r(\mathbb{K})$ than the one given by $\mathrm{SOS}_r[\boldsymbol{x}]$. This fact will be made formal in the next section, where we will see that the guarantees obtained using this approximation sequence are much stronger than the one obtained using the simple sequence of SOS polynomials.

Another approximation of non-negative polynomials on $\mathbb{K}$ is used and analysed in the literature (see (Lasserre, 2010; Slot and Laurent, 2020a; Slot, 2021), based on the notion of *preordering*. Instead of considering polynomials of the form $\sigma_0 + \sum_{j=1}^m f_j \ \sigma_j$ for sum of squares polynomials σ_j, polynomials of the form $\sum_{J\subset\{1,...,m\}} f_J \ \sigma_J$ are considered, where $f_J = \prod_{j\in J} f_j$ and the σ_J are sum of squares. For instance, Slot (2021) define $\mathcal{T}_r(\mathbb{K})$ as the set of all such polynomials such that $\deg(f_J \ \sigma_J) \leq 2r$ (see Eq. (7.5) for a formal definition). In a sense, this is just the quadratic module $\mathcal{Q}_r(\mathbb{K})$ but associated to the augmented family $(f_J)_{J\subset\{1,...,m\}}$ of inequalities which define the same semi-algebraic set $\mathbb{K}$. Of course, this family has 2^m instead of $m+1$ terms in the sum and we can see that $\mathcal{T}_r(\mathbb{K})$ is a larger set of non-negative functions than $\mathcal{Q}_r(\mathbb{K})$. Once again, the $\mathcal{T}_r(\mathbb{K})$ form an inner approximation of the set of non-negative functions :

$$\mathcal{T}_0(\mathbb{K}) \subset ..... \subset \mathcal{T}_r(\mathbb{K}) \subset ... \subset \mathcal{P}_+(\mathbb{K}). \tag{7.18}$$

The associated moments' point of view can of course be derived by considering localizing matrices for all $(f_J)_{J\subset\{1,...,m\}}$.

Apart from cases where the number of inequalities is very low, it is too costly to use these inner approximations of the cone $\mathcal{P}_+(\mathbb{K})$ in practice (as well as the moments' counterpart), as the number of PSD constraints exponentially increases in m. However, this approximation sequence has been more analyzed from a theoretical standpoint, and many guarantees can be obtained by considering this augmented sequence rather than $\mathcal{Q}_r(\mathbb{K})$, even though it is the one most used in practice.

Finally, note that $\mathcal{Q}_r(\mathbb{K})$ and $\mathcal{T}_r(\mathbb{K})$ coincide if $\mathbb{K}$ is only defined with one inequality (such as ball of fixed radius).

7.1 .2 Moment-SOS hierarchies of lower bounds for optimization, and their guarantees

The case of unconstrained optimization

Let us now consider the approximations by PSD cones described above in order to solve the unconstrained polynomial minimization problem Eqs. (7.8) and (7.9). Let f be a polynomial of degree $2r$ (note that if the degree of the polynomial is odd, the minimization problem trivially has solution $-\infty$), and let f_* be its minimum on $\mathbb{R}^d$.

We define the sequence of problems :

$$\begin{aligned} \rho_{r'} = &\inf \; L_{\boldsymbol{y}}(f) \\ &\text{subject to } \mathbf{M}_{r'}(\boldsymbol{y}) \succeq 0, \boldsymbol{y}_0 = 1 \end{aligned} \tag{7.19}$$

where we have relaxed the condition that $\boldsymbol{y} \in M(\mathbb{R}^d)$ with the condition that $\boldsymbol{y} \in M_{r'}(\mathbb{R}^d)$. The dual counterparts of these problems are

$$\begin{aligned} \rho^*_{r'} = &\sup \; c \\ &\text{subject to } f - c \in \mathrm{SOS}_{r'}[\mathcal{X}], \end{aligned} \tag{7.20}$$

where we have strengthened the condition that $f - c \in \mathcal{P}_+(\mathbb{R}^d)$ with the condition that $f - c \in \mathrm{SOS}_{r'}(\mathbb{R}^d)$.

The sequences $\rho_{r'}$ and $\rho^*_{r'}$ are *a)* increasing sequences (see Eq. (7.13) and Eq. (7.11)), *b)* are all *lower bounds* of f_*, and *c)* satisfy $\rho_{r'} = \rho^*_{r'}$, that is there is no duality gap (Lasserre, 2010).

In the dual problem, note that if the constraint is satisfied, necessarily, $f - c \in \mathrm{SOS}_r[\boldsymbol{x}]$. Thus, there is no need for the full sequence of problems in the unconstrained case, the case $r' = r$: there is actually **no hierarchy** in the unconstrained case. To solve Eq. (7.19) with $r' = r$, Lasserre (2010) uses interior point methods (Nesterov and Nemirovskii, 1994).

Once an optimal ρ_r is obtained (and we know that $\rho_r \leq f_*$), two questions remain.

(a) Can we say that $\rho_r = f_*$?

(b) Can we find a minimizer $x_* \in \mathbb{R}^d$ such that $f(x_*) = f_*$?

The following result provides a sufficient condition, based on the "stabilization of the rank of the moment matrix", in order to guarantee that $\rho_r = f_*$ and that a minimizer x_* can be extracted.

Theorem 7.1 (Lasserre (2010), Theorem 5.5)**.** *Let ρ_r be a solution to Eq. (7.19) if there is an optimal solution $\boldsymbol{y}_*$, and if $\mathrm{rank}(\mathbf{M}_r(\boldsymbol{y}_*)) = \mathrm{rank}(\mathbf{M}_{r-1}(\boldsymbol{y}_*))$, then $\rho_r = f_*$, $f - f_*$ is a sum of squares, an a minimizer of f can be extracted using Algorithm 4.2 by Lasserre (2010).*

Unfortunately however, unconstrained minimization for polynomials suffers from the following alternative :

(i) if the rank of the moment matrix stagnates, in which case $f - f_*$ is a sum of squares polynomial, $\rho_r = f_*$, and a minimizer can be extracted;

(ii) if the moment stagnation property is not satisfied, and we obtain a lower bound ρ_r but without any additional information on f_* or a minimizer.

Note that since only sum of squares can hope to satisfy the moment stagnation property, all non-negative polynomials which are not sums of squares (and they exist, see Chapter 2 by Lasserre (2010)) are excluded from the first point, and will not be minimized. Since there is no guarantee *a priori* that a non-negative polynomial is a sum of squares, there is no *a priori* guarantee that ρ_r will be close to f_*.

The reason for this is the lack of a proper Positivstellensatz for the unconstrained case, that is we cannot prove that there exists a sequence $c_n \uparrow f_*$ such that $f - c_n$, which is stricly positive on $\mathbb{R}^d$, can be decomposed as a sum of squares. While such results do not exist for general polynomials (except in the $d = 1$ case), they exists under certain mild assumptions for the constrained case. Moreover, in practice, we usually look for the minimizer in a large enough ball, and adding a constraint of the form $B^2 - \sum_{i=1}^d x_i^2 \geq 0$ is enough to make these assumptions hold (it is sometimes called the *archimedean* assumption).

Finally, note that in the one dimensional case, all non-negative polynomials are sum of squares, and this method works very well.

Constrained optimization and moment-SOS hierarchies

Let us now consider the moment-SOS approximation sequences in Sec. 7.1 .1 to solve the constrained polynomial minimization problem, and show that there are better guarantees in the constrained setting than in the unconstrained one. Let f be a polynomial function, and $r = \lfloor \deg(f)/2 \rfloor$ such that that f is either of degree $2r$ or $2r - 1$. Let $\mathbb{K}$ be defined by the polynomial inequalities $f_j \geq 0$ for $1 \leq j \leq m$, and $r_j = \lfloor \deg(f_j)/2 \rfloor$ such that f_j is of degree either $2r_j$ or $2r_j - 1$. Let $f_{\mathbb{K}}$ be the minimum of f on $\mathbb{K}$. Let $r_0 = \min_j(r_j)$. For any $r' \geq r_0$, we define the sequence of moment problems

$$\begin{aligned} \rho_{r'} &= \inf\ L_{\boldsymbol{y}}(f) \\ \text{subject to } &\mathbf{M}_{r'}(\boldsymbol{y}) \succeq 0, \mathbf{M}_{r'-r_j}(f_j \boldsymbol{y}) \succeq 0,\ \boldsymbol{y}_0 = 1 \end{aligned} \tag{7.21}$$

where we have relaxed the condition that $\boldsymbol{y} \in M(\mathbb{K})$ with the condition that $\boldsymbol{y} \in M_{r'}(\mathbb{K})$. The dual counterpart of this sequence of moment problems is the following sequence of problems,

$$\begin{aligned} \rho^*_{r'} &= \sup\ c \\ \text{subject to } & f - c = \sigma_0 + \sum_{j=1}^m \sigma_j f_j, \\ & \sigma_0 \in \mathrm{SOS}_{r'}[\boldsymbol{x}],\ \sigma_j \in \mathrm{SOS}_{r'-r_j}[\boldsymbol{x}],\ 1 \leq j \leq m, \end{aligned} \tag{7.22}$$

where we have strengthened the condition that $f - c \in \mathcal{P}_+(\mathbb{K})$ with the condition that $f - c \in \mathcal{Q}_{r'}(\mathbb{K})$.

Once again, Eqs. (7.11) and (7.13) guarantee that $\rho_{r'}$ and $\rho^*_{r'}$ are both increasing sequences. Moreover, we have

$$\forall r' \geq r_0,\ \rho^*_{r'} \leq \rho_{r'} \leq f_{\mathbb{K}}. \tag{7.23}$$

These problems are often referred to as the (Lasserre) **moment-SOS hierarchies** (of lower bounds), as they form a sequence of problems which are both increasingly harder and more precise.

The following algorithm (Lasserre, 2010) can be applied to attempt to minimize f, with a parameter R which is the highest order of the hierarchy we are prepared to go to. It is based on the equivalent of Theorem 7.1 in the constrained case, in order to establish if $\rho_{r'} = f_{\mathbb{K}}$ (Lasserre (2010), Theorem 5.7).

- Solve the problem Eq. (7.21) and obtain $\rho_{r'}$.
- If there is no optimal $\boldsymbol{y}_*$, then $\rho_{r'} \leq f_{\mathbb{K}}$; if $r' < R$, we set $r' \leftarrow r' + 1$
- If there is an optimal $\boldsymbol{y}_*$, and $\operatorname{rank}(\mathbf{M}_r(\boldsymbol{y})) > \operatorname{rank}(\mathbf{M}_{r-r_0}(\boldsymbol{y}))$ then $\rho_{r'} \leq f_{\mathbb{K}}$; if $r' < R$, we set $r' \leftarrow r' + 1$
- If there is an optimal $\boldsymbol{y}_*$ and $\operatorname{rank}(\mathbf{M}_r(\boldsymbol{y})) = \operatorname{rank}(\mathbf{M}_{r-r_0}(\boldsymbol{y}))$, then $f_{\mathbb{K}} = \rho_{r'}$ and one can extract the minimizer using algorithm 4.2 by Lasserre (2010).

Guarantees of the moment-SOS hierarchies

The key difference with the unconstrained case is that the moment-SOS hierarchy converges to the actual minimizer $f_{\mathbb{K}}$. There are essentially two types of results describing this convergence.

- The first results (which were also the first resutls obtained on this subject) describe assumptions under which the hierarchy converges (*i.e.*, $\rho_{r'} \uparrow f_{\mathbb{K}}$).
- The second results are more precise and describe the speed of convergence of the hierarchy towards $f_{\mathbb{K}}$ as a functions of the order r' of the hierarchy. They can have stronger assumptions in than the ones needed for the convergence, in order to prove faster rates. Note that most of these results are obtained using the pre-ordering approximation Eq. (7.5) and is not exactly the hierarchy described in Eq. (7.21). We mention them more precisely in the paragraph below.

Convergence results. Let (A) denote Assumption 2.1 by Lasserre (2010). It is satisfied in many cases, and can be enforced as soon as the set $\mathbb{K}$ is localized inside a ball of known radius $B > 0$ (or in the unconstrained case, as soon as we know we are looking for the minima of f in a ball of given radius). In that case, adding the constraint $B^2 - \sum_{i=1}^d x_i^2 \geq 0$ directly guarantees that (A) is satisfied. The following theorem is the cornerstone to prove the (asymptotic) convergence of the Lasserre moment-SOS hierarchy.

Theorem 7.2 (Putinar Positivstellensatz, (Putinar, 1993))**.** *Under assumption (A) on $\mathbb{K}$, any strictly positive polynomial belongs to $\mathrm{SOS}_{r'}(\mathbb{K})$ for some $r' \in \mathbb{N}$, i.e., is of the form*

$$\begin{aligned} &\sigma_0 + \sum_{j=1}^m \sigma_j f_j, \\ &\sigma_0 \in \mathrm{SOS}_{r'}[\boldsymbol{x}],\ \sigma_j \in \mathrm{SOS}_{r'-r_j}[\boldsymbol{x}],\ 1 \leq j \leq m. \end{aligned} \tag{7.24}$$

Indeed, once such a result is established, then for any $c < f_*$, $f - c$ is strictly positive and hence belongs to a certain $\mathrm{SOS}_{r'}(\mathbb{K})$ and thus $\rho_{r'}^* \geq c$, showing that $\rho_r^* \uparrow f_{\mathbb{K}}$ and hence also $\rho_r \uparrow f_{\mathbb{K}}$ (see Lasserre (2010), theorem 4.1).

Under the assumption that f has a unique minimizer on $\mathbb{K}$ as well as a few additional assumptions, it is also possible to design a sequence $x_{r'} \to x_*$ from the hierarchy solutions $(\boldsymbol{y}_*)_{r'}$, if they exist (see Lasserre (2010), Theorem 5.6).

Under stronger assumptions than the ones above, *finite convergence* of the Lasserre hierarchy has been obtained by Nie (2014) in Theorem 1.1. These assumptions are based on first and second order sufficiency conditions at the minimizers, and are analoguous to constraint qualification assumptions in non-linear optimization (Gilbert, 2020), developped for polynomials by Marshall (2006). Under these assumptions, one obtains that $f - f_{\mathbb{K}} \in \mathcal{Q}_{r'}(\mathbb{K})$ for some large enough r'_0 (although no bound is provided on the size of r'_0), and hence $\rho_{r'} = f_{\mathbb{K}}$ for all $r' \geq r_0$.

Other results on finite convergence of the Lasserre hierarchy have been established for certain classes of polynomials. For instance, Lasserre (2009) proves finite convergence strictly convex polynomials and gives the order of convergence of the Hierarchy for SOS-convex polynomials (that is polynomials whose Hessians are sum of squares, as defined by Helton and Nie (2010)). The strict convexity assumption is has been weakened in further works, as the one by Klerk and Laurent (2011).

Rates of convergence. Under assumption (A), Nie and Schweighofer (2007) proved that the convergence $\rho_r \uparrow f_{\mathbb{K}}$ happens at a rate of order $O(1/\log(r')^c)$, where c is a constant which depends on $\mathbb{K}$.

More recently, better convergence bounds have been provided in specific cases, such as when $\mathbb{K}$ is the unit ball, as stated below.

Theorem 7.3 (Theorem 3 by Slot (2021))**.** *Assume $\mathbb{K}$ is the d-dimensional unit ball and let f be a polynomial of degree r. Then for any $r' \geq 2rd$, we have*

$$f_{\mathbb{K}} - \rho_{r'} \leq \frac{C_{r,d}}{(r')^2} \left(\max_{\mathbb{K}} f - \min_{\mathbb{K}} f \right) \tag{7.25}$$

Guarantees using the preordering approximation sequence

It is important to note that many convergence rates obtained for the Lasserre hierarchy have been obtained in a setting slightly different than the hierarchy presented in Eq. (7.22). Instead, works such as those by Laurent and Slot (2021); Slot (2021) derive rates of convergence for the hierarchy defined by the "preordering" approximation sequence in Eq. (7.18) (*i.e.,* replacing the constraint that $f - c \in \mathcal{Q}_{r'}(\mathbb{K})$ by $f - c \in \mathcal{T}_{r'}(\mathbb{K})$ in Eq. (7.22)).

A summary of those results can be found in Table 1. of the work by Slot (2021). In particular, rates of order $O(1/(r')^2)$ such as the one reported in Theorem 7.3 are obtained for both the simplex $\mathbb{K} = \Delta_{d-1} = \{x \in \mathbb{R}^d \ : \ x \geq 0, \ x^\top \mathbf{1} = 1\}$ and the hypercube $\mathbb{K} = [-1, 1]^d$.

7.1 .3 A hierarchy of upper bounds

In this section, we briefly present another approximating sequence of the cone of non-negative polynomials on a set $\mathbb{K}$, introduced by Lasserre (2011). This sequence forms an outer approximation and leads to a hierarchy of upper bounds for the minimum. It also comes with theoretical guarantees on the convergence and rate of convergence of the hierarchy.

Outer approximation of the cone of non-negative polynomials

For simplicity, let $\mathbb{K}$ be a compact set of $\mathbb{R}^d$ (for generalizations to non-compact sets, see the original paper by Lasserre (2011), where one can deal with the non-compact case with a slight modification). Note that here, $\mathbb{K}$ does not have to be a semi-algebraic set. The outer approximation of the cone of non-negative polynomials relies on a **fixed reference measure** μ,

whose support is $\mathbb{K}$. Given this reference measure, one considers the following sets for $r' \in \mathbb{N}$:

$$\mathcal{C}^r_{r'}(\mathbb{K}) = \left\{ f \in \mathbb{R}_{2r}[\boldsymbol{x}] \ : \ \forall q \in \mathrm{SOS}_{r'}[\boldsymbol{x}], \ \int_{\mathbb{K}} q(x) f(x) \ d\mu(x) \right\} \tag{7.26}$$

Once again, this is an abuse of notation because this set depends not only on $\mathbb{K}$ but on μ. Moreover, this set can be defined for any function f and not only for a polynomial. It is clear that we have nested sequence of outer approximations approximations of the set of non-negative polynomials on $\mathbb{K}$ of degree at most $2r$ which we write $\mathcal{C}^r(\mathbb{K})$.

$$\mathcal{C}^r_0(\mathbb{K}) \supset \ldots \supset \mathcal{C}^r_{r'}(\mathbb{K}) \supset \ldots \supset \mathcal{C}^r(\mathbb{K}). \tag{7.27}$$

Note that $\mathcal{C}^r_{r'}(\mathbb{K})$ can also be written using a localization matrix (see below Eq. (7.14) for the definition),

$$\mathcal{C}^r_{r'}(\mathbb{K}) = \left\{ f \in \mathbb{R}_r[\boldsymbol{x}] \ : \ M_{r'}(fy) \succeq 0, \ y_\alpha = \int_{\mathbb{K}} x^\alpha \mu(dx) \right\} \tag{7.28}$$

This means that if the moments (y_α) are computable for the measure μ, we can have a computable PSD representation of the cone $\mathcal{C}^r_{r'}(\mathbb{K})$.

Hierarchy of upper bounds

Given a polynomial function $f \in \mathbb{R}_{2r}[\boldsymbol{x}]$, if we relax Eq. (7.9) by replacing $f - c \in \mathcal{P}_+(\mathbb{K})$ by $f - c \in \mathcal{C}^r_{r'}(\mathbb{K})$, we get the following hierarchy of upper bounds indexed by r' :

$$\begin{aligned} \lambda_{r'} = & \sup \ c \\ & \text{subject to } f - c \in \mathcal{C}^r_{r'}(\mathbb{K}). \end{aligned} \tag{7.29}$$

Note that this correspond to the dual problem in the previous sections. The dual of Eq. (7.29), which corresponds to the primal problem in the previous sections (*i.e.*, a moment problem), can be written as

$$\begin{aligned} \lambda^*_{r'} = & \inf \int_{\mathbb{K}} f \sigma d\mu \\ & \text{subject to } \int_{\mathbb{K}} \sigma d\mu = 1, \ \sigma \in \mathrm{SOS}_{r'}[\boldsymbol{x}]. \end{aligned} \tag{7.30}$$

Theorem 4.2 by Lasserre (2011) shows in particular that if $f_{\mathbb{K}} > -\infty$ and $\mathbb{K}$ has non-empty interior, there is no duality gap (and hence $\lambda_{r'} = \lambda^*_{r'}$). As explained by Lasserre (2011), these optimization problems correspond to a generalized eigenvalue problem between the localizing matrices $M_{r'}(y)$ and $M_{r'}(fy)$. Moreover, this hierarchy also comes with a way of approximating the minimizer, when $\mathbb{K}$ is convex (see Theorem 4.2 by Lasserre (2011)).

Convergence guarantees

In terms of convergence guarantees, the moment-SOS hierarchy of upper bounds comes with many results. The first, which can be found as Theorem 4.1 by Lasserre (2011), establishes the convergence $\lambda_{r'} \downarrow f_{\mathbb{K}}$.

Many convergence rates have been established, and are summarized in Table 2 by Slot (2021). Since the choice of μ defines the approximation, these results depend on the choice of μ In particular, Slot and Laurent (2021) show that under very mild assumptions on $\mathbb{K}$ (such as convexity, or being a semi-algebraic set with dense interior), if μ is the Lebesgue measure, then the convergence rate is of order $O(\log(r')^2/(r')^2)$. For specific cases like the unit ball, the unit simplex or the unit hypercube, $O(1/(r')^2)$ rates have also been shown by Slot and Laurent (2020b).

7.1 .4 Summary

In this section, we have presented different moment-SOS hierarchies in order to solve global polynomial optimization on unconstrained and constrained domains.

We started by presenting the first moment-SOS hierarchy which provides lower bounds on the global optimum. It based on a series of moment relaxations or SOS strengthenings (depending on the primal or dual point of view) , which consist in approximating the set moments or non-negative polynomials on $\mathbb{K}$ by a sequence of PSD representable cones, on which the associated moment or SOS problems can be optimized.

Note that using interior point methods, the complexity of optimizing the hierarchy until degree r' is roughly or order $d^{r'}$ (this complexity can be reduced using sparsity, see Lasserre (2010), section 4.6.2).

We have seen that the proposed methods return either *a)* a minimizer and the minimum under a rank condition, or *b)* a lower bound $\rho_{r'}$ for $f_{\mathbb{K}}$ otherwise.

In the unconstrained case, if a lower bound is returned, there is no guarantee on its proximity to f_*.

In the constrained setting (which is usually naturally the case), under very mild assumption, it can be shown that the hierarchy asymptotically converges, that is $\rho_{r'} \uparrow f_{\mathbb{K}}$ (finite convergence can also be shown in some convex or locally convex cases). Moreover, convergence rates of order $1/\log(r')^c$ and $1/(r')^2$ in certain specific settings can be proven.

We then presented a second moment-SOS hierarchy of upper bounds, based on the choice of a base measure on $\mathbb{K}$. This moment-SOS hierarchy enjoys very nice convergence properties (close to $O(1/(r')^2)$ in many settings). However, it is less obvious to use in practice, as one must compute the moments of the base measure.

In terms of complexity of minimizing a polynomial function on a subset of $\mathbb{R}^d$ with error ϵ, we would need a degree in the hierarchy of order $1/\sqrt{\varepsilon}$, which leads to a crude estimation of the complexity of order $d^{1/\sqrt{\epsilon}}$, and which is therefore exponential in $1/\sqrt{\varepsilon}$.

7.2 Global optimization through kernel sums of squares

In this section, we present the work by Rudi, Marteau-Ferey, and Bach (2020), which proposes an algorithm to perform global optimization through sums of squares of kernels. This method is

analyzed in the context of approximating a function f with a given regularity.

Here, we take a different approach from the paper, whose verbatim can be found in chapter 8. We start by presenting the method generally, without guarantees, in Sec. 7.2 .1, before applying it in the setting considered by Rudi, Marteau-Ferey, and Bach (2020) where guarantees can be obtained, in Sec. 7.2 .2.

Notations. In this section, we will consider a RKHS $\mathcal{H}$ on a set $\mathcal{X}$ associated to the kernel k, whcih we will assume to be separable (*i.e.*, there is a dense at most countable sequence in $\mathcal{H}$). We denote with $\mathcal{S}(\mathcal{H})$ the set of bounded symmetric operators on $\mathcal{H}$. In particular, in this section, we will consider the set $\mathcal{S}_1(\mathcal{H})$ of trace class symmetric operators on $\mathcal{H}$ (Weidmann, 1980) equipped with the trace norm (note that we could take other spectral norms, such as the Hilber-Schmidt norm, as explained in chapter 5). We will denote with $\mathcal{S}_{1,+}(\mathcal{H})$, or simply $\mathcal{S}_+(\mathcal{H})$ (the norm considered will always be the trace norm in this section), the set of postivie semidefinite trace-class symmetric operators on $\mathcal{H}$.

7.2 .1 A generic methods based on kernel sum of squares

In this section, we present the main ingredients of the method developed by Rudi, Marteau-Ferey, and Bach (2020). We present these ingredients in full generality, without assuming that $\mathcal{X} = \mathbb{R}^d$ for example, but without providing any guarantees at this stage.

The method relies on three main steps.

(i) As in the polynomial case, we start by considering a kernel equivalent of the moment-SOS relaxation.

(ii) We then consider a regularized discretization of this relaxation, in the spirit of empirical risk minimization.

(iii) Finally, using the kernel representer theorem, we effectively solve this regularized discretized version.

Primal and dual relaxations

The first step is to consider the following primal and dual relaxations of the measure problem Eqs. (1.25) and (7.2). It is easier to see things in the dual : we replace the positivity constraint $f - c \geq 0$ by the constraint that $f - c$ be a PSD model introduced in part II. This is done in Eq. (8.3) in the work by Rudi, Marteau-Ferey, and Bach (2020).

To align with the notations of the previous section, define

$$\begin{aligned} \mathrm{SOS}_0^{\mathcal{H}} &= \{f : \mathcal{X} \to \mathbb{R} \ : \ f(x) = \langle k_x, Ak_x \rangle_{\mathcal{H}}, \ x \in \mathcal{X}, \ A \in \mathcal{S}_+(\mathcal{H}), \ \mathrm{rank}(A) < \infty\}, \\ \mathrm{SOS}^{\mathcal{H}} &= \{f : \mathcal{X} \to \mathbb{R} \ : \ f(x) = \langle k_x, Ak_x \rangle_{\mathcal{H}}, \ x \in \mathcal{X}, \ A \in \mathcal{S}_+(\mathcal{H})\}. \end{aligned} \tag{7.31}$$

Note that as explained in chapter 5, $\mathrm{SOS}_0^{\mathcal{H}}$ is simply the set of finite sums of squares of functions in $\mathcal{H}$, while $\mathrm{SOS}^{\mathcal{H}}$ is its completion for the trace norm, defined as $\|f\|_1 = \min\{\mathrm{Tr}(A) \ : \ f(x) = \langle k_x, Ak_x \rangle_{\mathcal{H}}, \ x \in \mathcal{X}, \ A \in \mathcal{S}_+(\mathcal{H})\}$. Strengthening the constraint $f - c \geq 0$ by $f - c \in \mathrm{SOS}^{\mathcal{H}}$, we get the following strengthening of the dual :

$$\begin{aligned} &\rho^*_{\mathcal{H}} = \sup \ c \\ &\text{subject to } f(x) - c = g(x), \ x \in \mathcal{X}, \ g \in \mathrm{SOS}^{\mathcal{H}} \end{aligned} \tag{7.32}$$

This corresponds to the following relaxation of the primal measure problem :

$$\begin{aligned} \rho_{\mathcal{H}} &= \inf \int_{\mathcal{X}} f(x)\mu(dx) \\ \text{subject to } \mu \in \mathscr{M}(\mathcal{X}),\ &\int_{\mathcal{X}} \mu(dx) = 1,\ \int_{\mathcal{X}} k_x \otimes_{\mathcal{H}} k_x\ \mu(dx) \succeq 0, \end{aligned} \tag{7.33}$$

where the positivity constraint of the measure is relaxed with the PSD constraint $\int_{\mathcal{X}} k_x \otimes_{\mathcal{H}} k_x\ \mu(dx) \succeq 0$ on signed measures.

As these problems are relaxations/strengthenings of the original problem, the following guarantees hold :

$$\rho^*_{\mathcal{H}} \leq \rho_{\mathcal{H}} \leq f_* = \inf_{x \in \mathcal{X}} f(x). \tag{7.34}$$

Note on polynomials. As we will see in Sec. 7.3 , the relaxation presented above corresponds exactly to the polynomial unconstrained moment-SOS relaxations, if we consider the RKHS $\mathcal{H} = \mathbb{R}_r[\boldsymbol{x}]$ (the RKHS structure can easily be obtained, see Sec. 7.3).

Note on moments. It is not as obvious to define the moment problem, as in the polynomial case, since we have assumed nothing on f at this point. It is possible to formulate the moment problem if f is assumed to belong in a certain space, as we will see in Sec. 7.3 . For the remaining of this section, we will keep to the measure problem as primal problem, without simplifying it as a moment problem as used to be the case in the polynomial setting.

Note on tightness. As we will see when applying this method with the Sobolev kernel in Sec. 7.2 .2, in the infinite dimensional RKHS setting (non-parametric), it is easier to show that Eq. (7.32) are actually tight, *i.e.*, $\rho_{\mathcal{H}} = \rho^*_{\mathcal{H}} = f_*$. However, as we will see in the next section, it will be crucial to have more than this type of tightness, and to have an *optimal solution* in the dual case Eq. (7.32) (*i.e.*, the existence of c_*, g_* which solve the problem and reach the supremum), in order to get good properties of the method, as we do not solve Eq. (7.32) directly, but approximate it. This corresponds to a well-specified assumption in part I, as we will take the same "regularized e.r.m." approach to solve the problem Eq. (7.32).

Evaluating the constraint on a finite number of points

While these primal and dual relaxation are PSD representable, they are not finite dimensional, because of the fact that $\mathcal{H}$ is potentially infinite dimensional. In a second step, we therefore adopt the same strategy as when performing empirical risk minimization, and instead of considering the equality in Eq. (7.32) to hold at every point of $\mathcal{X}$, we restrict it to hold only a a finite number of points. To that end, we give ourselves a sequence $(x_i)_{1 \leq i \leq n} \in \mathcal{X}^n$ of test points, corresponding to a subset $\mathcal{X}_n = \{x_i\ :\ 1 \leq i \leq n\} \subset \mathcal{X}$ of size n, and we solve the following regularized problem corresponding to Eq. (7.32) :

$$\begin{aligned} \rho^*_{n,\lambda} &= \sup\ c - \lambda \|g\|_1 \\ \text{subject to } &f(x_i) - c = g(x_i),\ 1 \leq i \leq n,\ g \in \mathrm{SOS}^{\mathcal{H}}, \end{aligned} \tag{7.35}$$

which we can recast as the following problem on semidefinite trace class operators :

$$\begin{aligned} &\rho^*_{n,\lambda} = \sup\ c - \lambda \operatorname{Tr}(A) \\ &\text{subject to } f(x_i) - c = \langle k_{x_i}, A k_{x_i} \rangle_{\mathcal{H}},\ 1 \leq i \leq n,\ A \in \mathcal{S}_+(\mathcal{H}). \end{aligned} \tag{7.36}$$

This corresponds to the following rewriting of the primal measure problem, where the measure is restricted to have support in $\mathcal{X}_n$:

$$\begin{aligned} &\rho_{n,\lambda} = \inf \int_{\mathcal{X}_n} f(x)\mu(dx) \\ &\text{subject to } \mu \in \mathscr{M}(\mathcal{X}_n),\ \int_{\mathcal{X}_n} \mu(dx) = 1\ \int_{\mathcal{X}_n} k_x \otimes_{\mathcal{H}} k_x\ \mu(dx) + \lambda \mathbf{I}_{\mathcal{H}} \succeq 0. \end{aligned} \tag{7.37}$$

Writing a signed measure with support in $\mathcal{X}_n$ as $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, we get the following dual of Eq. (7.36)

$$\begin{aligned} &\rho_{n,\lambda} = \inf \sum_{i=1}^n \alpha_i f(x_i) \\ &\text{subject to } (\alpha_i) \in \mathbb{R}^n,\ \sum_{i=1}^n \alpha_i = 1\ \sum_{i=1}^n \alpha_i k_{x_i} \otimes_{\mathcal{H}} k_{x_i} + \lambda \mathbf{I}_{\mathcal{H}} \succeq 0. \end{aligned} \tag{7.38}$$

Difference with SOS hierarchies. We want to highlight that it is this modification of the problem Eq. (7.32) which makes this method fundamentally different from the polynomial method. Indeed, by going from Eq. (7.32) to Eq. (7.35) the following phenomemons appear.

(i). Take $\lambda = 0$. The restriction to the set $\mathcal{X}_n$ makes the primal problem harder (we restrict the set of measures to sums of diracs on $\mathcal{X}_n$) and the dual problem easier (as we only enforce the equality constraint on a finite subset of $\mathcal{X}$). Thus, the operation of restricting to $\mathcal{X}_n$ is a strengthening of the primal, and a relaxation of the dual : $\rho_{n,0} \geq \rho_{\mathcal{H}}$ and $\rho^*_{n,0} \geq \rho^*_{\mathcal{H}}$. In the case where we do not know tightness a priori of problems Eqs. (7.32) and (7.33), we cannot say anything about $\rho^*_{n,0}, \rho_n$ compared to f_*, as we used to have $\rho_{\mathcal{H}}, \rho^*_{\mathcal{H}} \leq f_*$. Even in the case where Eqs. (7.32) and (7.33) can be shown to be tight *a priori*, if the class of functions $\mathrm{SOS}^{\mathcal{H}}$ is too large (and it will usually be the case as soon as $\mathcal{H}$ is infinite dimensional, for example in the Sobolev kernel setting of chapter 8), the problem is still complicated because we need regularization. Indeed, in the case where the space $\mathrm{SOS}^{\mathcal{H}}$ is large, it is possible that it can interpolate any function $f - c$ with $c \leq \min_{1 \leq i \leq n} f(x_i)$. In that case, we have $\rho^*_{n,0} = \min_{1 \leq i \leq n} f(x_i)$, which is a trivial upper bound on the minimum and can be obtained without such an elaborate method.

(ii). As soon as $\mathcal{H}$ and therefore $\mathrm{SOS}^{\mathcal{H}}$ is too large, regularization is necessary to avoid overfitting or interpolation. As the space $\mathrm{SOS}^{\mathcal{H}}$ is equipped with a natural norm given by the trace norm, it is natural to regularize the dual problem (with the SOS constraint) with that norm. This corresponds to a form of strengthening of the dual and of relaxation of the dual, and we therefore have $\rho_{n,\lambda} \leq \rho_{n,0}$ and $\rho^*_{n,\lambda} \leq \rho^*_{n,0}$. Since the bounds go in the other way around from the bounds between the discretized and non discretized problems, there is no way, even when the original

SOS problem is tight, to know if $\rho^*_{n,\lambda}, \rho_{n,\lambda}$ are lower or upper bounds for f_*. However, as is the e.r.m. setting, the interaction between n and λ helps derive asymptotic upper bounds between $\rho^*_{\mathcal{H}}$ and $\rho^*_{n,\lambda}$ under certain conditions, as we will see in Sec. 7.2 .2.

Algorithm

Looking at Eq. (7.35), we see that we can apply proposition 5.3 in chapter 5, *i.e.*, the representer theorem for PSD models. Following the reasoning in chapter 8, define Φ_i the $i-th$ column of a matrix such that $\Phi^\top \Phi = \mathbf{K}$, where $\mathbf{K}$ is the kernel matrix $\mathbf{K} = (k(x_i, x_j))_{1 \leq i,j \leq n}$. The problem can be written in the dual as

$$\begin{aligned} &\rho^*_{n,\lambda} = \sup\ c - \lambda \operatorname{Tr}(A) \\ &\text{subject to } f(x_i) - c = \Phi_i^\top A \Phi_i,\ 1 \leq i \leq n,\ A \succeq 0, \end{aligned} \tag{7.39}$$

and in the primal as

$$\begin{aligned} &\rho_{n,\lambda} = \inf \sum_{i=1}^n \alpha_i f(x_i) \\ &\text{subject to } (\alpha_i) \in \mathbb{R}^n,\ \sum_{i=1}^n \alpha_i = 1,\ \sum_{i=1}^n \alpha_i \Phi_i \Phi_i^\top + \lambda \mathbf{I}_{\mathbb{R}^n} \succeq 0. \end{aligned} \tag{7.40}$$

Both of these problems are linear semidefinite programs in the cone $\mathbb{S}_+(\mathbb{R}^n)$. If, for example, we solve the primal formulation using interior point methods, we can obtain an ε-solution in time $O(n^{3.5} \log \frac{1}{\varepsilon})$. They are detailed in chapter 8, in algorithm 6 and Sec. 8.6 .

Remark on extraction. Contrary to the polynomial case, where a stabilization of the rank was a sufficient condition to find a minimizer, it is not direct to a sufficient condition to extract a minimizer associated to the estimation $\rho_{n,\lambda}$ or $\rho^*_{n,\lambda}$ of the minimum. One way to go if $\mathcal{X}$ is equipped with a metric $d_{\mathcal{X}}$ is to look for x_* as the solution to the following problem :

$$\begin{aligned} x_* &= \arg\min_{x \in \mathcal{X}} \sum_{i=1}^n \alpha_i d_{\mathcal{X}}(x, x_i)^2 \\ &= \int_{\mathcal{X}} d_{\mathcal{X}}(x, x')^2 \mu_n(dx') \\ &= \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x_1, x_2)^2 \mu_n(dx_1) \delta_x(dx_2), \end{aligned} \tag{7.41}$$

Where δ_x is just a dirac in $\mathcal{X}$, and $\mu_n = \sum_{i=1}^n \alpha_i \delta_{x_i}$ is the signed measure with support in $\mathcal{X}_n$ which is the solution to Eq. (7.37). In other words, we look for x_* such that the dirac measure δ_{x_*} is as close as possible in Wasserstein-2 norm (Santambrogio, 2015) to the signed measure μ_n. In the case where $\mathcal{X} = \mathbb{R}^d$, the solution to Eq. (7.41) is simply $x_* = \sum_{i=1}^n \alpha_i x_i$. The way chosen by Rudi, Marteau-Ferey, and Bach (2020) is essentially this one. Note that in order for this strategy to be well motivated, it is important that the minimizer be unique.

7.2 .2 Applying this method to the minimization of regular functions in the Euclidean space

In this section, we explain how Rudi, Marteau-Ferey, and Bach (2020) apply the method described in Sec. 7.2 .1 in the setting of minimizing a function f belonging to $C_b^r(\Omega)$, where Ω is an open domain of $\mathbb{R}^d$, and where $C_b^r(\Omega)$ denotes the set of r continuously differentiable functions with bounded derivatives.

The goal of this work from a high level viewpoint is to leverage regularity in global optimization, even if it is performed using only function values.

Problem setting and worst case bounds

Formally, the high level objective consists in

(i) finding an ε-approximation of the minimum, *i.e.*, $\widehat{c}$ such that $|\widehat{c} - f_*| \leq \varepsilon$;

(ii) finding an ε-approximation of the minimizer, *i.e.*, $\widehat{x}$ such that $|f(\widehat{x}) - f_*| \leq \varepsilon$.

Given $\varepsilon > 0$, we wish to find such approximation with the lowest number of function calls (*i.e.*, the smallest possible number of x_i, $1 \leq i \leq n$ in Sec. 7.2 .1), and with worst-case guarantees over all functions f in some relevant class of functions $\mathcal{F} \subset C_b^r(\Omega)$. We want the method to find $\widehat{x}$ to satisfy :

$$\sup_{f \in \mathcal{F}} \left\{ f(\hat{x}) - \min_{x \in \Omega} f(x) \right\} \leq \varepsilon. \tag{7.42}$$

Novak (2006) argues that when having access to n function values, the problem of optimizing f is as hard as the problem of approximating f uniformly (see introduction and section 1.3 by Novak (2006) and Fig. 7.1 for a visual representation). In the setting where $f \in C_b^r(\Omega)$, section 1.3.9 by Novak (2006) shows that the problem of optimization is indeed as hard as approximation, and that $n =\propto \varepsilon^{-d/r}$ points are needed in order to obtain an ε-approximation (the proportionality constants may depend exponentially in d). The goal in the work by Rudi, Marteau-Ferey, and Bach (2020) is to match these bounds, that is exhibit an algorithm which needs only $n \propto \varepsilon^{-d/r}$ evaluation points to achieve ε error.

We say that such an algorithm breaks the curse of dimensionality in the exponent using smoothness in the sense that the number of samples needed in the worst case is not $n \propto \varepsilon^{-d}$ (as is the case in the Lischitz case and basic algorithms), but of order $\varepsilon^{-d/r}$. Note that the curse of dimensionality sometimes denotes the fact that we need much more regularity to optimize in high dimensions : in that case, of course, we do not break that specific curse, but rather add to its statement.

Assumptions

The work by Rudi, Marteau-Ferey, and Bach (2020) is done in the following setting. Assume Ω is a bounded open subset of $\mathbb{R}^d$, that $f \in C_b^r(\Omega)$ for some $r \geq 0$, and that the following assumption is satisfied.

Assumption 7.1. *There exists $\varepsilon > 0$ such that $\{x \in \mathcal{X} \ : \ f(x) - f_* \leq \varepsilon\}$ is compact.*

Assumption 7.1 states that close enough to the optimum, the level sets are compact. This implies *a)* the existence of at least one minimizer, and *b)* that there is no minimizer near the boundary.

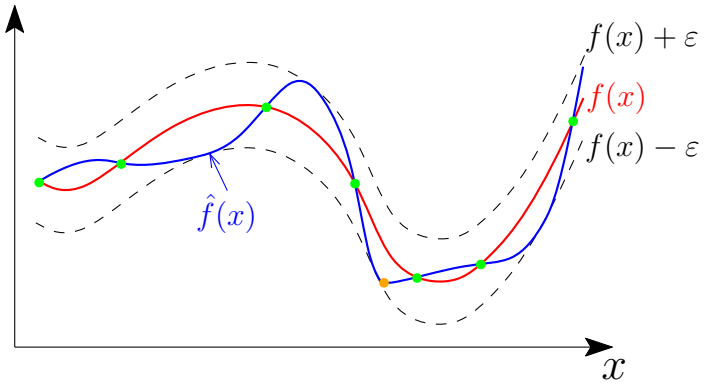

Figure 7.1: Optimization is as hard as approximation.

Finally, Rudi, Marteau-Ferey, and Bach (2020) take the Sobolev kernel k_s with $s > d/2$ (see Eq. (1.38)) to apply the method presented in Sec. 7.2 .1. Note that this choice of kernel is useful because of the direct link with differentiability, and makes the analysis clear. However, as can be seen in chapter 6, one can actually approximate such a kernel using a Gaussian kernel, with adapted bandwidth, and which is easier to implement in practice.

Unde these assumptions, we can easily prove that Eq. (7.32) actually solves the minimization problem, *i.e.*,

$$\rho^*_{\mathcal{H}} = \rho_{\mathcal{H}} = f_*. \tag{7.43}$$

This is a main difference with the polynomial setting, where Eq. (7.43) only happened if $f - f_*$ was a sum of squares of polynomials. In this setting, the space $\mathrm{SOS}^{\mathcal{H}}$ is large enough to contain any close approximation of $f - f_*$. However, this does not give us any information on $\rho_{n,\lambda}$, which is the result of the method by Rudi, Marteau-Ferey, and Bach (2020).

Similarly to the well-specified setting presented in part I, in order to obtain guarantees on $\rho_{n,\lambda}$, it is useful for the problem to be well specified, *i.e.*, that the solutions to Eqs. (7.32) and (7.33) be reached at certain values μ_* for the primal or (c_*, g_*) for the dual.

Guarantees on the relaxation

The first step of the work by Rudi, Marteau-Ferey, and Bach (2020) is therefore to show that under certain additional assumptions, there is an exact solution to Eq. (7.32). To do so, the following assumption is made on the set of minimizers.

Assumption 7.2. *All minimizers of f are strict global minimizers, i.e., the Hessian at all minimizers is positive definite.*

The following theorem is a simplification of Theorem 8.3. In particular, note that the condition $r \geq s + 2$ will imply that $r > d/2 + 2$.

Main theorem 5: Rudi et al. (2020), Theorem 3.

Assume that $r - 2 \geq s$ and recall that $\mathcal{H}$ is the reproducing kernel Hilbert space associated to the Sobolev kernel k_s. Under the assumptions above (including Assumptions 7.1 and 7.2) on the kernel and the function, there exists a finite rank operator $A_* \in \mathcal{S}_+(\mathcal{H})$ with rank at most $d|Z| + 1$ (where $|Z|$ is the size of the set of minimizers), such that :

$$\forall x \in \Omega, \ f(x) - f_* = \langle k_x, A_* k_x \rangle_{\mathcal{H}}. \tag{7.44}$$

In particular, this shows that Eq. (7.32) has an optimal solution (c_*, g_*) with $c_* = f_*$ and $g_* = \langle k_{(\cdot)}, A_* k_{(\cdot)} \rangle_{\mathcal{H}}$.

The proof of this theorem is based on a local decomposition as a sum of squares around each global minimum, before gluing the decompositions together using standard gluing lemmas from differential geometry. Using the fact that the Hessian is positive definite, we perform an exact Taylor-type decomposition around each global minimum to the second order : that is why we lose an exponent 2 in the bounds, *i.e.*, $r \geq s + 2$. This approach is generalized in Marteau-Ferey, Bach, and Rudi (2022b) in order to go to functions which have a continuous set of zeros, and which are defined on a manifold (see Sec. 7.2 .3 for more details).

Guarantees after restriction to a finite number of points and near-optimal algorithm

The second step of the method presented by Rudi, Marteau-Ferey, and Bach (2020) is to guarantee the equality constraint in Eq. (7.32) only on a finite subset $\mathcal{X}_n = \{x_1, ..., x_n\} \subset \Omega$ and to regularize the problem. To analyse the performance of the resulting $\rho^*_{n,\lambda}$, the authors use the fill distance $h_{\mathcal{X}_n,\Omega}$ of $\mathcal{X}_n$ in Ω, defined as

$$h_{\mathcal{X}_n,\Omega} = \sup_{x \in \Omega} \inf_{x_i \in \mathcal{X}_n} \|x - x_i\|_{\mathbb{R}^d}. \tag{7.45}$$

The following simplified theorem is obtained for the performance of $\rho^*_{n,\lambda}$, which is the solution of Eq. (7.35), whose details can be found in Theorem 8.5 and has originally been proved as Theorem 5. by Rudi, Marteau-Ferey, and Bach (2020).

Main theorem 6: Rudi, Marteau-Ferey, and Bach (2020), Theorem 5.

Assume that the conditions of Main theorem 5 are satisfied, and that Ω is a ball of radius R for simplicity. Let $r' < s - d/2$ There exists constants $C_{r',d}, C'_{r',d}$ depeding only on r', d and a constant $h_{R,r'}$ depending only on R, r' such that if $g_* \in \mathcal{S}_+(\mathcal{H})$ is the optimal solution satisfying $f - f_* = g_*$, we have

$$|\rho^*_{n,\lambda} - f_*| \ \leq \ C_{r',d} \ |f|_{\Omega,r'} \ h^{r'}_{\mathcal{X}_n,\Omega} \ + \ \lambda \|g\|_1, \tag{7.46}$$

as soon as $h_{\mathcal{X}_n,\Omega} \leq h_{R,r'}$, $\lambda \geq C'_{r',d} h^{r'}_{\mathcal{X}_n,\Omega}$.

The proof of this theorem relies on bounds obtained through scattered data analysis (Wendland, 2004), and the fact that $\mathcal{H}_s$ embeds itself in $C^{r'}(\Omega)$ for any $r' < s - d/2$.

Assuming that Ω is a ball of radius R, and that we can generate i.i.d. samples from the uniform

measure on the ball, we can choose the points x_i randomly with the following guarantee on the fill distance.

Lemma 7.1 (Lemma 8.4)**.** *Let $\widehat{\mathcal{X}}_n = \{x_1, \dots, x_n\}$ independent points sampled from the uniform distribution on Ω. There exists a constant n_d depending only on d such that for any $\delta > 0$, if $n \geq n_d \log \frac{2}{\delta}$, then the following holds with probability at least $1 - \delta$:*

$$h_{\widehat{\mathcal{X}}_n, \Omega} \;\leq\; 11R\, n^{-\frac{1}{d}}\, (\log \tfrac{2^d n}{\delta})^{1/d}. \tag{7.47}$$

Note that Lemma 8.4 is a more general result for sets which go beyond the simple ball of radius R. In the case of a hypercube, one could simply take a grid $\mathcal{X}_n$, and we would have $h_{\mathcal{X}_n} \leq Cn^{-1/d}$ where n is the number of points in the grid.

We now state Theorem 8.6, in the case where we subsample uniformly n points from the uniform measure when Ω is a ball of radius R.

Main theorem 7: Rudi, Marteau-Ferey, and Bach (2020), Theorem 6

Assume that Ω is a ball of radius R, that $\mathcal{H}$ is the RKHS associated to the Sobolev kernel k_s for $s > d/2$. Let $r \geq s + 2$ and $0 \leq r' < s - d/2$. There exists constant n_0 and C_0 depending on s, r', d, r such that the following hold. Let $f \in C_b^r(\Omega)$ satisfying Assumptions 7.1 and 7.2, $\widehat{\rho}^*_{n,\lambda}$ be the solution of Eq. (7.35), and $\delta \in (0, 1]$. If $n \geq n_0 \;\log \frac{2}{\delta}$ and λ satisfies

$$\lambda \;\geq\; C_0\, n^{-\frac{r'}{d}}\, (\log \tfrac{2^d n}{\delta})^{\frac{r'}{d}}, \tag{7.48}$$

then, with probability at least $1 - \delta$,

$$|\widehat{\rho}^*_{n,\lambda} - f_*| \;\leq\; 3\,\lambda\, \left(\|g_*\| + |f|_{\Omega,r'}\right). \tag{7.49}$$

In particular, assuming $r > d/2 + 2$ and taking $s = r - 2$ and $r' = \lceil r - 3 - d/2 \rceil$ and taking λ_n as in the lower bound of Eq. (7.48), we have constants n_1, C_1 depending only on r, d such that as soon as $n \geq n_1 \;\log \frac{2}{\delta}$, it holds

$$|\widehat{\rho}^*_{n,\lambda} - f_*| \;\leq\; C_1\; \left(\|g_*\| + |f|_{\Omega,r'}\right)\; n^{-(r/d - 3/d - 1/2)}\; (\log \tfrac{2^d n}{\delta})^{r/d - 2/d - 1/2}. \tag{7.50}$$

This result shows, up to logarithmic terms, that with the right choice of regularization, if $f \in C_b^r(\Omega)$ and satisfies the additional assumptions Assumptions 7.1 and 7.2, the method presented in Sec. 7.2 .1 achieves an error of order $n^{-r/d+3/d+1/2}$ up to logarithmic and constant terms (which do not depend on n). This is to compare to the worst case bound obtained by Novak (2006), which was $n^{-r/d}$. We are short of a term $3/d + 1/2$, which is due to the fact that *a)* our decomposition as a sum of squares reduces the regularity, and *b)* that we use the very loose fact that $C_b^s(\Omega) \subset W_2^s(\Omega)$. However, this results still shows that in spirit, as soon as the regularity is of the order of the dimension, one does not pay the dimensionality in the rate (*i.e.*, approximately, $|\widehat{\rho}^*_{n,\lambda} - f_*| \leq Cn^{-1/2}$), although the dimension is still present in the constants.

To end this part on the guarantees on the discretized regularized problems, note that the original results, given in chapter 8, also focus on providing explicit bounds for the constants in the above theorems. That is why there is a particular condition on the domain Ω, (although that condition is more general than that of being a ball of radius R, see Assumption 8.1). The same reasoning could be applied on a domain satisfying a uniform cone condition, as in chapter 4 of Adams and

Fournier (2003), and extend these results formally to a wider class of domains, although keeping track of the constants would lead to long and painful computations.

Extraction

Until now, we have only talked about approximating a minimum of f, but not a minimizer. In order to be able to approximate a minimizer, Rudi, Marteau-Ferey, and Bach (2020) use Eq. (7.41), and set $\widehat{x} = \sum_{i=1}^{n} \alpha_i x_i$, where the α_i are the optimal primal variables. This is justified formally in Sec. 8.7 , by making the assumptions that

(i) there is a unique minimizer x_* of f;

(ii) there is a parabola of the form $\nu\|x - x_*\|^2$ with small enough ν such that this parabola is a lower bound for f.

Under these assumptions, Rudi, Marteau-Ferey, and Bach (2020) provide an algorithm to retrieve both an approximation of the minimum and of the minimizer, using the same principle as the one described above, and with convergence guarantees of the form $\nu\|\widehat{x}_{n,\lambda} - x_*\|^2 \leq C\ n^{-d/r+3/d+1/2}$ for a suitable choice of λ (see Theorem 8.8).

7.2 .3 Going to functions with continuous sets of zeros on manifolds

The method presented in Sec. 7.2 .1 was quite general, and did not at all need that $\mathcal{X}$ be a domain of $\mathbb{R}^d$ to be applied. However, the analysis we have performed in this section crucially relies on certain properties of RKHS on $\mathbb{R}^d$, which are needed to prove that the problem Eq. (7.32) is not only tight, but also that there exists solutions to these problems (see Main theorem 5).

The goal of the work by Marteau-Ferey, Bach, and Rudi (2022b) is to extend these results which rely on the fact that certain regular non-negative functions can be decomposed as sums of squares of regular functions. It was motivated by two main lines of applications of this method we would like to explore.

- The first was to be able to decompose functions whose sets of minimizers is not discrete (as implied by Assumption 7.2), but wich can have some smoothness properties, such as being a manifold. This has been motivated by applications to optimal transport, where such sets of minimizers naturally appear (see Vacher, Muzellec, Rudi, Bach, and Vialard (2021)).
- The second was to be able to move from $\mathbb{R}^d$ to any manifold M, in order to apply this algorithm to minimize functions on manifolds.

The high level ideas of the work by Rudi, Marteau-Ferey, and Bach (2020) are the following. To generalize from domains of $\mathbb{R}^d$ to manifolds, one simply notes that the result was obtained in the $\mathbb{R}^d$ case by gluing decompositions. The same reasoning can be applied for manifolds, and hence the real crux of the problem is to show that certain regular non-negative functions can be locally decomposed as sums of squares of regular function on $\mathbb{R}^d$ (by composing with a chart). In the work by Rudi, Marteau-Ferey, and Bach (2020), the strategy to show a local decomposition as sums of squares was to use the fact that the Hessian at the optimal point was positive definite (Assumption 7.2) to write and exact second order decomposition, which naturally made a sum of squares decomposition appear (such as the ones obtained through the traditional Morse lemma, see Milnor (1963)). In the work by Marteau-Ferey, Bach, and Rudi (2022b), this is adapted to the setting where the set of zeros is a submanifold of $\mathbb{R}^d$, and Assumption 7.2 is modified to allow a continuous set of zeros. Instead, it imposes that orthogonally to the manifold of zeros, the Hessian is positive definite. Both of these situations and their difference are illustrated in

Fig. 7.2. Denoting with T_xN the set of tangeant vectors to a manifold N at x, we prove the following theorem (see chapter 9 for more details and references).

Main theorem 8: Marteau-Ferey, Bach, and Rudi (2022b), Theorem 3.9

Let f be a non-negative function of class C^p for $p \geq 2$ and let $\mathcal{Z}$ denote the set of zeros of f. If $\mathcal{Z}$ is a compact sub-manifold of M of class C^1 such that

$$\forall x_0 \in \mathcal{Z},\ \forall h \in T_{x_0}M \setminus T_{x_0}\mathcal{Z},\ H(x_0)[h,h] > 0, \tag{7.51}$$

then there exists a finite number of functions $f_1, ..., f_N \in C^{p-2}(M)$ such that

$$\forall x \in M,\ f(x) = \sum_{i=1}^{N} f_i(x)^2. \tag{7.52}$$

In Figs. 7.2 and 7.3 , we give examples of what functions which satisfy this assumption look like, when defined on $\mathbb{R}^d$ and on a manifold. While we have not formally done so yet, this result allows to apply exactly the same methodology as for functions with strict minima in open sets of $\mathbb{R}^d$, with, we believe, the same guarantees.

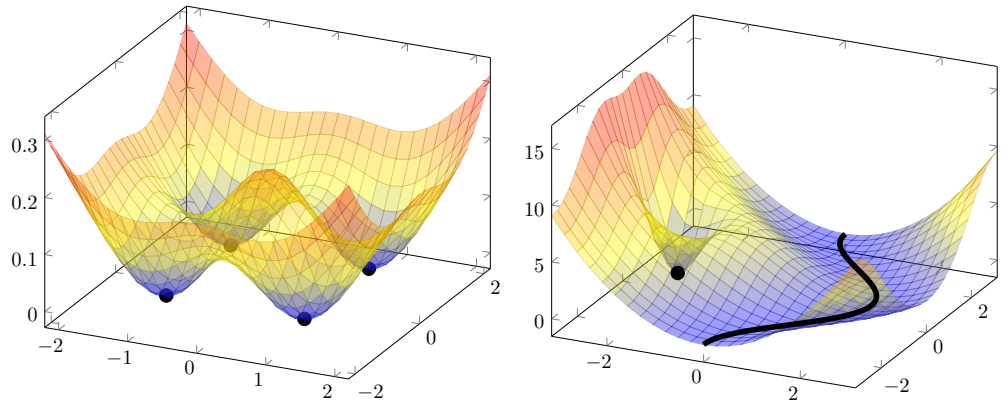

Figure 7.2: Plots of functions $z = f(x, y)$, where the zeros of f are highlighted in black. **left**: f satisfies Assumption 7.2, **right**: f does not satisfy Assumption 7.2 but satisfies the assumptions needed for a sum of squares decomposition in Marteau-Ferey, Bach, and Rudi (2022b).

7.3 Similarities and differences between the two approaches

We conclude this chapter by putting together Secs. 7.1 and 7.2 , and showing the structural similarities and differences between the moment-SOS hierarchy of lower bounds for polynomials and the method proposed by Rudi, Marteau-Ferey, and Bach (2020).

7.3 .1 Structure of the methods

First, we compare the structure of the methods, *i.e.*, the different key algorithmic steps to obtain an approximation of the minimum/minimizer.

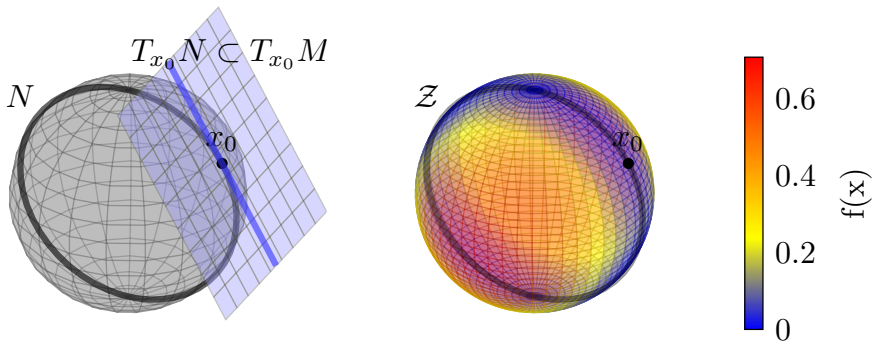

Figure 7.3: **Left:** Representation of the manifold $M = S^2$ as well as a sub-manifold N homeomorphic to a circle. The tangent spaces at a given point $x_0 \in N \subset M$ are represented as well. **Right:** Representation of a non-negative function on the sphere as a color map; it satisfies the assumptions of Marteau-Ferey, Bach, and Rudi (2022b) to be decomposed as a sum of squares, and its null space $\mathcal{Z}$ is represented in black.

The same relaxation

The first step of both methods is very similar : we relax the primal problem (or strengthen the dual) into a positive semidefinite problem. If we fix $r \in \mathbb{N}$, the set $\mathrm{SOS}_r[\boldsymbol{x}]$ of SOS polynomials of degree at most r (see Eq. (7.3)) corresponds exactly to the set $\mathrm{SOS}^{\mathcal{H}}$ in Eq. (7.31) where we set $\mathcal{H} = \mathbb{R}_r[\boldsymbol{x}]$ equipped with a kernel structure. For example, we can take $k(x_1, x_2) = \sum_{|\alpha| \leq r} {}_1 x^\alpha x_2^\alpha$, where the kernel structure is inherited from the linear representation v_r defined in Sec. 7.1 : $k(x_1, x_2) = v_r(x_1)^\top v_r(x_2)$ and where the norm of a polynomial $f = (f_\alpha)_{|\alpha| \leq r}$ in this RKHS is just the euclidean norm of the coefficients, *i.e.*, $\|f\|_{\mathcal{H}}^2 = \sum_{|\alpha| \leq r} f_\alpha^2$.

However, while we can identify this first relaxation step, there is a big difference between the two approaches at this point, owing to the necessity of solving opposite problems.

- In the polynomial case, the set $\mathcal{H} = \mathbb{R}_r[\boldsymbol{x}]$ is finite dimensional and hence is not large enough to perform a tight relaxation.
- On the contrary, in the standard kernel sum of squares, the set $\mathcal{H}$ can be taken in a way that $\rho_{\mathcal{H}}^* = \rho_{\mathcal{H}} = f_*$ (see Eq. (7.43)). However, the relaxed problem is therefore still infinite dimensional, and there is a need to decrease the dimension.

Hierarchichal structure for polynomials and the need for constraints

As highlighted above, since for a single r, the moment-SOS strengthening of $\mathcal{P}_+(\mathbb{R}^d)$ by $\mathrm{SOS}_r[\boldsymbol{x}]$ is not tight enough : we need to make r grow in roder to get closer and closer to $\mathcal{P}_+(\mathbb{R}^d)$, or, from another perspective, for ρ_r^* to get closer and closer to f_*. However, as we have explained in Sec. 7.1 , keeping to $\mathbb{R}^d$ and using an increasing r is not enough to approximate f_* well, *i.e.*, we do not necessarily have $\rho_r^* \uparrow f_*$.

However, it is the case in the constrained setting, where the polynomial contraints add some algebraic structure and allows to for ρ_r to go to f_*. Indeed, for polynomials, adding constraints loosens the rigidity of the polynomial structure, by allowing more cancellations between coefficients. This setting is the most widespread, as a lot of problems are constrained or can easily be constrained (when the minimum is roughly localized in a given region for instance).

Reducing dimension in the kernel sum of squares setting

On the contrary, in the kernel sum of squares case, there is no need for a hierarchy or for constraints for the proiblem to be tight. However, the initial problem is infinite dimensional and cannot be solved as such.

The technique developed by Rudi, Marteau-Ferey, and Bach (2020) is to reduce the problem of approximating the function $f - f_*$ by a sum of squares of functions g to a finite number of evaluation points $(x_1, ..., x_n)$, regularizing in order to avoid undesirable side effects of overfitting. This is a typical machine learning approach, as described in the general introduction.

The drawback of this approach is twofold. The theoretical analysis imposes a stronger assumption than simply tightness (*i.e.*, that $\rho^*_{\mathcal{H}} = f_*$), it imposes that an optimal solution exists. This assumption is much stronger than the tightness assumption, and proves to be one of the real challenges for analyzing these methods. Moreover, the *a posteriori* analysis of the returned solution is not obvious to do.

Solving a semidefinite program

The last structural point which appears in both moment-SOS polynomial hierarchies and in our setting is the need, in the end, to solve semidefinite programs. This can be quite challenging, and we see that the complexity grows in order d^r where r is the degree of the hierarchy used in the polynomial setting, while growing in order $n^{3.5}$ in the kernel sum of squares setting. This shows that these two methods have really different limiting conditions in terms of computation time. The SOS hierarchy is limited by the complexity of the problem through its dimension and the degree needed in the hierarchy, while the kernel SOS technique is limited by the quality of the approximation of the contraint $f - f_* = g$ using n points.

7.3 .2 Two different type of guarantees

The difference between these two methods is reflected in the guarantees we can obtain. Indeed, as polynomials are more rigid, the *a priori* convergence guarantees are relatively weak, and speed of convergence of the hierarchy cannot be analyzed. However, the objects involved are much more structured, and one can obtain good *a posteriori* guarantees. Symmetrically, in the kernel sum of squares setting, we obtain very good rates and non-asymptotic guarantees. However, as the model is implicit and large, it is not obvious to obtain *a posteriori* guarantees from a returned solution, in particular because the functions f we handle are much less structured.

Guarantees for the moment-SOS hierarchy

In the polynomial case, the *a priori guarantees* show convergence of the hierarchy of lower and upper bounds, and even finite convergence in some cases. Moreover, for some specific domains $\mathbb{K}$, this convergence can be shown to have error $1/(r')^2$, where r' is the degree of the hierarchy, thus showing that a degree of order $1/\sqrt{\varepsilon}$ is needed to reach ε error. However, as soon as the underlying dimension d is large, the number of variables grows exponentially as $d^{r'}$ in terms of the order r' of the hierarchy, thus making high orders for high precision ε unreachable (even though some sparsification techniques can be applied).

In practice, the *a posteriori* guarantees can be very strong. When the algorithm finishes, the user can have the following information :

- a lower bound ρ_r on f_*;

- if a rank condition holds, the knowledge that $\rho_r = f_*$ and a procedure to compute a minimizer;
- if not, a technique to approximate the minimizer if certain conditions are satisfied.

As explained above, the situation is quite different in the kernel sum of squares setting.

Guarantees for kernel sum of squares

There are strong *a priori* guarantees :

- fast rates of convergence with regularity for the minimum (i.e. of order $\varepsilon^{-C\ m/d}$ for a constant C in time and space, to be compared to the space complexity of order $d^{1/\sqrt{\varepsilon}}$ in the polynomial case);
- fast rates of convergence for the minimizer under certain additional assumptions.

Howoever, there are no a posteriori guarantees, except for the one given by the approximation of the minimizer (*i.e.*, we will know that $f(\widehat{x}) \geq f_*$).

In term of complexity, this methods has a real advantage, as it can always be run, and the user can hope that it will work well in favorable settings. Moreover, note that the rates of convergence derived in chapter 8 are expressed using certain theoretical quantities which give us insight into what makes a problem difficult or not : the norms of the elements in the sum of squares decomposition for example.

7.3 .3 Building connexions

As we have seen in the two previous sections, there are fundamental similarities and differences between these two methods : in one, one seeks to tackle the rigidity of polynomials while in the other, one tackles the "large dimension" of the kernel models.

However, the variety of fields in which the moment-SOS hierarchies have been applied also provides a plethora of interesting task for the kernel sum of squares to tackle, accompanied of course by additional challenges to solve. We will go a bit further into detail about these different directions in chapter 10, and just formulate a few questions which directly appear when looking at the analysis and comparison above.

Are there ways to keep a lower a posteriori bound in the kernel setting ?

One of the main advantages of polynomial optimization from the user point of view is that it provides lower bounds (which can directly be coupled to an upper bound if an approximation of the minimizer is returned). This, of course, cannot be done using the technique developed by Rudi, Marteau-Ferey, and Bach (2020) as the evaluation breaks the guarantee of being a lower bound. Therefore, can we find a way to get both better approximation properties for larger classes of functions with kernels, while keeping an upper bound ?

Constraints

We also have seen that adding constraints helps the optimization guarantees in the polynomial case. Does something similar exist in the kernel case (for example, maybe this reduces the norm of optimal solutions) ?

Moments

In the setting we described for the kernel sum of squares case, we did not once talk about moments, which were central objects in the polynomial case. This is something which would be very interesting to explore in greater detail, and can be defined, although in a way which is a bit more involved than in the polynomial setting.

Let k be a kernel and $\mathcal{H}$ the associated RKHS. We consider the RKHS $\mathcal{H}^{\otimes 2}$ associated to the kernel k^2, and note that there is a natural embedding $\mathrm{Op} : \mathcal{H}^{\otimes 2} \to \mathcal{S}_2(\mathcal{H})$ which extends the map $\sum_i \alpha_i k_{x_i}^2 \in \mathcal{H}^{\otimes 2} \mapsto \sum_i \alpha_i k_{x_i} \otimes_{\mathcal{H}} k_{x_i} \in \mathcal{S}_2(\mathcal{H})$, where $\mathcal{S}_2(\mathcal{H})$ is the set of Hilbert-Schmidt operators on $\mathcal{H}$ (in particular, this set contains the set $\mathcal{S}_1(\mathcal{H})$ of trace class operators). In that case, the set of moments would be

$$M^{\mathcal{H}} = \left\{ g \in \mathcal{H}^{\otimes 2} \ : \ \exists \mu \in \mathscr{M}(\mathcal{X}),\ g = \int_{\mathcal{X}} k_x^2 \mu(dx),\ \langle 1, g \rangle_{\mathcal{H}^{\otimes 2}} = 1,\ \mathrm{Op}(g) \in \mathcal{S}_+(\mathcal{H}) \right\}, \quad (7.53)$$

assuming of course that the function 1 belongs to $\mathcal{H}^{\otimes 2}$. We have yet to study this set of moments in the kernel setting, and its link with extraction procedures.

Chapter 8

Finding global minima via kernel approximations

This chapter is a verbatim of the work :

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations, 2020. URL https://arxiv.org/abs/2012.11978.

Contents

8.1 Introduction . . . 348
8.2 Outline of contributions . . . 349
8.3 Setting . . . 353
8.4 Equivalence of the infinite-dimensional problem . . . 356
8.5 Properties of the finite-dimensional problem . . . 360
8.6 Algorithm . . . 366
8.7 Finding the global minimizer . . . 368
8.8 Extensions . . . 370
8.9 Relationship with polynomial hierarchies . . . 373
8.10 Experiments . . . 375
8.11 Discussion . . . 380
8.A Additional notations and definitions . . . 384
8.B Fundamental results on scattered data approximation . . . 390
8.C Auxiliary results on RKHS . . . 392
8.D The constants of translation invariant and Sobolev kernels . . . 393
8.E Proofs for algorithm 6 . . . 399
8.F Global minimizer. Proofs. . . . 404
8.G Proofs for the extensions . . . 407
8.H Details on the algorithmic setup used in the benchmark experiments 411

8.1 Introduction

We consider the general problem of unconstrained optimization. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a possibly non-convex function. Our goal is to solve the following problem

$$\min_{x \in \mathbb{R}^d} f(x). \tag{8.1}$$

In particular, we will consider the setting where (a) the function is smooth, that is, $f \in C^m(\mathbb{R}^d)$ with $m \in \mathbb{N}_+$ (f m-times continuously differentiable), and (b) we are able to evaluate it on given points, without the need of computing the gradient. For this class of problems there are known lower-bounds (Novak, 2006; Nesterov, 2013) that show that it is not possible to achieve a global minimum with error ε with less than $O(\varepsilon^{-d/m})$ function evaluations. In this paper, we want to achieve this lower bound in terms of function evaluations, while having an optimization algorithm which has a running-time which is polynomial in the underlying dimension and the number of function evaluations.

Several methods are available to solve this class of problems. For example, the function f can be approximated from its values at n sampled points, and the approximation of the function globally minimized instead of f. If the approximation is good enough, then this can be optimal in terms of n, but computationally infeasible. Optimal approximations can be obtained by multivariate polynomials (Ivanov, 1971) or functions in Sobolev spaces (Novak and Woźniakowski, 2008), with potentially adaptive ways of selecting points where the function is evaluated (see, e.g., the work by Osborne, Garnett, and Roberts (2009) and references therein). Alternatively, when the function is itself a polynomial, algorithms based on the "sum-of-squares" paradigm can be used, but their computational complexity grows polynomially on $d^{r/2}$, where r is in the most favorable situations the order of the polynomial, but potentially larger when so-called hierarchies are used (Lasserre, 2001; Laurent, 2009; Lasserre, 2010).

It turns out that the analysis of lower-bounds on the number of function evaluations shows an intimate link between function interpolation and function minimization, i.e., the lower-bounds of one problem are the same for the other problem. However, existing methods consider a two-step approach where (1) the function is approximated optimally, and (2) the approximation is minimized. In this paper, we consider a joint approach where approximation and optimization are done *jointly*.

We derive an algorithm that casts the possibly non-convex problem in Eq. (8.1) in terms of a simple convex problem based on a non-parametric representation of non-negative functions via positive definite operators (Marteau-Ferey, Bach, and Rudi, 2020). As shown below, it can be considered as an infinite-dimensional counter-part to polynomial optimization with sums of squares, with two key differences: (1) the relaxation is always tight for the direct formulation, and (2) the computational cost does not depend on the dimension of the model (here infinite anyway), by using a subsampling algorithm and a computational trick common in statistics and machine learning.

The resulting algorithm with n sampled points will be able to achieve an error of $\varepsilon = O(n^{-m/d+3/d+1/2})$ as soon as $m \geq 3 + d/2$, with n function evaluations to reach the global minimum with precision ε, and a computational complexity of $O(n^{3.5} \log(1/\varepsilon))$ (with explicit constants). This is still not the optimal complexity in terms of number of function evaluations (which is $\varepsilon = O(n^{-m/d})$), but this is achieved with a polynomial-time algorithm in n. This is particularly interesting in the contexts where the function to be optimized is very smooth, i.e., $m \gg d$, possibly C^∞ or a polynomial. For example, if the function is differentiable at least $d + 3$ times, even if non-convex, the proposed algorithm finds the global minimum with error $O(n^{-1/2})$ and time $O(n^{3.5} \log n)$.

Note that the (typically exponential) dependence on the dimensionality d is only in the constants and tracked explicitly in the rest of the paper.

Moreover the algorithm is based on simple interior-point methods for semidefinite programming, directly implementable and based only on function evaluations and matrix operations. It can thus leverage multiple GPU architectures to reach large values of n, which are needed when the dimension grows.

8.2 Outline of contributions

In this section, we present our framework, our algorithm and summarize the associated guarantees.

Denote by $\zeta \in \mathbb{R}^d$ a global minimizer of f and assume to know a bounded open region $\Omega \subset \mathbb{R}^d$ that contains ζ. We start with a straightforward and classical convex characterization of the problem in Eq. (8.1), with infinitely many constraints:

$$\max_{c \in \mathbb{R}} \; c \quad \text{such that} \quad \forall x \in \Omega, \; f(x) \geq c. \tag{8.2}$$

Note that the solution c_* of the problem above corresponds to $c_* = f(\zeta) = f_*$, the global minimum of f. The problem above is convex, but typically intractable to solve, due to the dense set of inequalities that c must satisfy.

To solve Eq. (8.2) our main idea is to represent the dense set of inequalities in terms of a dense set of *equalities* and then to approximate them by subsampling.

Tight relaxation. We start by introducing a quadratic form $\langle \phi(x), A\phi(x) \rangle$ with A a self-adjoint positive semidefinite operator from $\mathcal{H}$ to $\mathcal{H}$, for a suitable map $\phi : \Omega \to \mathcal{H}$ and an infinite-dimensional Hilbert space $\mathcal{H}$, to define the following problem

$$\max_{c \in \mathbb{R}, \; A \in \mathbb{S}_+(\mathcal{H})} \; c \quad \text{such that} \quad \forall x \in \Omega, \; f(x) - c = \langle \phi(x), A\phi(x) \rangle \,, \tag{8.3}$$

where $\mathbb{S}_+(\mathcal{H})$ is the set of bounded self-adjoint positive semi-definite operators on $\mathcal{H}$.

The problem in Eq. (8.3) has a smaller optimized objective function than the problem in Eq. (8.2) because we constrain A to be positive semi-definite and any feasible point for Eq. (8.3) is feasible for Eq. (8.2). In fact, when f is a polynomial and $\phi(x)$ is composed of monomials of degree less than half the degree of f (and thus $\mathcal{H}$ finite-dimensional), then we recover the classical "sum-of-squares" relaxation of polynomial optimization. In that situation, the relaxation is tight only if $f - f_*$ is itself a sum-of-squares, which is known to not always be the case. Then, to make the relaxation tight in the limit, several hierarchies of polynomial optimization problems have been considered using polynomials of increasing degrees (Lasserre, 2001; Laurent, 2009; Lasserre, 2010).

In this paper, we consider a well-chosen infinite-dimensional space $\mathcal{H}$, and we prove that if f is smooth enough (i.e., m-times differentiable with $m > 3 + d/2$), under mild geometrical assumptions on f then there always exists a map ϕ, and a finite rank $A_* \in \mathbb{S}_+(\mathcal{H})$ for which the problem in Eq. (8.2) and the one above are equivalent, that is, the relaxation is tight.

Note that, the resulting ϕ, despite being infinite-dimensional, has an explicit and easy-to-compute ($O(d)$ in memory and time) inner product $k(x, x') = \langle \phi(x), \phi(x') \rangle$ that will be the only quantity required to run the algorithm. We will thus use Hilbert spaces $\mathcal{H}$ which are reproducing kernel

Algorithm 6 Global minimum. Given $f : \mathbb{R}^d \to \mathbb{R}$, $\Omega, n \in \mathbb{N}_+, \lambda > 0, s > d/2$.

$\widehat{X} \leftarrow \{x_1, \dots, x_n\}$ ▷ Sampled i.i.d. uniformly on Ω
$f_j \leftarrow f(x_j),\ \forall j \in [n]$

Features computation
$K_{ij} \leftarrow k(x_i, x_j)\ i, j \in [n]$ ▷ k Sobolev kernel of smoothness s, Eq. (8.7)
$R \leftarrow \texttt{cholesky}(K)$ ▷ upper triangular Cholesky
$\Phi_j = j$-th column of $R,\ \forall j \in [n]$

Solution of the approximate problem (use any algorithm in Sec. 8.6)
$\hat{c} \leftarrow \max_{c \in \mathbb{R}, B \in \mathbb{S}_+(\mathbb{R}^n)} c - \lambda \operatorname{Tr}(B)$ such that $\forall j \in [n], f_j - c = \Phi_j^\top B \Phi_j$
return $\hat{c}$

Hilbert spaces (Berlinet and Thomas-Agnan, 2011), such as Sobolev spaces (Adams and Fournier, 2003).

Subsampling. We approximate the problem above as follows. Given a finite set $\widehat{X} = \{x_1, \dots, x_n\}$ which is a subset of Ω, we restrict the equality in Eq. (8.3) to only $x_1, \dots, x_n$.

Unlike the case of polynomial optimization where subsampling is exact if n is large enough (Lasserre, Toh, and Yang, 2017), in our case subsampling leads to an error that decreases in n and depends on the regularity of f and of the map $x \mapsto \langle \phi(x), A\phi(x) \rangle$. While f is smooth enough by assumption, we need to control the regularity of the map induced by A, to guarantee that the constraints subsampled on $\widehat{X}$ well approximate the whole set of constraints on Ω. Then we consider a penalization term based on the trace of A and solve the following problem

$$\begin{aligned} \max_{c \in \mathbb{R},\ A \in \mathbb{S}_+(\mathcal{H})} \quad & c - \lambda \operatorname{Tr}(A) \\ \text{such that} \quad & \forall i \in \{1, \dots, n\},\ f(x_i) - c = \langle \phi(x_i), A\phi(x_i) \rangle \,, \end{aligned} \tag{8.4}$$

for some positive λ (with the implicit assumption that we optimize over operators A with finite trace). We show in this paper that solving Eq. (8.4) leads to an approximate optimum of the original problem in Eq. (8.2), when n is large enough and λ small enough. Note that the value of c which we obtain after subsampling in not anymore a lower bound on the global minimum, but we can provide both a priori and a posteriori certificates of optimality (see Sec. 8.8 .3).

Finite-dimensional algorithm. The problem in Eq. (8.4) is still formulated in an infinite-dimensional space. We can leverage the particular choice of penalty by the trace of A and the choice of Hilbert space to obtain a finite-dimensional algorithm. Indeed, for reproducing kernel Hilbert spaces, then, following (Marteau-Ferey, Bach, and Rudi, 2020), we only need to solve the problem in the finite-dimensional Hilbert space spanned by $\phi(x_1), \dots, \phi(x_n)$, that is, we only need to look at A of the form $A = \sum_{i,j=1}^n C_{ij} \phi(x_i) \otimes \phi(x_j)$ for some positive semi-definite *matrix* $C \in \mathbb{R}^{n \times n}$. We can then write $\operatorname{Tr}(A) = \operatorname{Tr}(CK)$, with $K \in \mathbb{R}^{n \times n}$ the matrix of dot-products with $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$, and $\langle \phi(x_i), A\phi(x_i) \rangle = (KCK)_{ii}$.

Consider the Cholesky decomposition of K as $K = R^\top R$, with $R \in \mathbb{R}^{n \times n}$ upper-triangular. We can directly solve for $B = RCR^\top$, noting that $KCK = R^\top BR$ and $\operatorname{Tr}(CK) = \operatorname{Tr}(B)$. We can thus use a representation in terms of finite-dimensional vectors $\Phi_1, \dots, \Phi_n \in \mathbb{R}^n$ defined as the

columns of R. We thus study the following problem,

$$\begin{aligned} \max_{c\in\mathbb{R},\ B\in\mathbb{S}_+(\mathbb{R}^n)} \quad & c - \lambda \operatorname{Tr}(B) \\ \text{such that} \quad & \forall i \in \{1,\dots,n\},\ f(x_i) - c = \Phi_i^\top B \Phi_i. \end{aligned} \tag{8.5}$$

From an algorithmic viewpoint the problem above can be solved efficiently since this is a semi-definite program. We show in Sec. 8.6 how we can apply Newton method and classical interior-point algorithms, leading to a computational complexity of $O(n^{3.5}\log(1/\varepsilon))$ in time and $O(n^2)$ in space.

Note that in the context of sum-of-squares polynomials, the relationship with reproducing kernel Hilbert spaces had been explored for approximation purposes after a polynomial optimization algorithm is used (Marx, Pauwels, Weisser, Henrion, and Lasserre, 2019). In this paper, we propose to leverage kernel methods *within* the optimization algorithm.

Why not simply subsampling the inequality? One straightforward algorithm is to subsample the dense set of *inequalities* in Eq. (8.2). Doing this will simply lead to outputting $\min_{i\in\{1,\dots,n\}} f(x_i)$. This last algorithm, while easy to implement and convergent, is very slow, with a rate of $O(n^{-2/d})$ (see the discussion in Sec. 8.11). Subsampling the dense set of *equalities* in Eq. (8.3) allows to use smooth interpolation tools. When $\lambda = 0$, the optimal value is also $\min_{i\in\{1,\dots,n\}} f(x_i)$ (if the kernel matrix is invertible, see Sec. 8.6), but for $\lambda > 0$, we can leverage smoothness as shown below.

Theoretical guarantees. From a theoretical viewpoint, denoting by $\hat{c}$ the minimizer of Eq. (8.5), we provide upper bounds for $|f_* - \hat{c}|$ with explicit constants and that hold under mild geometrical assumptions on f. We prove that the bound depends on how the points in $\widehat{X} = \{x_1,\dots,x_n\}$ are chosen. In particular we prove that when they are chosen uniformly at random on Ω, the problem in Eq. (8.5) achieves the global minimum with error ε with a precise dependence on n.

The results in this paper hold under the following assumptions.

Assumption 8.1 (Geometric properties on Ω and f)**.** *The following holds:*

(a) *Let $\Omega = \cup_{x\in S} B_r(x)$, where S is a bounded subset of $\mathbb{R}^d$ and $B_r(x)$ is the open ball of radius $r > 0$, centered in x.*

(b) *The function f is in $C^2(\mathbb{R}^d)$. Ω contains at least one global minimizer. The minimizers in Ω are isolated points with strictly positive Hessian and their number is finite. There is no minimizer on the boundary of Ω.*

Note that Assumption 8.1(a) can be easily relaxed to Ω having locally Lipschitz-continuous boundaries (Adams and Fournier, 2003, Section 4.9). Assumption 8.1(b) is satisfied if all global minimizers of f are in Ω, and are second-order strict local minimizers. Note that similar assumptions are made to show finite convergence for polynomial optimization hierarchies (Nie, 2014).

Theorem 8.1 (Main result, informal)**.** *Let $\Omega \subset \mathbb{R}^d$ be a ball of radius $R > 0$. Let $s > d/2$ and let k be Sobolev kernel of smoothness s (see Example 8.1). Let $f \in C^{s+3}(\mathbb{R}^d)$ and that satisfies Assumption 8.1(b). Let $\hat{c}$ be the result of algorithm 6 executed with $n \in \mathbb{N}_+$ points chosen uniformly at random in Ω and $\lambda > 0$. Let $\delta > 0$. There exist $n_{s,d,\delta}, C_{s,d} > 0$ such that, when $n > n_{s,d,\delta}$, and*

$$\lambda \ \geq\ C_{s,d}\, n^{-s/d+1/2}\, (\log \tfrac{n}{\delta})^{s/d-1/2},$$

then, with probability at least $1-\delta$,

$$|\hat{c} - f_*| \;\leq\; 3\,\lambda\,\left(\mathrm{Tr}(A_*) + |f|_{\Omega,\lceil s-d/2\rceil}\right),$$

where A_* *is any solution of Eq.* (8.3).

Note that A_* exists since $f \in C^{s+3}(\mathbb{R}^d)$ and it satisfies the geometrical mild condition in Assumption 8.1(b) (as we prove in Sec. 8.4), and that all constants can be made explicit (see Theorem 8.6). From the result above, and with $m = s+3$, for $s > d/2$, we can achieve an error of order $n^{-s/d+1/2}$, which translates to $\varepsilon = O(n^{-m/d+3/d+1/2})$ as soon as $m > d/2+3$. We pay the additional exponent 3 since we construct the candidate matrix representing the solution by requiring that each component of the Hessian of f, which is $m-2$ times differentiable belongs to the RKHS. This accounts for the 2 term, the last 1 is paid simply since s can be not integer. The rate for the class of functions $C^m(\Omega)$ is sub-optimal by an exponent $1/2+3/d$. In the following remark we are going to show that our algorithm achieves nearly-optimal convergence rates when the function to optimize is in a Sobolev space. Denote by $W_2^s(\Omega)$ the Sobolev space of squared-integrable functions of smoothness $s > 0$, i.e., the space of functions whose weak derivatives up to order s are square-integrable on Ω, (see the work by Adams and Fournier (2003)).

Remark 24 (Nearly optimal rates for Sobolev spaces.). *If* Ω *satisfies Assumption* 8.1(a), f *satisfies Assumption* 8.1(b) *and* $f \in W_2^s(\Omega)$, *with* $s > d/2+3$, *then algorithm 6 with Sobolev kernel of smoothnes* $s-3$ *achieves the convergence rate*

$$O\big(n^{-s/d+1/2+3/d}\big),$$

modulo logarithmic factors, as proven in Theorem 8.6. When d *is large, then the error exponent is asymptotically optimal, since the term* $3/d$ *becomes negligible, leading to the optimal exponent* $-s/d+1/2$ *(see, e.g., (Novak and Woźniakowski, 2008, Prop. 1.3.11)).*

Finding the global minimizer. In Sec. 8.7 we derive an extension of the problem in Eq. (8.5), with the goal of finding the global minimizer. Under the additional assumption that the minimizer is unique we obtain the similar rate as Theorem 8.5 for the localization of the global minimizer.

Warm restart scheme for linear rates. Applying a simple warm restart scheme, we prove, in Sec. 8.7 .2, that when f has a unique global minimum, then it is possible to achieve it with error ε, with a number of observations that is only logarithmic in ε

$$n = O(C_{d,m}\log(1/\varepsilon)),$$

for some constant $C_{d,m}$ that can be exponential in d (note that the added assumption of unique minimizer makes this result not contradict the lower bound in $\varepsilon^{-d/m}$).

Rates for functions with low smoothness $m \leq d/2$ **or functions that are not in** $\mathcal{H}$**.** In Sec. 8.8 .2 we study a variation of the problem in Eq. (8.5) that allows to have some error $\tau > 0$ on the constraints. When $f \in C^{m+2}(\Omega)$, by tuning τ appropriately with respect to λ, we show that algorithm 6 applied on this different formulation achieves an error in the order

$$O\big(\,n^{-\frac{m}{2d}(1-(d-m)/(2r-m))}\,\big),$$

where r is now the index of the Sobolev kernel and can be chosen arbitrarily large. The exponent of the rate above matches the optimal one for C^{m+2} functions (that is $-(m+2)/d$) up to a multiplicative factor of $\frac{1}{2}\frac{1}{1+2/m}$.

Relationship to polynomial optimization. When f is a polynomial of degree $2r$, then it is natural to consider $\phi(x)$ composed of all monomials of degree less than r, leading to a space $\mathcal{H}$ of dimension $\binom{d+r}{r}$. All polynomials can be represented as $f(x) = c + \phi(x)^\top A \phi(x)$ for some symmetric matrix A. When $A \succcurlyeq 0$, by using its eigendecomposition, we can see that the polynomial $x \mapsto \phi(x)^\top A \phi(x)$ is a sum-of-squares polynomial.

However, in general A may not be positive semi-definite, as non-negative polynomials are not all sum-of-squares. Moreover, even when there exists a matrix $A \succcurlyeq 0$, the corresponding c may not be the minimum of f (it only needs to be a lower bound)—see, e.g., the work by Lasserre (2001) and references therein.

If $f(x) - f_*$ is a sum of squares, then, with $\lambda = 0$ and $n = \binom{d+2r}{2r}$ points (to ensure that subsampling is exact), we exactly get the minimum of f, as we are solving *exactly* the usual optimization problem.

When $f(x) - f_*$ is not a sum of squares, then a variety of hierarchies have been designed when optimization is performed on a compact constraint set described with polynomial inequalities (such as taking an ℓ_2-ball for Ω), that augment the problem dimensionality to reach global convergence(Lasserre, 2001; Laurent, 2009; Lasserre, 2010). In Sec. 8.9 , we show how our framework fits with one these hierarchies, and also can provide computational gains.

Note that our framework, by looking directly at an infinite-dimensional space circumvents the need for hierarchies, and solves a single optimization problem. The difficulty is that it requires sampling. Moreover by using only kernel evaluations, we circumvent the explicit construction of a basis for $\mathcal{H}$, which is computationally cumbersome when d grows.

Organization of the paper. The paper is organized as follows: in Sec. 8.3 , we present the kernel setting our paper relies on; then, in Sec. 8.4 , we analyze the infinite-dimensional problem and show its equivalence with global minimization. Then, in Sec. 8.5 , we present our theoretical guarantee for the finite-dimensional algorithm, as summarized in Theorem 8.1. In Sec. 8.6 we present the dual algorithm based on self-concordant barriers and the damped Newton algorithm. In Sec. 8.7 , we present our extension to find the global minimizer, while in Sec. 8.8 , we provide certificates of optimality for potentially inexactly solved problems. In Sec. 8.9 , we discuss further relationships with polynomial hierarchies, and provide illustrative experiments in Sec. 8.10 . We conclude in Sec. 8.11 with a discussion opening up to many future problems.

8.3 Setting

In this section, we first introduce some definitions and notation about *reproducing Kernel Hilbert spaces* in Sec. 8.3 .1 (for more details, see the work by Aronszajn (1950); Paulsen and Raghupathi (2016)), and present our detailed assumptions in Sec. 8.3 .2. In Sec. 8.4 we show how our infinite-dimensional sum-of-squares representation can be built, and in Sec. 8.5 we provide guarantees on subsampling.

8.3 .1 Definitions and notation

In this section we denote by $u \cdot v$, $a \circ v$ respectively the pointwise multiplication between the functions u and v, and the composition between the functions a and v. We denote by $\mathbb{N}$ the set of natural numbers including 0, by $\mathbb{N}_+$ the set $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$ and $[n]$ the set $\{1, \dots, n\}$ for $n \in \mathbb{N}_+$. We will always consider $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$ if not specified otherwise. Moreover we denote by $B_r(z)$ the open ball $B_r(z) = \{x \in \mathbb{R}^d \mid \|x - z\| < r\}$. Let $\Omega \subseteq \mathbb{R}^d$ be an

open set. Let $\alpha \in \mathbb{N}^d$. We introduce the following *multi-index notation* $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $\partial_x^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ (Adams and Fournier, 2003). For $m \in \mathbb{N}$, and Ω an open set of $\mathbb{R}^d$, denote by $C^m(\Omega)$ the set of m-times differentiable functions on Ω with continuous m-th derivatives. For any function u defined on a superset of Ω and m times differentiable on Ω, define the following semi norm.

$$|u|_{\Omega,m} = \max_{|\alpha|=m} \sup_{x \in \Omega} \left|\partial^\alpha u(x)\right|. \tag{8.6}$$

Positive definite matrices and operators. Let $\mathcal{H}$ be a Hilbert space, endowed with the inner product $\langle \cdot, \cdot \rangle$. Let $A : \mathcal{H} \to \mathcal{H}$ be a linear operator and denote by A^* the adjoint operator, by $\mathrm{Tr}(A)$ the trace of A and by $\| \cdot \|_F$ the Hilbert-Schmidt norm $\|A\|_F^2 = \mathrm{Tr}(A^*A)$. We always endow $\mathbb{R}^p$ with the standard inner product $x^\top y = \sum_{i=1}^p x_i y_i$ for any $x, y \in \mathbb{R}^p$. In the case $\mathcal{H} = \mathbb{R}^p$, with the standard inner product, then $A \in \mathbb{R}^{p \times p}$ is a matrix and the Hilbert-Schmidt norm corresponds to the Frobenius norm. We say that $A \succeq 0$ or A is a *positive operator* (positive matrix if $\mathcal{H}$ is finite dimensional), when A is bounded, self-adjoint, and $\langle u, Au \rangle \geq 0$, $\forall u \in \mathcal{H}$. We denote by $\mathbb{S}_+(\mathcal{H})$ the space of positive operators on $\mathcal{H}$. Moreover, we denote by $A \succ 0$, or A strictly positive operator, the case $\langle u, Au \rangle > 0$ for all $u \in \mathcal{H}$ such that $u \neq 0$.

Kernels and reproducing kernel Hilbert spaces. For this section we refer to the work by Aronszajn (1950); Steinwart and Christmann (2008); Paulsen and Raghupathi (2016), for more details (see also Sec. 8.A .3, page 387). Let Ω be a set. A function $k : \Omega \times \Omega \to \mathbb{R}$ is called a *positive definite kernel* if all matrices of pairwise evaluations are positive semi-definite, that is, if it satisfies the following equation

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0, \qquad \forall n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in \Omega.$$

Given a kernel k, the *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$, with the associated inner product $\langle \cdot, \cdot \rangle$, is a space of real functions with domain Ω, with the following properties.

(a) The function $k_x = k(x, \cdot)$ satisfies $k_x \in \mathcal{H}$ for any $x \in \Omega$.

(b) The inner product satisfies $\langle f, k_x \rangle = f(x)$ for all $f \in \mathcal{H}$, $x \in \Omega$. In particular $\langle k_{x'}, k_x \rangle = k(x', x)$ for all $x, x' \in \Omega$.

In other words, function evaluations are uniformly bounded and continuous linear forms and the k_x are the evaluation functionals. The norm associated to $\mathcal{H}$ is the one induced by the inner product, i.e., $\|f\|^2 = \langle f, f \rangle$. We remark that given a kernel on Ω there exists a unique associated RKHS on Ω (Berlinet and Thomas-Agnan, 2011). Moreover, the kernel admits a characterization in terms of a *feature map* ϕ,

$$\phi : \Omega \to \mathcal{H}, \quad \text{defined as} \quad \phi(x) = k(x, \cdot) = k_x, \ \forall x \in \Omega.$$

Indeed according to the point (b) above, we have $k(x, x') = \langle \phi(x), \phi(x') \rangle$ for all $x, x' \in \Omega$. We will conclude the section with an example of RKHS that will be useful in the rest of the paper.

Example 8.1 (Sobolev kernel (Wendland, 2004))**.** *Let $s > d/2$, with $d \in \mathbb{N}_+$, and Ω be a bounded open set. Let*

$$k_s(x, x') = c_s \|x - x'\|^{s - d/2} \mathcal{K}_{s-d/2}(\|x - x'\|), \quad \forall x, x' \in \Omega, \tag{8.7}$$

where $\mathcal{K} : \mathbb{R}_+ \to \mathbb{R}$ the Bessel function of the second kind (see, e.g., 5.10 in the work by Wendland (2004)) and $c_s = \frac{2^{1+d/2-s}}{\Gamma(s-d/2)}$. The constant c_s is chosen such that $k_s(x, x) = 1$ for any $x \in \Omega$.

In the particular case of $s = d/2 + 1/2$, we have $k(x, x') = \exp(-\|x - x'\|)$. Note that a scale factor is often added as $k(x, x') = \exp(-\|x - x'\|/\sigma)$ in this last example. In such case, all bounds that we derive in this paper would then have extra factors proportional to powers of σ. To conclude, when Ω has locally Lipschitz boundary (a sufficient condition is Assumption 8.1(a)) then $\mathcal{H} = W_2^s(\Omega)$, where $W_2^s(\Omega)$ is the Sobolev space of functions whose weak-derivatives up to order s are square-integrable (Adams and Fournier, 2003). Moreover, in this case $\|\cdot\|_{\mathcal{H}}$ is equivalent to $\|\cdot\|_{W_2^s(\Omega)}$.

Reproducing kernel Hilbert spaces are classically used in fitting problems, such as appearing in statistics and machine learning, because of function evaluations $f \mapsto f(x)$ are bounded operators for any x, and optimization problems involving f only through function evaluations at a finite number of points $x_1, \dots, x_n$, and penalized with the norm $\|f\|$, can be solved by looking only a f of the form $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ (Aronszajn, 1950; Paulsen and Raghupathi, 2016). We will use an extension of this classical "representer theorem" to operators and spectral norms in Sec. 8.5 .

8.3 .2 Precise assumptions on reproducing kernel Hilbert space

On top of Assumption 8.1 (made on the function f and the set Ω), we make the following assumptions on the space $\mathcal{H}$ and the associated kernel k.

Assumption 8.2 (Properties of the space $\mathcal{H}$)**.** *Given a bounded open set $\Omega \subset \mathbb{R}^d$, let $\mathcal{H}$ be a space of functions on Ω with norm $\|\cdot\|_{\mathcal{H}}$, satisfying the following conditions*

(a) *$w|_\Omega \in \mathcal{H}$, $\forall w \in C^\infty(\mathbb{R}^d)$. Moreover there exists $\mathsf{M} \geq 1$ such that*

$$\|u \cdot v\|_{\mathcal{H}} \leq \mathsf{M}\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}, \quad \forall u, v \in \mathcal{H}.$$

(b) *$a \circ v \in \mathcal{H}$, for any $a \in C^\infty(\mathbb{R}^p)$, $v = (v_1, \dots, v_p)$, $v_j \in \mathcal{H}$, $j \in [p]$.*

(c) *Let $z \in \mathbb{R}^d, r > 0$ s.t. the ball $B_r(z)$ is in Ω. For any $u \in \mathcal{H}$, there exists $g_{r,z} \in \mathcal{H}$ s.t.*

$$g_{r,z}(x) = \int_0^1 (1 - t)u(z + t(x - z))dt, \quad \forall x \in B_r(z).$$

(d) *$\mathcal{H}$ is a RKHS with associated kernel k. For some $m \in \mathbb{N}_+$ and some $\mathsf{D}_m \geq 1$, the kernel k satisfies*

$$\max_{|\alpha|=m} \sup_{x,y\in\Omega} |\partial_x^\alpha \partial_y^\alpha k(x, y)| \leq \mathsf{D}_m^2 < \infty.$$

Assumptions 8.2(a) to 8.2(c) above require essentially that functions in $\mathcal{H}$ (a) can be multiplied by other functions in $\mathcal{H}$, or by infinitely smooth functions, and still be in $\mathcal{H}$; (b) that can be composed with infinitely smooth functions, or (c) integrated, and still be in $\mathcal{H}$. Moreover Assumption 8.2(d) requires that $\mathcal{H}$ is a RKHS with a kernel that is m-times differentiable. An interesting consequence of Assumption 8.2(d) is the following remark (for more details, see, e.g., (Steinwart and Christmann, 2008, Corollary 4.36)).

Remark 25. *Assumption 8.2(d) guarantees that $\mathcal{H} \subseteq C^m(\Omega)$ and $|u|_{\Omega,m} \leq \mathsf{D}_m\|u\|_{\mathcal{H}}$.*

Note that Assumptions 8.2(a) to 8.2(c) are the only required in Sec. 8.4 to prove the crucial decomposition in Theorem 8.2 and are satisfied by notable spaces (that are not necessarily RKHS) like $C^s(\Omega)$ or Sobolev spaces $W_p^s(\Omega)$ with $s > d/p$ and $p \in [1, \infty]$. Instead, Assumption 8.2(d) is required for the analysis of the finite-dimensional problem and in particular Theorems 8.4

and 8.5. In the following proposition we show that $W_2^s(\Omega)$ with $s > d/2$ and Ω satisfying Assumption 8.1(a) satisfy the whole of Assumption 8.2.

Proposition 8.1 (Sobolev kernels satisfy Assumption 8.2)**.** *Let Ω be a bounded open set of $\mathbb{R}^d$. The Sobolev kernel with $s > d/2$ recalled in Example 8.1 satisfies Assumption 8.2 for any $m \in \mathbb{N}_+, m < s - \frac{d}{2}$ and*

$$\mathsf{M} = (2\pi)^{d/2} 2^{s+1/2}, \qquad \mathsf{D}_m = (2\pi)^{d/4} \sqrt{\frac{\Gamma(m + d/2)\Gamma(s - d/2 - m)}{\Gamma(s - d/2)\Gamma(d/2)}}.$$

The proof of proposition above is in Sec. 8.D .2, page 395. We make a last assumption regarding the differentiability of f, namely that f and its second-derivatives are in $\mathcal{H}$.

Assumption 8.3 (Analytic properties of f)**.** *The function f satisfies $f|_\Omega \in C^2(\Omega) \cap \mathcal{H}$ and $\frac{\partial^2 f}{\partial x_i \partial x_j}|_\Omega \in \mathcal{H}$ for all $i, j \in [d]$.*

8.4 Equivalence of the infinite-dimensional problem

In Theorem 8.2 and Cor. 8.1, we provide a representation of $f - f_*$ in terms of an infinite-dimensional, *but finite-rank*, positive operator, under basic geometric conditions on f and algebraic properties of $\mathcal{H}$. In Theorem 8.3 we use this operator to prove that Eq. (8.3) achieves the global minimum of f. In this section we analyze the conditions under which the problem in (8.3) has the same solution as the one in Eq. (8.2).

The proof follows by explicitly constructing a bounded positive operator A_* (which will have finite trace) that satisfy $f(x) - f_* = \langle \phi(x), A_* \phi(x) \rangle$ for all $x \in \Omega$. Note that, by construction $f - f_*$ is a non-negative function. If $w := \sqrt{f - f_*} \in \mathcal{H}$ then $A_* = w \otimes w$ would suffice. However, denoting by $\zeta \in \Omega$ a global minimizer, note that $f(\zeta) - f_* = 0$ and the smoothness of $\sqrt{f - f_*}$ may degrade around ζ, making $\sqrt{f - f_*} \notin \mathcal{H}$ even if $f - f_* \in \mathcal{H}$.

Here we follow a different approach. In Lemma 8.1 we provide a decomposition that represents the function $f - f_*$ locally around each global optimum using the fact that it is locally strongly convex around the minimizers. In the proof of Theorem 8.2 we provide a decomposition of the function far from the optimal points; we then glue these different decompositions via bump functions.

Lemma 8.1. *Let $\mathcal{H}$ be a space of functions on Ω that satisfy Assumptions 8.2(a) to 8.2(c). Let $\zeta \in \Omega$ and $r, \gamma > 0$. Let $B_r(\zeta) \subset \Omega$ be a ball centered in ζ of radius r and $g \in C^2(\Omega)$ satisfy $g(\zeta) = 0$, $\nabla^2 g(x) \succcurlyeq \gamma I$ for $x \in B_r(\zeta)$ and $\frac{\partial^2}{\partial x_i \partial x_j} g \in \mathcal{H}$ for $i, j \in [d]$. Then, there exists $w_j \in \mathcal{H}, j \in [d]$ such that*

$$g(x) = \sum_{j=1}^{d} w_j(x)^2, \quad \forall x \in B_r(\zeta). \tag{8.8}$$

Proof. Let $x \in B_r(\zeta)$ and consider the function $h(t) = g(\zeta + t(x - \zeta))$ on $[0, 1]$. Note that $h(0) = g(\zeta)$ and $h(1) = g(x)$. Taking the Taylor expansion of h of order 1, we have $h(1) = h(0) + h'(0) + \int_0^1 (1 - t) h''(t) dt$, with $h(0) = g(\zeta)$, $h'(0) = (x - \zeta)^\top \nabla g(\zeta)$ and $h''(t) = (x - \zeta)^\top \nabla^2 g(\zeta + t(x - \zeta))(x - \zeta)$. Since $g(\zeta) = 0$ by construction and $\nabla g(\zeta) = 0$ since ζ is a local minimizer of g, we have $h(0) = h'(0) = 0$ leading to

$$g(x) = (x - \zeta)^\top R(x)(x - \zeta), \quad R(x) = \int_0^1 (1 - t) \nabla^2 g(\zeta + t(x - \zeta)) dt. \tag{8.9}$$

Note that for $x \in B_r(\zeta)$ we have $\nabla^2 g(x) \succcurlyeq \gamma I$ and so $R(x) \succcurlyeq \gamma I$. In particular, this implies that for any $x \in B_r(\zeta)$, $S(x) = \sqrt{R(x)}$ is well defined ($\sqrt{\cdot} : \mathbb{S}_+(\mathbb{R}^d) \to \mathbb{S}_+(\mathbb{R}^d)$ is the spectral square root, where for any $M \in \mathbb{S}_+(\mathbb{R}^d)$ and any eigen-decomposition $M = \sum_{j=1}^d \lambda_j u_j u_j^\top$, $\sqrt{M} = \sum_{j=1}^d \sqrt{\lambda_j} u_j u_j^\top$). Thus,

$$\forall x \in B_r(\zeta),\ g(x) = (x-\zeta)^\top S(x) S(x)(x-\zeta) = \sum_{i=1}^d \left(e_i^\top S(x)(x-\zeta)\right)^2.$$

The following steps prove the existence of $w_i \in \mathcal{H}$ such that $w_i|_{B_r(\zeta)} = e_i^\top S(\cdot)(\cdot - \zeta)$. Let $(e_1, ..., e_d)$ be the canonical basis of $\mathbb{R}^d$ and $\mathbb{S}(\mathbb{R}^d)$ be the set of symmetric matrices on $\mathbb{R}^d$ endowed with Frobenius norm, in the rest of the proof we identify it with the isometric space $\mathbb{R}^{d(d+1)/2}$ (corresponding of taking the upper triangular part of the matrix and reshaping it in form of a vector).

Step 1. *There exists a function $\overline{R} : \Omega \to \mathbb{S}(\mathbb{R}^d)$, such that*

$$\forall i, j \in [d],\ e_i^\top \overline{R} e_j \in \mathcal{H} \text{ and } \overline{R}|_{B_r(\zeta)} = R.$$

This is a direct consequence of the fact that $\frac{\partial^2}{\partial x_i \partial x_j} g \in \mathcal{H}$ for all $i \leq j \in [d]$, of Assumption 8.2(c) and of the definition of R in Eq. (8.9).

Step 2. *There exists a function $\overline{S} : \Omega \to \mathbb{S}(\mathbb{R}^d)$ such that*

$$\forall i, j \in [d],\ e_i^\top \overline{S} e_j \in \mathcal{H} \text{ and } \forall x \in B_r(\zeta),\ \overline{S}(x) = \sqrt{R}(x).$$

Let $\tau := \sup_{x \in B_r(\zeta)} \|R(x)\|_{\text{op}} = \|\overline{R}(x)\|_{\text{op}}$, which is well defined because R is continuous since $g \in C^2(\Omega)$. Define the compact set $K = \{T \in \mathbb{S}(\mathbb{R}^d) \mid \gamma I \preceq T \preceq \tau I\}$ and the open set $U = \{T \in \mathbb{S}(\mathbb{R}^d) \mid \frac{\gamma}{2} I \prec T \prec 2\tau I\}$. Note that $K \subset U \subset \mathbb{S}(\mathbb{R}^d)$.

Fix $i, j \in [d]$ and consider the function $\theta_{i,j} : U \to \mathbb{R}$ defined by $\theta_{i,j}(M) = e_i^\top \sqrt{M} e_j$. Since the square root $\sqrt{\cdot} : \mathbb{S}_+(\mathbb{R}^d) \to \mathbb{S}_+(\mathbb{R}^d)$ is infinitely differentiable (see e.g. the explicit construction in the work by Del Moral and Niclas (2018) Thm. 1.1) and $U \subset \mathbb{S}_+(\mathbb{R}^d)$ then $\theta_{i,j}$ is infinitely differentiable on U, i.e., $\theta_{i,j} \in C^\infty(U)$. By proposition 8.10, since K is a compact set in U, there exists $\overline{\theta}_{i,j} \in C_0^\infty(\mathbb{S}(\mathbb{R}^d))$ such that $\forall T \in K,\ \overline{\theta}_{i,j}(T) = \theta_{i,j}(T)$.

Define $\overline{S}(x) = \sum_{i,j \in [d]} (\overline{\theta}_{i,j} \circ \overline{R})(x) e_i e_j^\top$ for any $x \in \Omega$. Applying Assumption 8.2(b), $e_i^\top \overline{S} e_j = \overline{\theta}_{i,j} \circ \overline{R} \in \mathcal{H}$ since the $\overline{R}_{k,l} \in \mathcal{H},\ k, l \in [d]$ and $\overline{\theta}_{i,j}$ is in $C_0^\infty(\mathbb{S}(\mathbb{R}^d))$. Moreover, by construction, for any $x \in B_r(\zeta)$, we have $\overline{R}(x) = R(x) \in K$ and so

$$\overline{S}_{i,j}(x) = \overline{\theta}_{i,j}(\overline{R}(x)) = \theta_{i,j}(R(x)) = e_i^\top \sqrt{R(x)} e_j.$$

Note that here, we have applied proposition 8.10 and Assumption 8.2(b) to $\mathbb{S}(\mathbb{R}^d)$ and not to $\mathbb{R}^{d(d+1)/2}$; this can be made formal by using the linear isomorphism between $\mathbb{S}(\mathbb{R}^d)$ endowed with the Frobenius norm and $\mathbb{R}^{d(d+1)/2}$ endowed with the Euclidean norm.

Step 3. *There exists a function $\overline{h} = (\overline{h}_j)_{j \in [d]} : \Omega \to \mathbb{R}^d$ such that*

$$\forall j \in [d], \overline{h}_j \in \mathcal{H} \text{ and } \forall x \in B_r(\zeta),\ \overline{h}(x) = x - \zeta.$$

Fix $j \in [n]$. Define $\overline{B}_r(\zeta) = K \subset U = B_{2r}(\zeta)$ and apply proposition proposition 8.10 to $x \in U \mapsto e_j^\top (x - \zeta)$ to get $h_j \in C_0^\infty(\mathbb{R}^d)$ which coincides with $e_j^\top (\cdot - \zeta)$ on K hence on $B_r(\zeta)$. Applying Assumption 8.2(a), the restriction $\overline{h}_j = h_j|_\Omega$ is in $\mathcal{H}$, and hence $\overline{h} = \sum_{j \in [d]} \overline{h}_j e_j$ satisfies the desired property.

Step 4. *The $w_i = e_i^\top \overline{S}\ \overline{h},\ i \in [d]$ have the desired property.*

It is clear that the w_i are in $\mathcal{H}$ as a linear combination of products of functions in $\mathcal{H}$ (see Assumption 8.2(a)), since $w_i = \sum_{j\in[d]} \overline{S}_{ij}(x)\overline{h}_j(x)$ for any $x \in \Omega$. Moreover,

$$\sum_{i\in[d]} w_i^2 = \overline{h}^\top \overline{S}^\top \left(\sum_{i=1}^{d} e_i e_i^\top\right) \overline{S}\ \overline{h} = \overline{h}^\top \overline{S}^2 \overline{h}.$$

Using the previous points,

$$\forall x \in B_r(\zeta),\ \sum_{i\in[d]} w_i^2(x) = \overline{h}^\top(x)\overline{S}^2(x)\overline{h}(x) = (x-\zeta)^\top R(x)(x-\zeta) = g(x).$$

□

Now we are going to use the local representations provided by the lemma above to build a global representation in terms of a finite-rank positive operator. Indeed far from the global optima the function $f - f_*$ is strictly positive and so we can take a smooth extension of the square root to represent it and glue it with the local representations around the global optima via bump functions as follows.

Theorem 8.2. *Let Ω be a bounded open set and let $\mathcal{H}$ be a space of functions on Ω that satisfy Assumptions 8.2(a) to 8.2(c). Let f satisfy Assumptions 8.1(b) and 8.3. Then there exist $w_1, \dots, w_q \in \mathcal{H}$ with $q \le dp + 1$ and $p \in \mathbb{N}_+$ the number of minimizers in Ω, such that*

$$f(x) - f_* = \sum_{j\in[q]} w_j(x)^2, \qquad \forall\ x \in \Omega. \tag{8.10}$$

Proof. Let $Z = \{\zeta_1, \dots, \zeta_p\}$, $p \in \mathbb{N}_+$ be the *non-empty* set of global minima of f, according to Assumption 8.1(b). Denote by $f_* = \min_{x\in\Omega} f(x)$ the global minimum of f, and by $g : \Omega \to \mathbb{R}$ the function $g = f|_\Omega - f_*\mathbf{1}|_\Omega$ where $\mathbf{1}$ is the function $\mathbf{1}(x) = 1$ for any $x \in \mathbb{R}^d$. Assumption 8.3 implies that $\nabla^2 g = \nabla^2 f|_\Omega$ is continuous, an that $\frac{\partial^2 g}{\partial x_i \partial x_j} \in \mathcal{H}$ for any $i, j \in [d]$. Moreover, $g \in \mathcal{H}$. Indeed, by construction $f_*\mathbf{1}$ is in $C^\infty(\mathbb{R}^d)$, and since $\mathcal{H}$ satisfies Assumption 8.2(a), $f_*\mathbf{1}|_\Omega \in \mathcal{H}$. Since $f|_\Omega \in \mathcal{H}$ by Assumption 8.3, then $g \in \mathcal{H}$.

Step 1. *There exists $r > 0$ and $\alpha > 0$ such that (i) the $B_r(\zeta_l),\ l \in [p]$ are included in Ω and (ii) for any $x \in \bigcup_{l\in[p]} B_r(\zeta_l)$, it holds $\nabla^2 g(x) \succeq \alpha I$.*

By Assumption 8.1(b), for all $\zeta \in Z$, $\nabla^2 g(\zeta) \succ 0$. Since $\nabla^2 g$ is continuous, Z is a finite set, and Ω is an open set, there exists a radius $r > 0$ and $\alpha > 0$ such that for all $l \in [p]$, $B_r(\zeta_l) \subset \Omega$ and $\nabla^2 g|_{B_r(\zeta_l)} \succeq \alpha I$. For the rest of the proof, fix r, α satisfying this property. For any $X \subset \Omega$ denote with $\mathbf{1}_X$ the indicator function of a X in Ω. We define $\chi_0 = \mathbf{1}_{\Omega\setminus\bigcup_{l\in[p]} B_{r/2}(\zeta_l)}$, and $\chi_l = \mathbf{1}_{B_r(\zeta_l)},\ l \in [p]$.

Step 2. *There exists $w_0 \in \mathcal{H}$ s.t. $w_0^2\chi_0 = g\chi_0$.*

Ω is bounded and by Assumption 8.1(b), the set of global minimizers of f included in Ω is finite and there is no mimimizer of f on the boundary, i.e., there exists $m_1 > 0$ and a compact $K \subset \Omega$ such that $\forall x \in \Omega \setminus K,\ g(x) \ge m_1$.

Moreover, f has no global optima on the compact $K \setminus \bigcup_{\zeta\in Z} B_{r/2}(\zeta)$ since the set of global optima is Z, hence the existence of $m_2 > 0$ such that $\forall x \in K \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l),\ g(x) \ge m_2$. Taking

$m = \min(m_1, m_2)$, it holds $\forall x \in \Omega \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l),\ g(x) \geq m > 0$. Since $f \in C^2(\Omega)$, f is also bounded above on Ω hence the existence of $M > 0$ such that $g \leq M$. Thus

$$\forall x \in \Omega \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l),\ g(x) \in I \subset (m/2, 2M), \qquad I = [m, M].$$

Since $\sqrt{\cdot} \in C^\infty((m/2, 2M))$, $(m/2, 2M)$ is an open subset of $\mathbb{R}$ and I is compact, applying proposition 8.10, there exists a smooth extension $s_I \in C_0^\infty(\mathbb{R})$ such that $s_I(t) = \sqrt{t}$ for any $t \in I$. Now since $g \in \mathcal{H}$ and $s_I \in C_0^\infty(\mathbb{R})$, by Assumption 8.2(b), $w_0 := s_I \circ g \in \mathcal{H}$. Since $\forall x \in \Omega \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l),\ g \in I$, this shows $g\chi_0 = w_0^2\chi_0$.

Step 3. *For all $l \in [p]$, there exists $(w_{l,j})_{j\in[d]} \in \mathcal{H}^d$ s.t. $g(x)\chi_l = \sum_{j=1}^d w_{l,j}^2\ \chi_l$.*

This is an immediate consequence of Lemma 8.1 noting that $\nabla g(x) \geq \alpha I$ on $B_r(\zeta_l)$.

Step 4. *There exists $b_l \in C^\infty(\mathbb{R}^d)$ s.t. $b_l = b_l\ \chi_l$ for all $l \in \{0, 1, \ldots, p\}$ and $\sum_{l=0}^p b_l^2 = 1$.*

This corresponds to Lemma 8.7, Sec. 8.A .4, page 388 applied to the balls $B_r(\zeta_l),\ l \in [p]$.

Step 5. Using all the previous steps

$$\begin{aligned} g = \sum_{l=0}^p g\ b_l^2 &= \sum_{l=0}^p g(\chi_l\ b_l)^2 = \sum_{l=0}^p (\chi_l\ g)\ (\chi_l b_l^2) \\ &= (\chi_0 w_0^2)\ (\chi_0\ b_0^2) + \sum_{l=1}^p \Big(\chi_l \sum_{j=1}^d w_{l,j}^2\Big)\ \chi_l\ b_l^2 \\ &= ([b_0\ \chi_0]\ w_0)^2 + \sum_{l=1}^p \sum_{j=1}^d ([b_l\ \chi_l]\ w_{l,j})^2 = (b_0\ w_0)^2 + \sum_{l=1}^p \sum_{j=1}^d (b_l\ w_{l,j})^2. \end{aligned}$$

Applying Assumption 8.2(a) to each function inside the squares in the previous expressions yields the result.

□

A direct corollary of the theorem above is the existence of $A_* \in \mathbb{S}_+(\mathcal{H})$ when $\mathcal{H}$ is a reproducing kernel Hilbert space satisfying the assumptions of Theorem 8.2.

Corollary 8.1. *Let k be a kernel whose associated RKHS $\mathcal{H}$ satisfies Assumptions 8.2(a) to 8.2(c) and let f satisfy Assumptions 8.1(b) and 8.3, then there exists $A_* \in \mathbb{S}_+(\mathcal{H})$ with $\operatorname{rank}(A_*) \leq d|Z| + 1$ such that $f(x) - f^* = \langle \phi(x), A_*\phi(x)\rangle$ for all $x \in \Omega$.*

Proof. By Theorem 8.2 we know that if f satisfies Assumptions 8.1(b) and 8.3 w.r.t. a space $\mathcal{H}$ that satisfies Assumptions 8.2(a) to 8.2(c), there exists $w_1, \ldots, w_q \in \mathcal{H}$ with $q \leq d|Z| + 1$ such that $f(x) - f^* = \sum_{j\in[q]} w_j^2(x)$ for any $x \in \Omega$. Since $\mathcal{H}$ is a reproducing kernel Hilbert space, for any $h \in \mathcal{H},\ x \in \Omega$ we have $h(x) = \langle \phi(x), h\rangle_{\mathcal{H}}$. Moreover, by the properties of the outer product in Hilbert spaces, for any $h, v \in \mathcal{H}$, it holds $(\langle h, v\rangle_{\mathcal{H}})^2 = \langle h, (v \otimes_{\mathcal{H}} v)h\rangle$.

Thus, for any $x \in \Omega, j \in [q]$, it holds $w_j(x)^2 = \langle \phi(x), (w_j \otimes w_j)\phi(x)\rangle$ and hence

$$\forall x \in \Omega,\ f(x) - f^* = \langle \phi(x), A_*\phi(x)\rangle\,, \qquad A_* = \sum_{j\in[q]} w_j \otimes w_j.$$

□

To conclude the section we prove the problem in Eq. (8.3) admits a maximizer whose non-negative operator is of rank at most $d|Z|+1$.

Theorem 8.3. *Let $\Omega \subset \mathbb{R}^d$ be an open set, k be a kernel, $\mathcal{H}$ the associated RKHS, and $f : \mathbb{R}^d \to \mathbb{R}$. Under Assumptions 8.1 to 8.3, the problem in Eq. (8.3) admits an optimal solution (c_*, A_*) with $c_* = f_*$, and A_* a positive operator on $\mathcal{H}$ with rank at most $d|Z|+1$.*

Proof. Let p_0 be the maximum of Eq. (8.2). Since $A \succeq 0$ implies $\langle \phi(x), A\phi(x)\rangle \geq 0$ for all $x \in \Omega$, the problem in Eq. (8.2) is a relaxation of Eq. (8.3), where the constraint $f(x) - c = \langle \phi(x), A\phi(x)\rangle$ is substituted by $f(x) - c \geq 0, \forall x \in \Omega$. Then $p_0 \geq p_*$ if a maximum p_* exists for Eq. (8.3). Moreover if there exists A that satisfies the constraints in Eq. (8.3) for the value $c_* = f_*$, then $p_0 = p_*$ and (c_*, A) is a maximizer of Eq. (8.3). The proof is concluded by applying Cor. 8.1 that shows that there exists A satisfying the constraints in Eq. (8.3) for the value $c = f_*$. □

In Cor. 8.1 and Theorem 8.3 we proved the existence of an infinite-dimensional trace-class positive operator A_* that satisfies $\langle \phi(x), A_*\phi(x)\rangle = f(x) - f_*$ for any $x \in \Omega$ and maximizing Eq. (8.3). The proof is quite general, requiring some geometric properties on f, the fact that f and its second derivatives belong to $\mathcal{H}$ and some algebraic properties of the space $\mathcal{H}$, in particular to be closed to multiplication with a C^∞ function, to integration, and to composition with a C^∞ map. The generality of the proof does not allow to derive an easy characterization of the trace of A_*.

8.5 Properties of the finite-dimensional problem

In the previous section we proved that there exists a finite rank positive operator A_* minimizing Eq. (8.3). In this section we study the effect of the discretization of Eq. (8.3) on a given a set of distinct points $\widehat{X} = \{x_1, \dots, x_n\}$. First, we derive Theorem 8.4 which is fundamental to prove Theorem 8.5, and is our main technical result (we believe it can have a broader impact beyond the use in this paper as discussed in Sec. 8.11). Given a smooth function g on Ω, in Theorem 8.4 we prove that if there exists a matrix $B \in \mathbb{S}_+(\mathbb{R}^n)$ such that $g(x_i) = \Phi_i^\top B \Phi_i$ for $i \in [n]$ (the vectors $\Phi_j \in \mathbb{R}^n$ are defined before Eq. (8.5)), then the inequality $g(x) \geq -\varepsilon$ holds for any $x \in \Omega$ for an ε depending on the smoothness of the kernel, the smoothness of g and how well the points in $\widehat{X}$ cover Ω. We denote by $h_{\widehat{X},\Omega}$ the *fill distance* (Wendland, 2004),

$$h_{\widehat{X},\Omega} = \sup_{x \in \Omega} \min_{i \in [n]} \|x - x_i\|, \tag{8.11}$$

corresponding to the maximum distance between a point in Ω and the set $\widehat{X}$. In particular, if the kernel and g are m-times differentiable, Theorem 8.4 proves that $g(x) \geq -\varepsilon$ holds with $\varepsilon = O(h^m_{\widehat{X},\Omega})$ which is an improvement when $m \gg 2$ with respect to standard discretization results that guarantee exponents of only 1 or 2. Then in Lemma 8.3 we show that there exists a finite-dimensional positive definite matrix $B \in \mathbb{S}_+(\mathbb{R}^n)$ such that $\mathrm{Tr}(B) \leq \mathrm{Tr}(A_*)$ and $\Phi_i^\top B \Phi_i = \langle \phi(x_i), A_*\phi(x_i)\rangle$ for all $i \in [n]$. Finally, in Theorem 8.5, we combine Lemma 8.3 with Theorem 8.4, to show that the problem in Eq. (8.5) provides a solution that is only $O(h^m_{\widehat{X},\Omega})$ distant from the solution of the infinite dimensional problem in Eq. (8.3).

To start we recall some basic properties of Φ_i and $\phi(x_i)$, for $i \in [n]$, already sketched in Sec. 8.2 . In particular, the next proposition shows that, by construction, $\Phi_i^\top \Phi_j = \phi(x_i)^\top \phi(x_j)$ for any $i, j \in [n]$ and more generally that the map V that maps $f \in \mathcal{H} \mapsto R^{-\top}(\langle \phi(x_1), f\rangle, \dots, \langle \phi(x_n), f\rangle) \in \mathbb{R}^n$ is a partial isometry and that $\Phi_i = V\phi(x_i)$. The map V will be crucial to characterize the properties of the finite dimensional version of the operator A_*

Lemma 8.2 (Characterizing Φ_j in terms of ϕ)**.** *Let k be a kernel satisfying Assumption 8.2(a). There exists a linear operator $V : \mathcal{H} \to \mathbb{R}^n$ such that*

$$\Phi_i = V\phi(x_i), \qquad \forall i \in [n].$$

Moreover V is a partial isometry: VV^* is the identity on $\mathbb{R}^n$, $P = V^*V$ is a rank n projection operator satisfying $P\phi(x_i) = \phi(x_i), \forall i \in [n]$.

The proof of Lemma 8.2 is given in Sec. 8.C .1 in page 393 and is based on the fact that the kernel matrix K is positive definite and invertible when k is *universal* (Steinwart and Christmann, 2008), property that is implied by Assumption 8.2(a), and that R is an invertible matrix that satisfies $K = R^\top R$.

8.5 .1 Uniform inequality from scattered constraints

In this section we derive Theorem 8.4. Here we want to guarantee that a function g satisfies $g(x) \geq -\varepsilon$ on Ω, by imposing some constraints on $g(x_i)$ for $i \in [n]$. If we use the most natural discretization, that consists in the constraints $g(x_i) \geq 0$, by Lipschitzianity of g we can guarantee only $\varepsilon = |g|_{\Omega,1} h_{\widehat{X},\Omega}$ (recall the definition of $|\cdot|_{\Omega,m}$ for $m \in \mathbb{N}$ from Eq. (8.6)). In the case of *equality constraints*, instead, standard results for *functions with scattered zeros* (Wendland, 2004) (recalled in Sec. 8.B) guarantee for all $x \in \Omega$

$$|u(x)| \leq \varepsilon, \quad \varepsilon = C h^m_{\widehat{X},\Omega} |u|_{\Omega,m},$$

when u is m-times differentiable and satisfies $u(x_i) = 0$ for any $i \in [n]$ (see the work by Wendland (2004); Narcowich, Ward, and Wendland (2003) or Theorem 8.13 for more details). Thus, in this case the discretization leverages the degree of smoothness of u, requiring much less points to achieve a given ε than in the inequality case.

The goal here is to derive a guarantee for *inequality constraints* that is as strong as the one for the equality constraints. In particular, given a function g defined on Ω and that satisfies $g(x_i) - \Phi_i B \Phi_i = 0$ on $\widehat{X}$, with $B \succeq 0$, we first derive a function u defined on the whole Ω and matching $g(x_i) - \Phi_i B \Phi_i$ on $\widehat{X}$. This is possible since we know that $\Phi_i = V\phi(x_i)$, by Lemma 8.2, then $u(x) = g(x) - \langle \phi(x), V^*BV\, \phi(x) \rangle$ satisfies $u(x_i) = g(x_i) - \Phi_i B \Phi_i$ for any $i \in [n]$. Finally, we apply the results for functions with scattered zeros on u. The desired result is obtained by noting that, since $\langle \phi(x), V^*BV\, \phi(x) \rangle \geq 0$ for any $x \in \Omega$, by construction, then for all $x \in \Omega$

$$-g(x) \leq -g(x) + \langle \phi(x), V^*BV\, \phi(x) \rangle \leq |g(x) - \langle \phi(x), V^*BV\, \phi(x) \rangle| = |u(x)| \leq \varepsilon,$$

i.e., $g(x) \geq -\varepsilon$ for all $x \in \Omega$ with $\varepsilon = C h^m_{\widehat{X},\Omega} |u|_{\Omega,m}$. In the following theorem we provide a slightly more general result, that allows for $|g(x_i) - \Phi_i B \Phi_i| \leq \tau$ with $\tau \geq 0$.

Theorem 8.4 (Uniform inequality from scattered constraints)**.** *Let Ω satisfy Assumption 8.1(a) for some $r > 0$. Let k be a kernel satisfying Assumptions 8.2(a) and 8.2(d) for some $m \in \mathbb{N}_+$. Let $\widehat{X} = \{x_1, \ldots, x_n\} \subset \Omega$ with $n \in \mathbb{N}_+$ such that $h_{\widehat{X},\Omega} \leq r \min(1, \frac{1}{18(m-1)^2})$. Let $g \in C^m(\Omega)$ and assume there exists $B \in \mathbb{S}_+(\mathbb{R}^n)$ and $\tau \geq 0$ such that*

$$|g(x_i) - \Phi_i^\top B \Phi_i| \leq \tau, \quad \forall i \in [n], \tag{8.12}$$

where the Φ_i's are defined in Sec. 8.2 . The following statement holds:

$$g(x) \geq -(\varepsilon + 2\tau)\ \forall x \in \Omega, \quad \text{where} \quad \varepsilon = C h^m_{\widehat{X},\Omega}, \tag{8.13}$$

and $C = C_0(|g|_{\Omega,m} + \mathsf{M}\mathsf{D}_m \operatorname{Tr}(B))$ with $C_0 = 3\frac{\max(\sqrt{d}, 3\sqrt{2d}(m-1))^{2m}}{m!}$. The constants $m, \mathsf{M}, \mathsf{D}_m$, defined in Assumptions 8.2(a) and 8.2(d), do not depend on $n, \widehat{X}, h_{\widehat{X},\Omega}, B$ or g.

Proof. Let the partial isometry $V : \mathcal{H} \to \mathbb{R}^n$ and the projection operator $P = V^*V$ be defined as in Lemma 8.2. Given $B \in \mathbb{S}_+(\mathbb{R}^n)$ satisfying Eq. (8.12), define the operator $A \in \mathbb{S}_+(\mathcal{H})$ as $A = V^*BV$ and the functions $u, r_A : \Omega \to \mathbb{R}$ as follows

$$r_A(x) = \langle \phi(x), A\phi(x) \rangle, \quad u(x) = g(x) - r_A(x), \quad \forall x \in \Omega.$$

Since $\Phi_i = V\phi(x_i)$ for all $i \in [n]$, then for all $i \in [n]$:

$$r_A(x_i) = \langle \phi(x_i), V^*BV\phi(x_i) \rangle = (V\phi(x_i))^\top B(V\phi(x_i)) = \Phi_i^\top B\Phi_i,$$

and hence $u(x_i) = g(x_i) - \Phi_i^\top B\Phi_i$. Thus, $|u(x_i)| \leq \tau$ for any $i \in [n]$. This allows to apply one of the classical results on functions with scattered zeros (Narcowich, Ward, and Wendland, 2003; Wendland, 2004) to bound $\sup_{x\in\Omega} |u(x)|$, which we derived again in Theorem 8.13 to obtain explicit constants. Since we have assumed $h_{\widehat{X},\Omega} \leq r/\max(1, 18(m-1)^2)$, applying Theorem 8.13, the following holds

$$\sup_{x\in\Omega} |u(x)| \leq 2\tau + \varepsilon, \qquad \varepsilon = c\, R_m(u)\, h_{\widehat{X},\Omega}^m,$$

where $c = 3\max(1, 18(m-1)^2)^m$ and $R_m(v) = \sum_{|\alpha|=m} \frac{1}{\alpha!} \sup_{x\in\Omega} |\partial^\alpha v(x)|$ for any $v \in C^m(\Omega)$ using the multi-index notation (recalled in Sec. 8.3 .1). Since $r_A(x) = \langle \phi(x), A\phi(x) \rangle \geq 0$ for any $x \in \Omega$ as $A \in \mathbb{S}_+(\mathcal{H})$, it holds :

$$g(x) \geq g(x) - r_A(x) = u(x) \geq -|u(x)| \geq -(2\tau + \varepsilon), \qquad \forall x \in \Omega. \tag{8.14}$$

The last step is bounding $R_m(u)$. Recall the definition of $|\cdot|_{\Omega,m}$ from Eq. (8.6). First, note that $A = V^*BV$ is finite rank (hence trace-class). Applying the cyclicity of the trace and the fact that VV^* is the identity on $\mathbb{R}^n$, it holds

$$\operatorname{Tr}(A) = \operatorname{Tr}(V^*BV) = \operatorname{Tr}(BVV^*) = \operatorname{Tr}(B).$$

Since k satisfies Assumption 8.2(a), by Lemma 8.9, page 392, $r_A \in \mathcal{H}$ and $\|r_A\|_{\mathcal{H}} \leq \mathsf{M} \operatorname{Tr}(A) = \mathsf{M} \operatorname{Tr}(B)$ where M is fixed in Assumption 8.2(a). Moreover, since the kernel k satisfies Assumption 8.2(d) with m and D_m, then $|v|_{\Omega,m} \leq \mathsf{D}_m \|v\|_{\mathcal{H}}$, for any $v \in \mathcal{H}$ as recalled in Remark 25. In particular, this implies $|r_A|_{\Omega,m} \leq \mathsf{D}_m \|r_A\|_{\mathcal{H}} \leq \mathsf{D}_m \mathsf{M} \operatorname{Tr}(B)$. To conclude, note that, by the multinomial theorem,

$$R_m(u) = \sum_{|\alpha|=m} \frac{1}{\alpha!} \sup_{x\in\Omega} \left|\partial^\alpha u(x)\right| \leq \sum_{|\alpha|=m} \frac{1}{\alpha!} |u|_{\Omega,m} = \frac{d^m}{m!} |u|_{\Omega,m}.$$

Since $|u|_{\Omega,m} \leq |g|_{\Omega,m} + |r_A|_{\Omega,m}$, combining all the previous bounds, it holds

$$\varepsilon \leq C_0 \left(|g|_{\Omega,m} + \mathsf{D}_m \mathsf{M} \operatorname{Tr}(B)\right) h_{\widehat{X},\Omega}^m, \qquad C_0 = 3\frac{d^m \max(1, 18(m-1)^2)^m}{m!}.$$

The proof is concluded by bounding ε in Eq. (8.14) with the inequality above. □

In the theorem above we used a domain satisfying Assumption 8.1(a) and a version of a bound for functions with scattered zeros (that we derived in Theorem 8.13 following the analysis in the work by Wendland (2004)), to have explicit and relatively small constants. However, by using different bounds for functions with scattered zeros, we can obtain the same result as Theorem 8.4, but with different assumptions on Ω (and different constants). For example, we can use Corollary 6.4 in the work by Narcowich, Ward, and Wendland (2003) to obtain a result that holds for $\Omega = [-1, 1]^d$ or Theorem 11.32 with $p = q = \infty, m = 0$ in the work by Wendland (2004) to obtain a result that holds for Ω with locally Lipschitz-continuous boundary.

8.5 .2 Convergence properties of the finite-dimensional problem

Now we use Theorem 8.4 to bound the error of Eq. (8.5). First, to apply Theorem 8.4 we need to prove the existence of at least one finite-dimensional $B \succeq 0$ that satisfies the constraints of Eq. (8.5) and such that the trace of B is independent of n and $h_{\widehat{X},\Omega}$. This is possible since we proved in Theorem 8.3 that there exists at least one finite rank operator A that solves Eq. (8.3) and thus satisfies its constraints, of which the ones in Eq. (8.5) constitute a subset. In the next lemma we construct $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$, such that $\langle \phi(x_i), A\phi(x_i)\rangle = \Phi_i^\top \overline{B} \Phi_i$. In particular, $\overline{B} = VA_*V^* = R^{-\top} C R^{-1}$, with $C_{i,j} = \langle \phi(x_i), A_*\phi(x_j)\rangle$ for $i,j \in [n]$, where A_* is one solution of Eq. (8.3) with minimum trace-norm, since the bound in Theorem 8.4 depends on the trace of the resulting matrix.

Lemma 8.3. *Let Ω be an open set and $\{x_1,\dots,x_n\} \subset \Omega$ with $n \in \mathbb{N}_+$. Let $g : \Omega \to \mathbb{R}$ and k be a kernel on Ω. Denote by $\mathcal{H}$ the associated RKHS and by ϕ the associated canonical feature map. Let $A \in \mathbb{S}_+(\mathcal{H})$ satisfy $\operatorname{Tr}(A) < \infty$ and $\langle \phi(x), A\phi(x)\rangle = g(x)$, $x \in \Omega$. Then there exists $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$ such that $\operatorname{Tr}(\overline{B}) \leq \operatorname{Tr}(A)$ and $g(x_i) = \Phi_i^\top \overline{B} \Phi_i$, $\forall i \in [n]$.*

Proof. Let $V : \mathcal{H} \to \mathbb{R}^n$ be the partial isometry defined in Lemma 8.2 and $P = V^*V$ be the associated projection operator. Define $\overline{B} \in \mathbb{R}^{n\times n}$ as $\overline{B} = VAV^*$. Since by Lemma 8.2, $\Phi_i = V\phi(x_i)$ and P satisfies $P\phi(x_i) = \phi(x_i)$ for $i \in [n]$,

$$\begin{aligned}\Phi_i^\top \overline{B} \Phi_i &= (V\phi(x_i))^\top (VAV^*)(V\phi(x_i)) = \left\langle V^*V\phi(x_i), AV^*V\phi(x_i)\right\rangle \\ &= \left\langle P\phi(x_i), AP\phi(x_i)\right\rangle = \langle \phi(x_i), A\phi(x_i)\rangle \quad \forall i \in [n].\end{aligned}$$

Note that $\overline{B}$ satisfies: (a) $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$, by construction; (b) the requirement $\Phi_i^\top \overline{B}\Phi_i = g(x_i)$, indeed $\Phi_i^\top \overline{B}\Phi_i = \langle \phi(x_i), A\phi(x_i)\rangle$ and $\langle \phi(x), A\phi(x)\rangle = g(x)$ for any $x \in \Omega$; (c) $\operatorname{Tr}(\overline{B}) \leq \operatorname{Tr}(A)$, indeed, by the cyclicity of the trace,

$$\operatorname{Tr}(\overline{B}) = \operatorname{Tr}(VAV^*) = \operatorname{Tr}(AV^*V) = \operatorname{Tr}(AP).$$

The proof is concluded by noting that, since $A \succeq 0$ and $\|P\|_{\text{op}} \leq 1$ because P is a projection, then $\operatorname{Tr}(AP) \leq \|P\|_{\text{op}} \operatorname{Tr}(|A|) = \|P\|_{\text{op}} \operatorname{Tr}(A) \leq \operatorname{Tr}(A)$. □

We are now ready to prove the convergence rates of Eq. (8.5) to the global minimum. We will use the bound for the inequality on scattered data that we derived Theorem 8.4 and the fact that there exists $\overline{B} \succeq 0$ that satisfies the constraints of Eq. (8.5) with a trace bounded by $\operatorname{Tr}(A_*)$ as we proved in the lemma above (that is in turn bounded by the the trace of the operator explicitly constructed in Theorem 8.2). The proof is organized as follows. We will first show that Eq. (8.5) admits a minimizer, that we denote by $(\hat{c}, \hat{B})$. The existence of $\overline{B}$ allows to derive a lower-bound on $\hat{c} - f_*$. Using Theorem 8.4 on the constraints of Eq. (8.5) and evaluating the resulting inequality in one minimizer ζ of f allows to find an upper bound on $\hat{c} - f_*$ and an upper bound for $\operatorname{Tr}(\hat{B})$.

Theorem 8.5 (Convergence rates of Eq. (8.5) to the global minimum)**.** *Let Ω be a set satisfying Assumption 8.1(a) for some $r > 0$. Let $n \in \mathbb{N}_+$ and $\widehat{X} = \{x_1,\dots,x_n\} \subset \Omega$ with fill distance $h_{\widehat{X},\Omega}$. Let k be a kernel and $\mathcal{H}$ the associated RKHS satisfying Assumption 8.2 for some $m \in \mathbb{N}_+$. Let f be a function satisfying Assumption 8.1(b) and Assumption 8.3 for $\mathcal{H}$. The problem in Eq. (8.5) admits a solution. Let $(\hat{c}, \hat{B})$ be any solution of Eq. (8.5), for a given $\lambda > 0$. The following holds*

$$|\hat{c} - f_*| \;\leq\; 2\eta\, |f|_{\Omega,m} \;+\; \lambda \operatorname{Tr}(A_*), \qquad \eta \;=\; C_0\, h_{\widehat{X},\Omega}^m, \tag{8.15}$$

when $h_{\widehat{X},\Omega} \leq r\min(1, \frac{1}{18(m-1)^2})$ *and* $\lambda \geq 2\mathsf{M}\mathsf{D}_m\eta$. *Here* $C_0 = 3\frac{\max(\sqrt{d},3\sqrt{2d}(m-1))^{2m}}{m!}$, D_m, M *are defined in Assumption 8.2 and* A_* *is given by Theorem 8.3. Moreover, under the same conditions*

$$\mathrm{Tr}(\hat{B}) \;\leq\; 2\;\mathrm{Tr}(A_*) \;+\; 2\tfrac{\eta}{\lambda}\,|f|_{\Omega,m}. \tag{8.16}$$

Proof. We divide the proof in few steps.

Step 0. Problem Eq. (8.5) admits always a solution.

(a) On the one hand, c *cannot be larger than* $c_0 = \min_{i\in[n]} f(x_i)$, otherwise there would be a point x_j for which $f(x_j) - c < 0$ and so the constraint $\Phi_j^\top B\Phi_j = f(x_j) - c$ would be violated, since does not exist any positive semi-definite matrix for which $\Phi_j^\top B\Phi_j < 0$.

(b) On the other, *there exists an admissible point.* Indeed let (c_*, A_*) be the solution of Eq. (8.3) such that A_* has minimum trace norm. By Theorem 8.3, we know that this solution exists with $c_* = f_*$, under Assumptions 8.1 to 8.3. Then, by Lemma 8.3 applied to $g(x) = f(x) - c_*$ and $A = A_*$, given $\widehat{X} = \{x_1, \ldots, x_n\}$ we know that there exists $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$ satisfying $\mathrm{Tr}(\overline{B}) \leq \mathrm{Tr}(A_*)$ such that the constraints of Eq. (8.5) are satisfied for $c = c_*$. Then $(c_*, \overline{B})$ is admissible for the problem in Eq. (8.5).

Thus, since there exists an admissible point for the constraints of Eq. (8.5) and its functional cannot be larger than c_0 without violating one constraint, the SDP problem in Eq. (8.5) admits a solution (see the work by Boyd and Vandenberghe (2004)).

Step 1. Consequences of existence of A_***.** Let $(\hat{c}, \hat{B})$ be one minimizer of Eq. (8.5). The existence of the admissible point $(c_*, \overline{B})$ proven in the step above implies that

$$\hat{c} - \lambda\,\mathrm{Tr}(\hat{B}) \geq c_* - \lambda\,\mathrm{Tr}(\overline{B}) \geq f_* - \lambda\,\mathrm{Tr}(A_*),$$

from which we derive,

$$\lambda\,\mathrm{Tr}(\hat{B}) - \lambda\,\mathrm{Tr}(A_*) \leq \Delta, \quad \Delta := \hat{c} - f_*. \tag{8.17}$$

Step 2. $\boldsymbol{f|_\Omega \in C^{m+2}(\Omega)}$**.** Assumption 8.3 guarantees that $f|_\Omega \in C^2(\Omega)$ and that for all $i, j \in [d]$, $\frac{\partial}{\partial x_i \partial x_j} f|_\Omega \in \mathcal{H}$. Since under Assumption 8.2(d), $\mathcal{H} \subset C^m(\Omega)$ by Remark 25, we see that $\frac{\partial}{\partial x_i \partial x_j} f|_\Omega \in C^m(\Omega)$ for all $i, j \in [d]$ and hence $f|_\Omega \in C^{m+2}(\Omega)$.

Step 3. L^∞ **bound due to the scattered zeros.** Let $(\hat{c}, \hat{B})$ be one minimizer of Eq. (8.5) and define $\hat{g}(x) = f(x) - \hat{c}$ for all $x \in \Omega$. Note that $\hat{g}(x_i) = \Phi_i^\top \hat{B}\Phi_i$ for $i \in [n]$. Moreover, $\hat{g} \in C^m(\Omega)$ because $f \in C^m(\Omega)$ and $\hat{c}$ is a constant. Considering that $h_{\widehat{X},\Omega} \leq \frac{r}{\max(1,18(m-1)^2)}$, by assumption, then all the conditions in Theorem 8.4 are satisfied for $g = \hat{g}$, $\tau = 0$ and $B = \hat{B}$. Applying Theorem 8.4, we obtain,

$$\forall x \in \Omega,\; f(x) - \hat{c} = \hat{g}(x) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{M}\mathsf{D}_m\,\mathrm{Tr}(\hat{B})), \qquad \eta = C_0 h_{\widehat{X},\Omega}^m,$$

where C_0 is defined in Theorem 8.4. Since the inequality above holds for any $x \in \Omega$, by evaluating it in one global minimizer $\zeta \in \Omega$, we have $f(\zeta) = f_*$ and hence

$$-\Delta = f_* - \hat{c} = f(\zeta) - \hat{c} = \hat{g}(\zeta) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{M}\mathsf{D}_m\,\mathrm{Tr}(\hat{B})).$$

Since $\hat{g} = f - \hat{c}\mathbf{1}_\Omega$, and since for any $m \in \mathbb{N}_+$, $|\mathbf{1}_\Omega|_{\Omega,m} = 0$, we have $|\hat{g}|_{\Omega,m} \leq |f|_{\Omega,m} + |\mathbf{1}_\Omega|_{\Omega,m} = |f|_{\Omega,m}$. Injecting this in the previous equation yields

$$\Delta \leq \eta|f|_{\Omega,m} + \eta\mathsf{M}\mathsf{D}_m\,\mathrm{Tr}(\hat{B}). \tag{8.18}$$

Conclusion. Combining Eq. (8.18) with Eq. (8.17), and since $\lambda \geq 2\mathsf{MD}_m\eta$ by assumption,

$$\tfrac{\lambda}{2}\operatorname{Tr}(\hat{B}) \leq (\lambda - \mathsf{MD}_m\eta)\operatorname{Tr}(\hat{B}) \leq \eta|f|_{\Omega,m} + \lambda\operatorname{Tr}(A_*).$$

Note that Eq. (8.16) is obtained from the one above, by dividing by $\frac{\lambda}{2}$. Finally the inequality Eq. (8.15) is derived by bounding Δ from below as $\Delta \geq -\lambda\operatorname{Tr}(A_*)$ by Eq. (8.17), since $\operatorname{Tr}(\hat{B}) \geq 0$ by construction, and bounding it from above as

$$\Delta \leq 2\eta|f|_{\Omega,m} + \lambda\operatorname{Tr}(A_*),$$

obtained by combining Eq. (8.18) with Eq. (8.16) and with the assumption $\mathsf{MD}_m\eta \leq \frac{\lambda}{2}$. □

The result above holds for any kernel satisfying Assumption 8.2 and any function f, Ω satisfying the geometric conditions in Assumption 8.1 and with $f \in C^2(\Omega)$ and $\frac{\partial^2 f}{\partial x_i \partial x_j} \in \mathcal{H}$ for $i, j \in [d]$. The latter requirement is quite easy to verify for example when $\mathcal{H}$ contains $C^s(\Omega)$ and $f \in C^{s+2}(\Omega)$ for some $s > 0$ as in the case of $\mathcal{H}$ being a Sobolev space with $s > d/2$. Moreover the proposed result holds for any discretization $\widehat{X}$ (random, or deterministic). We would like to conclude with the following remark on the sufficiency of the assumptions on f.

Remark 26 (Sufficiency of Assumptions 8.1(b) and 8.3)**.** *Assumptions Assumptions 8.1(b) and 8.3 are sufficient for Theorems 8.3 and 8.5 to hold. However, by inspecting their proof it is clear that they hold by requiring only the existence of a trace-class operator $A_* \in \mathbb{S}_+(\mathcal{H})$ such that $f(x) - f_* = \langle \phi(x), A_*\phi(x)\rangle$ for any $x \in \Omega$, where $f_* = \inf_{x\in\Omega} f(x)$. Note that this is implied by Assumptions 8.1(b) and 8.3 via Cor. 8.1.*

In the next subsection we are going to apply the theorem above to the specific setting of algorithm 6.

8.5 .3 Result for Sobolev kernels and discussion

In this we are going to apply Theorem 8.5 to algorithm 6 which corresponds to $\mathcal{H}$ be the Sobolev space of smoothness s and the points $\widehat{X}$ selected independently and uniformly at random. First, in the next lemma we bound in high probability the fill distance $h_{\widehat{X},\Omega}$ with respect to the number of points n that we sample, i.e., the cardinality of $\widehat{X}$.

Lemma 8.4 (Random sets of points)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded set with diameter $2R$, for some $R > 0$, and satisfying Assumption 8.1(a) for a given $r > 0$. Let $\widehat{X} = \{x_1, \ldots, x_n\}$ independent points sampled from the uniform distribution on Ω. When $n \geq 2(\frac{6R}{r})^d \left(\log\frac{2}{\delta} + 2d\log\frac{4R}{r}\right)$, then the following holds with probability at least $1 - \delta$:*

$$h_{\widehat{X},\Omega} \;\leq\; 11R\, n^{-\frac{1}{d}}\, (\log\tfrac{n}{\delta} + d\log\tfrac{2R}{r})^{1/d}.$$

The proof of Lemma 8.4 is in Sec. 8.E .1, page 401 and is a simpler version (with explicit constants) of more general results (Penrose, 2003, Thm. 13.7). In the next theorem we apply the bound in the lemma above with the explicit constants for Sobolev spaces derived in proposition 8.1 to Theorem 8.5. The derivation of the theorem below is in Sec. 8.E .2, page 403.

Theorem 8.6 (Convergence rates of algorithm 6 to the global minimum)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded set with diameter $2R$, for some $R > 0$, and satisfying Assumption 8.1(a) for a given $r \in (0, R]$ (e.g. if Ω is a ball with radius R, then $r = R$). Let s satisfying $s > d/2$. Let k be Sobolev kernel of smoothness s (see Example 8.1). Assume that f satisfies Assumption 8.1(b) and that $f|_\Omega \in W_2^{s+2}(\Omega)$. Let $\hat{c}$ be the result of algorithm 6 executed with $n \in \mathbb{N}_+$ points chosen*

uniformly at random in Ω *and* $\lambda > 0$. *Let* $\delta \in (0,1]$. *When* $m \in \mathbb{N}_+$ *satisfies* $m < s - d/2$ *and* $n \geq \max(4, 15(m-1))^{2d} \left(\frac{R}{r}\right)^d \left(2\log\frac{2}{\delta} + 4d\log\frac{20R\ m}{r}\right)$ *choose any* λ *satisfying*

$$\lambda \ \geq \ n^{-\frac{m}{d}} \ (\log\tfrac{2^d n}{\delta})^{\frac{m}{d}} \ R^m C_{m,s,d},$$

where $C_{m,s,d} = 11^m C_0 \max(1, \mathsf{MD}_m)$ *with* C_0 *defined in Theorem 8.5 and* MD_m *defined in proposition 8.1. Note that* $C_{m,s,d}$ *is explicitely bounded in the proof in page 403 in terms of* s, m, d. *Then, with probability at least* $1-\delta$, *the following holds*

$$|\hat{c} - f_*| \ \leq \ 3\,\lambda\,(\mathrm{Tr}(A_*) + |f|_{\Omega,m})\,.$$

A direct consequence of the theorem above, already stated in Remark 24, is the nearly-optimality of algorithm 6 for the cases of Sobolev functions. Indeed by applying Theorem 8.6 with m equal to the largest integer strictly smaller than $s - d/2$ we have that $m \geq s - d/2 - 1$, and so algorithm 6 achieves the global minimum with a rate that is $O(n^{-\frac{s}{d}+\frac{1}{2}+\frac{1}{d}})$. The lower bounds from information based complexity state that, by observing the functions in n points, it is not possible to find the minimum with error smaller than $n^{-\frac{s}{d}+\frac{1}{2}}$ for functions in $W_2^s(\Omega)$ (see, e.g., the work by Novak (2006), Prop. 1.3.11, page 36). Since in Theorem 8.6 we assume f belongs to $W_2^{s+2}(\Omega)$, the optimal rate would be $n^{-\frac{s}{d}+\frac{1}{2}-\frac{2}{d}}$ so we are a factor $n^{3/d}$ slower than the optimal rate. Note that this factor is negligible if the function is very smooth, i.e., $s \gg d$, or d is very large. An interesting corollary that corresponds to Theorem 8.1, can be derived considering that $C^{s+2}(\Omega) \subseteq W_2^{s+2}(\Omega)$, since Ω is bounded.

8.6 Algorithm

We need to solve the following optimization problem:

$$\max_{B \succcurlyeq 0, c \in \mathbb{R}} c - \ \lambda\,\mathrm{Tr}(B) \ \text{ such that } \ f(x_i) - c - \Phi_i^\top B \Phi_i = 0, \ \forall i \in [n].$$

This is a semi-definite programming problem with n constraints and a semi-definite constraint of size n. It can thus be solved with precistion ε in time $O(n^{3.5}\log(1/\varepsilon))$ and memory $O(n^2)$ by standard software packages (Boyd and Vandenberghe, 2004). However, to allow applications to $n = 1000$ or more, and on parallel architectures, we provide a simple Newton algorithm, which relies on penalization by a self-concordant barrier, that is, we aim to solve

$$\max_{B \succcurlyeq 0, c \in \mathbb{R}} c - \ \lambda\,\mathrm{Tr}(B) + \frac{\varepsilon}{n}\log\det(B) \ \text{ such that } \ f(x_i) - c - \Phi_i^\top B \Phi_i = 0, \ \forall i \in [n],$$

for which we know that at optimum, the deviation with the optimal value is at most ε (Nemirovski, 2004, Sec. 4.4). By standard Lagrangian duality, we get, with $\Phi \in \mathbb{R}^{n\times n}$ the matrix with rows $\Phi_1, \ldots, \Phi_n$, so that $\Phi\Phi^\top = K$:

$$\begin{aligned}
&\sup_{B\succcurlyeq 0, c} \inf_{\alpha \in \mathbb{R}^n} c + \sum_{i=1}^n \alpha_i \big(f(x_i) - c - \Phi_i^\top B \Phi_i\big) - \lambda\,\mathrm{Tr}(B) + \frac{\varepsilon}{n}\log\det(B) \\
= \ &\inf_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i f(x_i) - \frac{\varepsilon}{n}\log\det\left(\Phi^\top \mathrm{Diag}(\alpha)\Phi + \lambda I\right) + \frac{\varepsilon}{n}\log\frac{\varepsilon}{n} - \varepsilon \ \text{ s. t. } \ \alpha^\top 1_n = 1.
\end{aligned}$$

With the barrier term, this thus defines a dual function $H(\alpha)$, and we get the following gradient

$$H'(\alpha)_i = f_i - \frac{\varepsilon}{n}\,\Phi_i^\top \Big(\Phi^\top \mathrm{Diag}(\alpha)\Phi + \lambda I\Big)^{-1}\Phi_i = f_i - \frac{\varepsilon}{n\alpha_i}\Big[K(K + \lambda \mathrm{Diag}(\alpha)^{-1})^{-1}\Big]_{ii},$$

and Hessian

$$H''(\alpha)_{ij} = \frac{\varepsilon}{n}\big[\Phi_i^\top \Big(\Phi^\top \mathrm{Diag}(\alpha)\Phi + \lambda I\Big)^{-1}\Phi_j\big]^2,$$

which can be rewritten

$$H''(\alpha)_{ij} = \frac{\varepsilon}{n\alpha_j\alpha_i}\Big[K(K+\lambda \mathrm{Diag}(\alpha)^{-1})^{-1}\Big]_{ij}\Big[K(K+\lambda \mathrm{Diag}(\alpha)^{-1})^{-1}\Big]_{ji}.$$

We can then compute the step for the Damped Newton algorithm: $\alpha^+ = \alpha - \frac{1}{1+\sqrt{\frac{n}{\varepsilon}\lambda(\alpha)}}\Delta$, where $\Delta = H''(\alpha)^{-1}H'(\alpha) - \frac{1_n^\top H''(\alpha)^{-1}H'(\alpha)}{1_n^\top H''(\alpha)^{-1}1_n}H''(\alpha)^{-1}1_n$ and $\lambda(\alpha)^2 = \Delta^\top H''(\alpha)\Delta$ is the Newton decrement (which can serve as a stopping criterion). Note that the algorithm is always feasible, without a need for any eigenvalue decomposition. The overall complexity is $O(n^3)$ per iteration due to matrix inversions and linear systems. Note that the conditioning of these linear systems is at least as bad as the conditioning of the kernel matrix K. Fortunately, for the s-th Sobolev kernels in dimension d, the m-th eigenvalue of the kernel matrix typically decay as $m^{-2s/d}$ (Bach, 2017a, Sec. 2.3).

Retrieving c and B. From an optimal α, we can recover $B = \frac{\varepsilon}{n}\big(\Phi^\top \mathrm{Diag}(\alpha)\Phi + \lambda I\big)^{-1} = \frac{\varepsilon}{n\lambda}\big(I - \Phi^\top(\Phi\Phi^\top + \lambda \mathrm{Diag}(\alpha)^{-1})^{-1}\Phi\big)$ and $c = \frac{1}{n}H'(\alpha)^\top 1_n$ (since c is the Lagrange multiplier for the constraint $\alpha^\top 1_n = 1$). Thus, computing the model for a test point, can be done as $\frac{\varepsilon}{n\lambda}\big(k(x,x) - q(x)^\top(K+\lambda \mathrm{Diag}(\alpha)^{-1})^{-1}q(x)\big)$, where $q(x)_i = k(x,x_i)$. Alternatively, when Φ is invertible, we can use $q(x)^\top \Phi^{-\top} B \Phi^{-1} q(x)$.

Retrieving a minimizer. Given the dual solution, based on our localizing arguments presented in Sec. 8.7 , a good candidate solution will be

$$\hat{z} = \sum_{i=1}^n \alpha_i x_i \tag{8.19}$$

A more principled way to find a minimizer is provided in Sec. 8.7 , of which the equation above corresponds to the limit solution of Eq. (8.23) for $\nu \to 0$ (see Sec. 8.7 .1).

Number of iterations. In order to reach a Newton decrement $n^{1/2}\varepsilon^{-1/2}\lambda(\alpha) \leqslant \kappa$, a number of steps equal to a universal constant times $\frac{n}{\varepsilon}[H(\alpha_0) - H(\alpha_*)] + \log\log\frac{1}{\kappa}$ is sufficient (Nemirovski, 2004).

When initializing with $\alpha_0 = \frac{1}{n}1_n$, we have $H(\alpha_0) = \frac{1}{n}\sum_{i=1}^n f_i - \frac{\varepsilon}{n}\log\det\big(K + n\lambda I\big) + \frac{\varepsilon}{n}\log\varepsilon - \varepsilon$, and $H(\alpha_*) \geqslant c_* - \lambda \operatorname{Tr}(A_*) - \varepsilon$. This leads to a number of Newton steps less than

$$\frac{n}{\varepsilon}\big[\langle f\rangle - \inf f\big] + \log\det\big(K + n\lambda I\big) + \frac{n}{\varepsilon}\lambda \operatorname{Tr}(A_*) + \log\varepsilon + \log\log\frac{1}{\kappa}.$$

In our experiments, we do not perform path following (that would lead the classical interior-point method) and instead fixed value $\varepsilon = 10^{-3}$, and a few hundred Newton steps.

Behavior for $\lambda = 0$. If the kernel matrix K is invertible (which is the case almost surely for Sobolev kernels and points sampled independently from a distribution with a density with respect to the Lebesgue measure), then we show that for $\lambda = 0$, the optimal value of of the finite-dimensional problem in Eq. (8.5) is equal to $\min_{i\in[n]} f(x_i)$. Since $f(x_i) \geq c + \Phi_i^\top B \Phi_i$ implies $f(x_i) \geq c$, the optimal value has to be less than $\min_{i\in[n]} f(x_i)$. We therefore just need to find a feasible B that achieves it. Since K is assumed invertible (and thus its Cholesky factor as well), we can simply take $B = R^{-\top}\mathrm{Diag}[(f(x_j) - \min_{i\in[n]} f(x_i))_j]R^{-1}$.

8.7 Finding the global minimizer

In this section we provide and study the problem in Eq. (8.23), that is a variation of the problem in Eq. (8.5), and allows to find also the minimizer of f as we prove in Theorem 8.8. As in Sec. 8.2 we start from a convex representation of the optimization problem and then we derive our sampled version, passing by an intermediate infinite-dimensional problem that is useful to derive the theoretical properties of the method. While the problem in Eq. (8.2) can be seen as finding the largest constant c such that $f - c$ is still non-negative, in the problem below we find the parabola of the form $p_{z,\gamma}(x) = \frac{\nu}{2}\|x\|^2 - \nu x^\top z + c = \frac{\nu}{2}\|x - z\|^2 + c - \frac{\nu}{2}\|z\|^2$ with the highest vertex such that $f - p_{z,c}$ is still non-negative. Since the height of the vertex of $p_{z,c}$ corresponds to $c - \frac{\nu}{2}\|z\|^2$, the resulting optimization problem is the following,

$$\max_{c\in\mathbb{R}, z\in\mathbb{R}^d} c - \tfrac{\nu}{2}\|z\|^2 \quad \text{such that} \quad f(x) - \tfrac{\nu}{2}\|x\|^2 + \nu x^\top z - c \geq 0 \quad \forall x \in \Omega. \tag{8.20}$$

It is easy to see that if $f \in C^2(\mathbb{R}^d)$ has a unique minimizer ζ that belongs to Ω and is locally strongly convex around ζ then there exists a $\nu > 0$ such that the problem above achieves an optimum (c_*, z_*) with $z_* = \zeta$ and $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$. In particular, to characterize ν explicitly we introduce the stronger assumption below.

Assumption 8.4 (Geometric assumption to find global minimizer)**.** *The function $f : \mathbb{R}^d \to \mathbb{R}$ has a unique global minimizer in Ω.*

If f satisfies Assumptions 8.1(b) and 8.4, denote with ζ the unique minimizer of f in Ω and with $f_* = f(\zeta)$ the corresponding minimum.

Remark 27. *Under Assumptions 8.1(b) and 8.4 f can be lower bounded by a parabola with value f_* at ζ, i.e., there exists $\beta > 0$ such that*

$$\forall x \in \Omega, \;\; f(x) - f_* \geq \tfrac{\beta}{2}\|x - \zeta\|^2. \tag{8.21}$$

The remark above is derived in Sec. 8.F .1, page 404. In what follows, whenever f satisfies Assumptions 8.1(b) and 8.4, then β will be assumed to be the supremum among the value satisfying Eq. (8.21). Now we are ready to summarize the reasoning above on the fact that Eq. (8.20) achieves the minimizer of f.

Lemma 8.5. *Suppose f satisfies Assumptions 8.1 and 8.4. Let ζ be the unique minimizer of f in Ω and $f_* = f(\zeta)$ be the corresponding minimum. Let $\beta > 0$ such that Eq. (8.21) holds. If $\nu < \beta$ then the problem in Eq. (8.20) has a unique solution (c_*, z_*) such that $z_* = \zeta$ and $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$.*

The lemma above guarantees that the problem in Eq. (8.20) achieves the global minimum and the global minimizer of f, when f satisfies the geometric conditions Assumptions 8.1 and 8.4. Now, as we did for Eq. (8.2), we consider the following problem of which Eq. (8.20) is a tight relaxation.

$$\begin{aligned} \max_{c\in\mathbb{R}, z\in\mathbb{R}^d, A\in\mathbb{S}_+(\mathcal{H})} \; & c - \tfrac{\nu}{2}\|z\|^2 \\ \text{such that } & f(x) - \tfrac{\nu}{2}\|x\|^2 + \nu x^\top z - c \;\; = \langle \phi(x), A\phi(x)\rangle \quad \forall x \in \Omega. \end{aligned} \tag{8.22}$$

Indeed, since $\langle \phi(x), A\phi(x)\rangle \geq 0$ for any $x \in \Omega$ and $A \in \mathbb{S}_+(\mathcal{H})$, for any triplet (c, z, A) satisfying the constraints in the problem above, the couple (c, z) satisfies the constraints in Eq. (8.20). The contrary may be not true in general. In the next theorem we prove that when $\mathcal{H}$ satisfies Assumption 8.2 and Ω, f satisfy Assumptions 8.1, 8.3 and 8.4, then the relaxation is tight

and, in particular, when $\nu < \beta$, there exists a finite rank operator A_* such that the triplet $(f_* + \frac{\nu}{2}\|\zeta\|^2, \zeta, A_*)$ is optimal.

Theorem 8.7. *Let $\Omega \subset \mathbb{R}^d$ be an open set, k be a kernel, $\mathcal{H}$ the associated RKHS, and $f : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumptions 8.1 to 8.3, and Assumption 8.4. Let β satisfying Eq. (8.21). For any $\nu < \beta$, the problem in Eq. (8.22) admits an optimal solution (c_*, z_*, A_*) with $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$, $z_* = \zeta$, and A_* a positive semi-definite operator on $\mathcal{H}$ with rank at most $d + 1$.*

The proof of the theorem above is essentially the same of Theorem 8.3 and is reported for completeness in Sec. 8.F .2, page 404. In particular, to prove the existence of A_* we applied Cor. 8.1 to the function $f(x) - \frac{\nu}{2}\|x - \zeta\|^2$ that still satisfies Assumptions 8.1 and 8.3 when f does and $\nu < \beta$. Now we are ready to consider the finite-dimensional version of Eq. (8.22). Given a set of points $\widehat{X} = \{x_1, \dots, x_n\}$ with $n \in \mathbb{N}_+$,

$$\begin{aligned} \max_{c\in\mathbb{R}, z\in\mathbb{R}^d, B\in\mathbb{S}_+(\mathbb{R}^n)} \; & c - \tfrac{\nu}{2}\|z\|^2 - \lambda \operatorname{Tr}(B) \\ \text{such that} \quad & \forall i \in [n],\ f(x_i) - \tfrac{\nu}{2}\|x_i\|^2 + \nu x_i^\top z - c = \Phi_i^\top B \Phi_i. \end{aligned} \tag{8.23}$$

For the problem above we can derive similar convergence guarantees as for Eq. (8.5) and also a convergence of the estimated minimizer z to ζ, as reported in the following theorem.

Theorem 8.8 (Convergence rates of Eq. (8.23) to the global minimizer)**.** *Let Ω be a set satisfying Assumption 8.1(a) for some $r > 0$. Let $\widehat{X} = \{x_1, \dots, x_n\} \subset \Omega$ with fill distance $h_{\widehat{X},\Omega}$. Let k be a kernel satisfying Assumption 8.2 for some $m \geq 2$ and f satisfying Assumptions 8.1, 8.3 and 8.4. The problem in Eq. (8.23) admits a solution. Denote by $(\hat{c}, \hat{z}, \hat{B})$ any solution of Eq. (8.23), for a given $\lambda > 0$. Then for any $\nu < \beta$,*

$$\tfrac{\nu}{2}\|\hat{z} - \zeta\|^2 \leq 3\eta(|f|_{\Omega,m} + \nu) + 2\lambda \operatorname{Tr}(A_*), \qquad \eta \;=\; C\, h_{\widehat{X},\Omega}^m, \tag{8.24}$$

when $h_{\widehat{X},\Omega} \leq \frac{r}{18(m-1)^2}$ and $\lambda \geq 2\mathsf{M}\mathsf{D}_m\eta$. Here $C = 3\frac{(3\sqrt{2d}(m-1))^{2m}}{m!}$ and D_m, M are defined in Assumption 8.2. A_* is from Theorem 8.7. Moreover under the same conditions

$$|\hat{c} - \tfrac{\nu}{2}\|\hat{z}\|^2 - f_*| \;\leq\; 2\eta\,|f|_{\Omega,m} \;+\; \lambda\,\operatorname{Tr}(A_*) \;+\; 2\eta\nu, \tag{8.25}$$

$$\operatorname{Tr}(\hat{B}) \;\leq\; 2\,\operatorname{Tr}(A_*) \;+\; 2\tfrac{\eta}{\lambda}\,|f|_{\Omega,m} \;+\; 2\nu\tfrac{\eta}{\lambda}. \tag{8.26}$$

The proof of the theorem above is similar to the one of Theorem 8.5 and it is stated for completeness in Sec. 8.F .3, page 406. The same comments to Theorem 8.5 that we reported in the related section and the rates for Sobolev functions, apply also in this case. In the next section we describe the algorithm to solve the problem in Eq. (8.23).

8.7 .1 Algorithm

We can use the same dual technique as presented in Sec. 8.6 , and obtain a dual problem to Eq. (8.23) with the additional penalty $\frac{\varepsilon}{n}\log\det B$. The dual problem can readily be obtained as (up to constants)

$$\inf_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \alpha_i f(x_i) - \frac{\varepsilon}{n}\log\det\left(\Phi^\top \operatorname{Diag}(\alpha)\Phi + \lambda I\right) + \frac{\nu}{2}\Big(- \sum_{i=1}^n \alpha_i \|x_i\|_2^2 + \Big\| \sum_{i=1}^n \alpha_i x_i \Big\|_2^2 \Big),$$

such that $\alpha^\top 1_n = 1$, with the optimal z that can be recovered as $z = \sum_{i=1}^n \alpha_i x_i$. We note that when ν tends to zero, we recover the dual problem from Sec. 8.6 , and we keep the candidate above in Ω even when $\nu = 0$.

8.7 .2 Warm restart scheme for linear rates

It is worth noting that Theorem 8.8 provides strong guarantees on the distance $\|\hat{z} - \zeta\|$ where $\hat{z}$ is the solution of the problem Eq. (8.23) and ζ the global optimum of f. This suggests that we can implement a warm restart scheme that leverage the additional knowledge of the position of ζ. Assume indeed that Ω is a ball of radius R centered in z_0. For $t = 1, \dots, T$ with $T = \lceil \log \frac{1}{\varepsilon} \rceil$, we apply Eq. (8.23) to a set $\widehat{X}_t$ that contains enough points sampled uniformly at random in the ball $B_{r_{t-1}}(z_{t-1})$ such that Theorem 8.8 guarantees that $\|z_t - \zeta\| \leq r_{t-1}/e$ where z_t is the solution of Eq. (8.23). The cycle is repeated with $r_t = r_{t-1}/e$ and the new center be z_t. By plugging the estimate of Lemma 8.4 for $h_{\widehat{X}_t, B_{r_{t-1}}(z_{t-1})}$ in Theorem 8.8 for each step t, we obtain a total number of points n to achieve $\|z_T - \zeta\| \leq \varepsilon$ with probability $1 - T\delta$, that is

$$n = O\Big(C_{d,m}^{d/m} \Big(\frac{\mathcal{F}}{\nu}\Big)^{d/m} R^d \log \frac{1}{\varepsilon}\Big)$$

modulo logarithmic terms in n and δ, where $C_{d,m} = 3^m C \mathsf{MD}_m$ with C defined in Theorem 8.8 and $\mathcal{F} = |f|_{\Omega,m} + \nu + \operatorname{Tr}(A_*)$. This means that under the additional assumption of a unique minimizer in Ω, we achieve a convergence rate that is only logarithmic in ε, moreover when $m \gg d$ also the dependence with respect to $C_{d,m}$ (which is exponential in m and d in the case of the Sobolev kernel) and $\mathcal{F}$ improves, since d/m tends to 0.

8.8 Extensions

In this section we deal with two aspects: (a) the effect of solving approximately the problem in Eq. (8.5), and (b) how can we certify explicitly (no dependence on quantities of theoretical interest as $\operatorname{Tr}(A_*)$) how close is a given (approximate) solution to the optimum; (c) we will also cover the case when the function f does not have a positive definite representer A_* in $\mathbb{S}_+(\mathcal{H})$ but in a larger space. This allows to cover the cases of $f \in C^s(\mathbb{R}^d)$ with $s \leq d/2 + 2$.

8.8 .1 Approximate solutions

In this section we extend Theorem 8.5 to consider the case when we solve Eq. (8.5) in an approximate way. In particular, let $\lambda > 0, n \in \mathbb{N}_+$ and $\widehat{X} = \{x_1, \dots, x_n\}$. Denote by $p_{\lambda,n}$ the optimal value achieved by Eq. (8.5) for such λ, n. We say that $(\tilde{c}, \tilde{B})$ is an *approximate solution* of Eq. (8.5) with parameters $\theta_1, \theta_2, \tau_1, \tau_2 \geq 0$ if it satisfies the following inequalities

$$p_{\lambda,n} \; - \; \tilde{c} + \lambda \operatorname{Tr}(\tilde{B}) \leq \theta_1 + \theta_2 \operatorname{Tr}(\tilde{B}), \tag{8.27}$$

$$|f(x_i) - \tilde{c} - \Phi_i^\top \tilde{B} \Phi_i| \leq \tau_1 + \tau_2 \operatorname{Tr}(\tilde{B}), \quad \forall i \in [n]. \tag{8.28}$$

Theorem 8.9 (Error of approximate solutions of Eq. (8.5))**.** *Let $(\tilde{c}, \tilde{B})$ be an approximate solution of Eq. (8.5) for a given $n \in \mathbb{N}_+, \lambda > 0$ as defined in Eqs. (8.27) and (8.28) w.r.t. $\tau_1, \tau_2, \theta_1, \theta_2 \geq 0$. Under the same assumptions and notation of Theorem 8.5 and Remark 26, when $\tau_2, \theta_2 \leq \frac{\lambda}{8}$*

$$|\tilde{c} - f_*| \; \leq \; 7(2\tau_1 + \eta |f|_{\Omega,m}) + 6(\theta_1 + \lambda \operatorname{Tr}(A_*)), \tag{8.29}$$

$$\operatorname{Tr}(\tilde{B}) \; \leq \; 8 \operatorname{Tr}(A_*) \; + \; 8\tfrac{\eta}{\lambda} |f|_{\Omega,m} \; + \; 8\tfrac{\theta_1 + 2\tau_1}{\lambda}. \tag{8.30}$$

The proof of the theorem above is reported for completeness in Sec. 8.G .1, page 407, and is a variation of the one of Theorem 8.5 where we used Theorem 8.4 with $\tau = \tau_1 + \tau_2 \operatorname{Tr}(\tilde{B})$ and we further bound $p_{\lambda,n}$ via Eq. (8.27). From a practical side, the theorem above allows to use a wide

range of methods and techniques to approximate the solution of Eq. (8.5). In particular, it is possible to use lower dimensional approximations of $\Phi_1, \dots, \Phi_n$ and algorithms based on early stopping as described in Sec. 8.11 , since $\tau_1, \tau_2, \theta_1, \theta_2$ will take into account the error incurred in the approximations. An interesting application of the theorem above, from a theoretical side is that it allows also to deal with situations where f does not have a representer A_* in $\mathbb{S}_+(\mathcal{H})$ as we are going to discuss in the next section.

8.8 .2 Rates for f with low smoothness

When $f \in C^{s+2}(\mathbb{R}^d)$ with $s \in \mathbb{N}$, but with a low smoothness, i.e., $s \leq d/2$, we can still apply our method to find the global minimum and obtain almost optimal convergence rates, as soon as it satisfies the geometric conditions in Assumption 8.1(b), as we are going to show in Theorem 8.11 and the following discussion.

In this section, for any function u defined on a super-set of Ω and s times differentiable on Ω, we define the following semi norm :

$$\|u\|_{\Omega,s} = \max_{|\alpha| \leq s} \sup_{x \in \Omega} \left|\partial^\alpha u(x)\right|. \tag{8.31}$$

We consider the following variation of the problem in Eq. (8.5):

$$\max_{c \in \mathbb{R},\ B \in \mathbb{S}_+(\mathbb{R}^n)} c - \lambda \operatorname{Tr}(B) \quad \text{such that} \quad \forall i \in [n],\ |f(x_i) - c - \Phi_i^\top B \Phi_i| \leq \tau. \tag{8.32}$$

The idea is that f, under the geometric conditions in Assumption 8.1(b), still admits a decomposition in the form $f(x) = \sum_{j \in [p]} w_j(x)^2$, $p \in \mathbb{N}_+$ for any $x \in \Omega$, but now with respect to functions with low smoothness $w_1, \dots, w_p \in C^s(\mathbb{R}^d)$. To prove this we follow the same proof of Sec. 8.4 noting that the assumptions to apply Lemma 8.1 and Theorem 8.2 are that f belongs to a normed vector space space that satisfy the algebraic properties in Assumptions 8.2(a) to 8.2(c) which does not have necessarily to be a RKHS. In particular, note that the space $\widetilde{\mathcal{H}} = \{f|_\Omega \ : \ f \in C^s(\mathbb{R}^d)\}$ of restriction to Ω of functions in $C^s(\mathbb{R}^d)$, endowed with the norm $\|\cdot\|_{\Omega,s}$ defined in Eq. (8.31) (and is always finite on $\widetilde{H}$ since Ω is bounded) satisfies such assumptions. The reasoning above leads to the following corollary of Theorem 8.2 (the details can be found in Sec. 8.G .2 page 408).

Corollary 8.2. *Let Ω be a bounded open set and $f \in C^{s+2}(\mathbb{R}^d)$, $s \in \mathbb{N}$, satisfying Assumption 8.1(b). Then there exist $w_1, \dots, w_p \in C^s(\mathbb{R}^d)$, $p \in \mathbb{N}_+$, such that*

$$\forall x \in \Omega,\ f(x) - f_* = \sum_{j \in [p]} w_j^2(x).$$

By using the decomposition above, when the kernel satisfies Assumption 8.2(a), we build an operator $A_\varepsilon \in \mathbb{S}_+(\mathcal{H})$ that approximates f with error $O(\epsilon^s)$ for any $\epsilon > 0$. First note that, for any bounded open set $\Omega \subset \mathbb{R}^d$ and any $s \leq r$, there exists C_1 and C_2 depending only on r, s, Ω such that for any $g \in C^s(\mathbb{R}^d)$ and $\varepsilon > 0$ there exists a smooth approximation $g_\epsilon \in C^\infty(\mathbb{R}^d)$ such that $\sup_{x \in \Omega} |g(x) - g_\varepsilon(x)| \leq C_1 \varepsilon^s \|g\|_{\Omega,s}$ and such that $\|g_\varepsilon\|_{\Omega,r} \leq C_2 \varepsilon^{-(r-s)} \|g\|_{\Omega,s}$ (see Thm. 5.33 by Adams and Fournier (2003) for the more general case of Sobolev spaces, or (Cheney and Light, 2009, Chapter 21) for explicit construction in terms of convolutions with smooth functions). Denote by w_j^ε the smooth approximation of w_j on Ω for any $j \in [p]$. Since we consider kernels

rich enough that the associated RKHS $\mathcal{H}$ contains smooth functions (see Assumption 8.2(a)), then we have that $w_j^\varepsilon|_\Omega \in \mathcal{H}$ for any $j \in [p]$. Then

$$A_\varepsilon = \sum_{j\in[p]} w_j^\varepsilon|_\Omega \otimes w_j^\varepsilon|_\Omega \;\in\; \mathbb{S}_+(\mathcal{H}).$$

The reasoning above is formalized in the next theorem (the proof is in Sec. 8.G .3, page 408).

Theorem 8.10. *Let $d, p, s \in \mathbb{N}$. Let Ω satisfy Assumption 8.1(a) and $f(x) = \sum_{j\in[p]} w_j^2(x)$, $x \in \Omega$ with $w_j \in C^s(\mathbb{R}^d)$ for $j \in [p]$. Let k_r be the Sobolev kernel of smoothness $r > \max(s, \frac{d}{2})$ and let $\mathcal{H}$ be the associated RKHS. Then, for any $\varepsilon \in (0, 1]$ there exist $A_\varepsilon \in \mathbb{S}_+(\mathcal{H})$ such that*

$$\operatorname{Tr}(A_\varepsilon) \;\le\; C\varepsilon^{-2(r-s)}, \qquad \sup_{x\in\Omega} |f(x) - f_* - \langle \phi(x), A_\varepsilon \phi(x)\rangle| \;\le\; C'\varepsilon^s, \tag{8.33}$$

where $C = pqw^2, C' = pq'w^2$, and $w = \max_{j\in[p]} \|w_j\|_{\Omega,s}$ and q, q' are constants that depend only on s, r, d, Ω and are defined in the proof.

Denote now by $(\tilde{c}, \tilde{B})$ one minimizer of Eq. (8.32), and consider the problem in Eq. (8.5) with respect to $f_\varepsilon(x) = \langle \phi(x), A_\varepsilon \phi(x)\rangle + f_*$, i.e.,

$$\max_{c\in\mathbb{R},\; B\in\mathbb{S}_+(\mathbb{R}^n)} c - \lambda \operatorname{Tr}(B) \quad \text{such that} \quad \forall i \in [n],\; f_\varepsilon(x_i) - c = \Phi_i^\top B \Phi_i, \tag{8.34}$$

and denote by $p^\varepsilon_{\lambda,n}$ its optimum. Since $f_\varepsilon(x_i) - c = \Phi_i^\top B\Phi_i$ implies $|f(x_i) - c - \Phi_i^\top B\Phi_i| \le \tau$ when $\tau \ge \sup_{x\in\Omega} |f(x) - f_\varepsilon(x)|$, then in this case Eq. (8.32) is a relaxation of Eq. (8.34) and we have that $p^\varepsilon_{\lambda,n} - \tilde{c} - \lambda \operatorname{Tr}(\tilde{B}) \le 0$. Then, to obtain guarantees on $(\tilde{c}, \tilde{B})$ (the solution of Eq. (8.32)) we can apply Theorem 8.9 to the problem in Eq. (8.34) with $\theta_1, \theta_2, \tau_2 = 0$ and $\tau_1 = \tau$ with the requirement $\tau \ge \sup_{x\in\Omega} |f(x) - f_\varepsilon(x)|$. The reasoning above is formalized in the following theorem and the complete proof is reported in Sec. 8.G .4, page 409.

Theorem 8.11 (Global minimum for functions with low smoothness)**.** *Let $s \in \mathbb{N}$. Let k_r be a Sobolev kernel with smoothness $r \ge s, r > d/2$ and $\mathcal{H}$ be the associated RKHS. Let $\Omega \subset \mathbb{R}^d$ satisfying Assumption 8.1(a) and $f \in C^{s+2}(\mathbb{R}^d)$, satisfying Assumption 8.1(b). The problem in Eq. (8.32) admits a minimizer. Denote by $(\tilde{c}, \tilde{B})$ any of its minimizers for a given $\lambda > 0, \tau > 0$. With the same notation and the same conditions on λ of Theorem 8.5, when $\tau = \lambda^{s/(2r-s)}$*

$$|\tilde{c} - f_*| \;\le\; C_{1,f}(\lambda + \lambda^{\frac{s}{2r-s}}), \qquad \operatorname{Tr}(\tilde{B}) \;\le\; C_{2,f}(1 + \lambda^{-(1-\frac{s}{2r-s})}).$$

with $C_{1,f}, C_{2,f}$ defined in the proof and depending only on f and r, s, d, Ω.

The result above allows to derive the following estimate on algorithm 6 applied on the problem in Eq. (8.32) in the case of a function f with low smoothness. Consider the application algorithm 6 to the problem in Eq. (8.32) to a function $f \in C^{s+2}(\Omega)$ satisfying Assumption 8.1(b), with a Sobolev kernel k_r, $r \ge s, r > d/2$, and with $\tau = \lambda^{s/(2r-s)}$, $\lambda = O(n^{-\frac{r}{d}+1/2})$ on a set of n points sampled independently and uniformly at random from $\Omega = B_1(0)$, the unit ball of $\mathbb{R}^d$. By combining the result of Theorem 8.11 with the condition on λ in Theorem 8.5 and with the upper bound on the fill distance in the case of points sampled uniformly at random in Lemma 8.4, we have that

$$|\tilde{c} - f_*| \;=\; O\Big(n^{-\frac{s}{2d}(1-\frac{d-s}{2r-s})}\Big),$$

modulo logarithmic factors, where $\tilde{c}$ is the solution of Eq. (8.32). The rate above must be compared with the optimal rates for global minimization of functions in $C^{s+2}(\Omega)$ via function

evaluations, that is $n^{-\frac{s+2}{d}}$ for any $s \in \mathbb{N}$ (Prop. 1.3.9, pag. 34 by Novak (2006)). In the low smoothness setting, i.e., $s \leq d/2$ when we choose $r \gg d/2$, then the term $1 - \frac{d-s}{2r-s} \to 1$ and so the exponent of the rate above differs from the optimal one by a multiplicative factor $1/2 + \frac{1}{s}$, leading essentially to a rate of $O(n^{-s/(2d)})$. However, the choice of a large r will impact the hidden constants that are not tracked in the analysis above. Then for a fixed n there is a trade-off in r between the constants and the exponent of the rate. So in practice it would be useful to select r by parameter tuning.

8.8 .3 Certificate of optimality

While in Theorem 8.5 we provide a bound on the convergence of Eq. (8.5) *a priori*, i.e., only depending on properties of $f, \Omega, \mathcal{H}$, in this section we provide a bound *a posteriori*, that is a *certificate of optimality*. Indeed, the next theorem quantifies $f(z) - f^*$ for a candidate minimizer z, in terms of only $(\hat{c}, \hat{B})$, an (approximate) solution of Eq. (8.5) and $|f|_{\Omega,m}$. A candidate minimizer based on Eq. (8.5) is provided in Eq. (8.19). In section Sec. 8.7 we study a different algorithm Eq. (8.23) that explicitly provides a minimizer and whose certificate is studied in Sec. 8.G .5.

Theorem 8.12 (Certificate of optimality a minimizer from Eq. (8.5))**.** *Let Ω satisfy Assumption 8.1(a) for some $r > 0$. Let k be a kernel satisfying Assumptions 8.2(a) and 8.2(d) for some $m \in \mathbb{N}_+$. Let $\widehat{X} = \{x_1, \ldots, x_n\} \subset \Omega$ with $n \in \mathbb{N}_+$ such that $h_{\widehat{X},\Omega} \leq \frac{r}{18(m-1)^2}$. Let $f \in C^m(\Omega)$ and let $\hat{c} \in \mathbb{R}, \hat{B} \in \mathbb{S}_+(\mathbb{R}^n)$ and $\tau \geq 0$ satisfying*

$$|f(x_i) - \hat{c} \; - \; \Phi_i^\top \hat{B} \Phi_i| \leq \tau, \quad i \in [n], \tag{8.35}$$

where the Φ_i's are defined in Sec. 8.2 . Let $f_* = \min_{x \in \Omega} f(x)$. Then the following holds

$$|f(z) - f_*| \leq f(z) - \hat{c} + \varepsilon + 2\tau, \; \forall z \in \Omega, \quad \text{where} \quad \varepsilon = C h^m_{\widehat{X},\Omega}, \tag{8.36}$$

and $C = C_0(|f|_{\Omega,m} + \mathsf{M}\mathsf{D}_m \operatorname{Tr}(\hat{B}))$. The constants C_0, defined in Theorem 8.4, and $m, \mathsf{M}, \mathsf{D}_m$, defined in Assumptions 8.2(a) and 8.2(d), do not depend on $n, \widehat{X}$, $h_{\widehat{X},\Omega}, \hat{c}, \hat{B}$ or f.

Proof. By applying Theorem 8.4 with $g(x) = f(x) - \hat{c}$, we have $f(x) - \hat{c} \geq -\varepsilon - 2\tau$ for any $x \in \Omega$. In particular this implies that $f(\zeta) - \hat{c} \geq -\varepsilon - \tau$. The proof is concluded by noting that $f(z) \geq f_*$ by definition of f_*. □

8.9 Relationship with polynomial hierarchies

The formulation as an infinite-dimensional sum-of-squares bears some strong similarities with polynomial hierarchies. There are several such hierarchies allowing to solve any polynomial optimization problem (Lasserre, 2001, 2007, 2011), but one has a clear relationship to ours. The goal of the following discussion is to shed light on the benefits in terms of condition number and dimensionality of the problem, deriving by using an infinite dimensional feature map in the finite dimensional problem, instead of an explicit finite-dimensional polynomial map as in the case considered by the papers cited above.

Adding small perturbations. We start this discussion from the following result from Lasserre (Lasserre, 2007), that is, for any multivariate non-negative polynomial f on $\mathbb{R}^d$, and for

any $\eta > 0$, there exists a degree $r(f, \eta)$ such that the function

$$f_\eta(x) = f(x) + \eta \sum_{k=0}^{r(f,\varepsilon)} \frac{1}{k!} \sum_{j=1}^{d} x_j^{2k}$$

is a sum of squares, and such that the ℓ_1-norm between the coefficients of f and f_η tends to zero (here this ℓ_1-norm is equal to $\eta d \sum_{k=0}^{r(f,\varepsilon)} \frac{1}{k!} \leqslant \eta d e$).

This implies that for the kernel $k_r(x, y) = \sum_{k=0}^{r} \frac{(x^\top y)^k}{k!}$, with feature map $\phi_r(x)$ composed of all weighted monomials of degree less than r, the function

$$f(x) + \eta \|\phi_r(x)\|_2^2 = f(x) + \eta k_r(x, x)$$

is a sum of squares, for any $r \geqslant r(f, \eta)$, with η arbitrarily close to zero (this can be obtained by adding the required squares to go from $\sum_{j=1}^{d} x_j^{2k}$ to $\|x\|^{2k} = (\sum_{j=1}^{d} x_j^2)^k$). This result implies that minimizing f arbitrarily precisely over any compact set K (such that $\sup_{x \in K} k_r(x, x)$ is finite), can be done by minimizing $f(x) + \eta k(x, x)$, with sum-of-squares polynomials of sufficiently large degree. We already showed that in this paper that if f satisfies the geometric condition in Assumption 8.1(b), our framework is able to find the global minimum by the finite dimensional problem in Eq. (8.5), which, in turn, is based on a kernel associated to an infinite dimensional space (as the Sobolev kernel, see Example 8.1). We now show how our framework can provide approximation guarantees and potentially efficient algorithms for the problem above even when Assumption 8.1(b) may not hold and we use a polynomial kernel of degree r (with r that may not be large enough). However, in this case the resulting problem would suffer of a possibly infinite condition number and a larger dimensionality than the one achievable with an infinite dimensional feature map.

Modified optimization problem. Given the representation of $x \mapsto f(x) - f_* + \eta \|\phi_r(x)\|_2^2$ as a sum-of-squares, we can explicitly model the function as

$$f(x) - c + \eta \|\phi_r(x)\|_2^2 = \langle \phi_r(x), A \phi_r(x) \rangle$$

with A positive definite and $\eta \geq 0$. Note that if r is greater than twice the degree of f this problem is always feasible by taking η sufficiently large. Moreover, for feasible (c, η, A), we have for any $x \in \Omega$,

$$f(x) \geq c - \eta \|\phi_r(x)\|^2 \geq c - \eta \sup_{y \in \Omega} \|\phi_r(y)\|_2^2.$$

Thus, a relaxation of the optimization problem is

$$\sup_{c \in \mathbb{R}, A \succcurlyeq 0, \eta \geqslant 0} \quad c - \eta \sup_{y \in \Omega} \|\phi_r(y)\|_2^2 \quad \text{s. t.} \quad \forall x \in \Omega, \ f(x) = c + \phi_r(x)^\top A \phi_r(x) - \eta \|\phi(x)\|_2^2.$$

Moreover, if we choose r larger than $r(f - f_*, \eta)$, we know that there exists a feasible A which is positive semi-definite, with $c = f_* - \eta \sup_{y \in \Omega} \|\phi_r(y)\|_2^2$, and thus the objective value is greater than $f_* - \eta \sup_{y \in \Omega} \|\phi_r(y)\|_2^2$. Thus, the objective value of the problem above converges to f_*, when η go to zero (and thus $r(f - f_*, \eta)$ goes to infinity), while always providing a lower bound. Note that if $f - f_*$ is a sum of squares, then the optimal value η can be taken to be zero, and we recover the initial problem.

Subsampling and regularization. At this point, since r is finite, subsampling $\binom{d}{2r}$ points leads to an equivalent finite-dimensional problem. We can also add some regularization to sub-sample the problem and avoiding such a large number of points. Note here that the kernel matrix will probably be ill-conditioned, and the problem computationally harder to solve and difficult ro regularize.

Infinite-degree polynomials. In the approach outlined above, we need to let r increase to converge to the optimal value. We can directly take $r = \infty$, since $k_r(x, y) = \sum_{k=0}^{r} \frac{(x^\top y)^k}{k!}$ tends to the kernel $\exp(x^\top y)$, and here use subsampling. Again, it may lead to numerical difficuties. However, we can use Sobolev kernels (with guarantees on performance and controlled conditioning of kernel matrices), on the function $f(x) + \eta e^{\|x\|_2^2}$ for which we now there exists a sum of squares representation as soon as f is a polynomial.

8.10 Experiments

In this section, we illustrate our results with experiments on synthetic data.

Finding hyperparameters. Given a function to minimize and a chosen kernel, there are three types of hyperparameters: (a) the number n of sample points, (b) the regularization parameter λ, and (c) the kernel parameters. Since n drives the running time complexity of the method, we will always set it manually, while we will estimate the other parameters (regularization and kernel), by "cross-validation" (i.e., selecting the parameters of the algorithm that lead to the minimum value of f at the candidate optimum, among a logarithmic range of parameters). This adds a few function evaluations, but allows to choose good parameters.

Functions to minimize. We consider first a simple functions defined in $\mathbb{R}^2$ with their global minimimizer on $[-1, 1]^d$, which is minus the sum of Gaussian bumps (see Fig. 8.1). To go to higher even dimensions with the possibility of computing the global minimum with high precision by grid search, we consider functions of the form $f(x) = f(x_1, x_2) + f(x_3, x_4) + \cdots + f(x_{d-1}, x_d)$. We also consider adding a high-frequency cosine on the coordinate directions representing a more general scenario for a non-convex function. Note that in this second setting the gradient based methods cannot work properly (while ours can) as we are going to see in the simulations.

All results are reported by normalizing function values so that the range of values is 1, that is, $\max_{x\in[-1,1]^d} f(x) = 1$ and $\min_{x\in[-1,1]^d} f(x) = 0$.

Baseline algorithms. We compare our algorithm with the exponential kernel and points sampled from a quasi-random sequence in $[-1, 1]^d$, such as the Halton sequence (Niederreiter, 1992), to:

- Random search: select a quasi-random sequence in $[-1, 1]^d$ and take the point with minimal function value.
- Random search with gradient descent: starting gradient descent for a certain number of iterations from quasi-random points, with a number of initialization divided by $d + 1$ and the number of gradient steps, to account for gradient evaluations based on $d + 1$ function evaluations (by finite-difference). The step-size for gradient descent is taken constant, but its values is optimized for smallest final value while providing a descent algorithm.

Illustration in two dimensions. We show in Fig. 8.1 a function in two dimensions, with sampled point in purple, the trajectory of the candidate optimum along Newton iterations in red, and the final model of the function. We also compare to gradient descent with random starting points. We consider two functions below, one without extra high-frequency component (top), and one with (bottom). We can make the following observations:

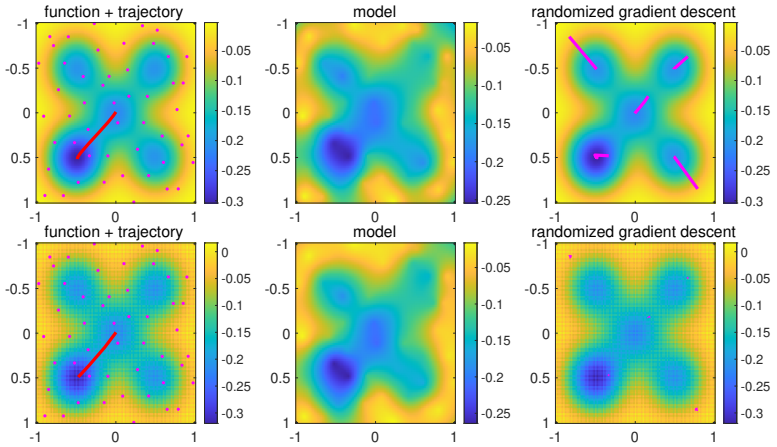

Figure 8.1: Top: 2D function without small-amplitude high-frequency components. Bottom: 2D function with small-amplitude high-frequency components. Left: sampled points and the trajectory of the proposed algorithm. Center: model reconstructed by the algorithm (see Eq. (8.37)). Right: the trajectory of gradient descent starting from random points. As it is possible to see, even a small local non-convexity prevents the random+GD algorithms to converge properly, while the proposed method is quite robust to it.

- Our algorithm outperforms random search, that is, it improves on the function values of the sampled points.
- For the smoother function, gradient descent performs quite well, but is not robust when high-frequency components are added.

Note that the proposed algorithm provides also a model of the function reconstructed starting from its evaluation on the sampled points. In particular, if $(\hat{c}, \hat{B})$ is a solution of the algorithm, the approximate function $\hat{g} \approx f - f^*$ corresponds to

$$\hat{g}(x) = \langle \phi(x), V^* B V \phi(x) \rangle = v(x)^\top R^{-1} \hat{B} R^{-\top} v(x), \quad \forall x \in \Omega \tag{8.37}$$

with $v(x) = (k(x_1, x), \ldots, k(x_n, x))$ for $x \in \Omega$ and where $V : \mathcal{H} \to \mathbb{R}^n$ is in Sec. 8.5 .

Higher dimensions. We compare the algorithms on a problem in dimension $d = 8$, as n increases, in order to assess how we approach the global optimum. We perform 4 replications with different random seeds for the sampling of points in $[-1, 1]^d$. The function to be minimized is built as described at the beginning of this section and is shifted and rescaled to have output in $[0, 1]$ and the minimum in 0. We can see that as n gets large, the performance of the proposed algorithm improves, and that with high frequency components, gradient descent with random restarts has worse performance and seem to show a slower rate overall, even in the case of the function without high-frequency components.

8.10 .1 Experiments on benchmarks for global optimization

In this section, we perform experiments using the algorithm described in the section above on the more than 200 global optimization problems in multiple dimensions constituting the well-known benchmark "Global Optimization Benchmarks" (Ali, Khompatraporn, and Zabinsky, 2005; Jamil and Yang, 2013; M. and C., 2005) `http://infinity77.net/global_optimization/index.html`.

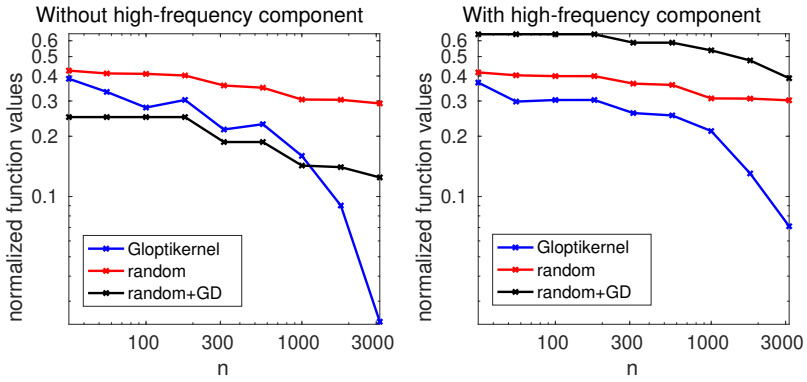

Figure 8.2: Multivariate case $d = 8$. Minimization error of our algorithm (gloptikernel) compared with random evaluations or random evaluations + GD. The function considered is built as described at the beginning of this section with domain $[-1, 1]^d$ and shifted and rescaled to have minimum in 0 and output in $[0, 1]$. Left: function without small-amplitude high-frequency components. Right: function with small-amplitude high-frequency components.

The functions to optimize come with their minimizer and their minimum to be used as a ground truth and with a region of interest where to look for the minimizers.

In this section, we present only the results for dimensions 4 and higher, as our method seems particularly interesting for these dimensions. The results for dimensions 2 and 3 can be found in Table 8.4, in the appendix, page 418. In Table 8.1, we report the results obtained by our algorithm. The algorithm we implemented is warm-restart scheme described in Sec. 8.7 .2. The implementation details are reported in Sec. 8.H . The algorithm was performed with $N = 200$ restart iterations, and was repeated 5 times (we select the best estimator out of the 5 restarts, to take into account of the high-probability factors). In Table 8.1, we report the following : (a) the problem name; (b) its dimension; (c) the number of iterations needed to achieve a threshold of 0.01 relative error; we define the relative error as $r(x) = \frac{f(x)-f(x_*)}{f(x_1)-f(x_*)}$; (d) the final absolute error $f(\widehat{x}) - f(x_*)$; (e) the number m of new function evaluations at each step (without counting those in order to select λ).

Note that the dimension of the optimization problem is $n = 3m$ and that the SDP constraint is also of size $n \times n$. Moreover, the choice we make to evaluate the relative error is in order to avoid very high values of the function f; comparing to $f(x_1)$ somewhat shows the importance of the iterative scheme.

Discussion, interpretation The set of functions on which we have tested our algorithm originally is a challenge allowing a maximum of 2000 function evaluations to reach a precision in absolute error of order 10^{-6}. We do not try to compete in this specific challenge, which models the fact that the number of function evaluations in certain real-life problems is very costly. In order to tackle this challenge, we would need to reduce the cost in function evaluations of certain steps such as that of the selection of λ (which we believe can be done without much difficulty).

Note that the fact that we achieve a relative error of 0.01 in almost all cases shows that the iterative scheme is indeed effective.

The performance on certain problems is bad, but this seems to be linked to the fact that the functions at hand have very high oscillations (hence high derivatives).

	d	iters thresh	final absolute error	fevs/iter
Colville	4	32	1.87E-03	31
Corana	4	1	0.00E+00	31
Shekel07	4	20	9.54E-07	31
PowerSum	4	2	3.26E-04	31
Ratkowsky01	4	90	3.69E+02	31
MieleCantrell	4	3	9.03E-13	31
Powell	4	6	2.85E-07	31
Shekel10	4	18	0.00E+00	31
Shekel05	4	18	1.91E-06	31
BiggsExp04	4	12	7.88E-05	31
Gear	4	2	1.18E-09	31
Kowalik	4	15	4.87E-05	31
DeVilliersGlasser01	4	4	1.06E+03	31
DeVilliersGlasser02	5	NaN	2.28E+03	36
Dolan	5	2	3.78E-13	36
BiggsExp05	5	3	2.64E-03	36
Trid	6	10	0.00E+00	41
Watson	6	11	1.09E-03	41
Hartmann6	6	8	0.00E+00	41
LennardJones	6	2	0.00E+00	41
Thurber	7	125	9.70E+03	46
Xor	9	NaN	6.99E-03	56
Paviani	10	23	1.03E-04	61
Cola	17	68	3.35E-01	96

Table 8.1: Results of our algorithm for functions in dimension greater than four

Poly1	$4x_1^2 + x_1x_2 - 4x_2^2 - 2.1x_1^4 + 4x_2^4 + x_1^6/3$
Poly2	$x_1^2x_4x_6x_7 + 4x_1x_2^2x_6x_8 + x_1x_2x_3x_4x_6 - x_2^4x_7 + 3x_2x_4^3x_7 + 3x_3x_4x_5x_6x_8 + x_3x_5x_7^2x_8 + \sum_{i=1}^{8} x_i^6$
Poly3	$-9x_2^2 + 8x_3x_7 + 2x_1x_4x_5 + 3x_3x_5x_6 + x_1^4 + x_2^4 + x_3^4 + x_4^4 + x_5^4 + x_6^4 + x_7^4 + x_1^6 + x_2x_3^5$
Poly4	$-15x_6 - 2x_1x_7^2 - 3x_2^2x_4 - x_3^2x_4 + x_1^4 + x_2^4 + x_3^4 + x_4^4 + x_5^4 + x_6^4 + x_7^4$
Poly5	$2x_5x_8 + 4x_1x_8x_9 + 4x_4x_6x_9 + x_1^4 + x_2^4 + x_3^4 + x_4^4 + x_5^4 + x_6^4 + x_7^4 + x_8^4 + x_9^4 + x_{10}^4$
Poly6	$-9x_2x_7x_{10} - 2x_3x_{11}x_{13} + 5x_5x_7x_{15} - 3x_9x_{11}x_{15} + \sum_{i=1}^{15} x_i^4$
Poly7	$8x_2x_8x_{11} + 3x_2x_{14}x_{15} - 5x_4x_7x_{13} - 13x_{12}^2x_{17} + \sum_{i=1}^{17} x_i^4$
Poly8	$-11x_2x_6x_{11} - 4x_3x_4x_{11} + 3x_4x_{10}x_{11} - x_5x_8x_{10} + \sum_{i=1}^{12} x_i^4$
Poly9	$12x_2x_4x_5x_8 + 5x_1x_2x_4x_5x_7 + 5x_2x_3x_4^2x_7 + x_1^6 + x_2^6 + x_3^6 + x_4^6 + x_5^6 + x_6^6 + x_7^6 + x_8^6 + x_9^6$

Table 8.2: Polynomials used in the experiments

Remark 28 (NaN values)**.** *NaN values simply mean that we never reach a relative precision of* 0.01*.*

8.10 .2 Comparison with SOS polynomials

In this section, we present a second set of experiments with the same setting as before but optimizing polynomial functions.

One of the reference algorithms in order to optimize polynomials (on semi-algebraic sets) is the Lasserre Hierarchy, implemented in the toolbox `gloptipoly 3` (Henrion, Lasserre, and Löfberg, 2007). Applying this toolbox on a minimization problem constrained on a hyper-rectangle will yield either a lower bound (if the hierarchy does not converge) or the exact minimum as well as a minimizer if the hierarchy does converge.

The idea of this section is not to compete with the Lasserre hierarchies, which are taylored for polynomials. Rather, we wish to compare both methods, and show that they can complement each other in the case where gloptipoly does not converge, by providing an approximation of the minimizer with a certificate (i.e. with an upper bound on its distance to the optimum). This is particularly interesting in high dimensions, or with polynomials with high degree : in that case, the size of the polynomial problem becomes untractable (for our computer at least), while our algorithm still runs and returns a solution.

In this experiment, we consider polynomials whose expression can be found below, and wish to find a minimizer for these polynomials in the hypercube $[-2, 2]^d$. Note that this domain is chosen such that we can easily sample from it. In particular, Lasserre hierarchies can adapt to much more flexible situations in terms of constraints.

Selection of polynomials Almost all of the polynomials considered in these experiments are of the form

$$P(x) = \sum_{i=1}^{d} x_i^{2k} + Q(x), \deg(Q) \leq 2k - 1.$$

We randomly select a few non zero indices for the Q as well as a random integer. The exact expressions of all the polynomials used can be found below in Table 8.2.

Results and Discussion. We report the results of these experiments in Table 8.3. The following columns are reported: (a) the name of the polynomial function; (b) the dimension d of

	d	deg	cv	relax	PSD dim glopti	gap	PSD dim (ours)
Poly1	2	6	True	3	10x10+4x(6x6)	0.0000E+00	63x63
Poly3	7	6	True	4	330x330+14x(120x120)	5.0819E-02	138x138
Poly4	7	4	False	4	330x330+14x(120x120)	1.9073E-05	138x138
Poly2	8	6	True	3	165x165+16x(45x45)	1.7166E-04	153x153
Poly9	9	6	False	3	455x455+20x(91x91)	5.1392E-02	168x168
Poly5	10	4	False	3	286x286+20x(66x66)	2.4796E-05	183x183
Poly8	12	4	False	2	91x91+24x(13x13)	3.6388E-02	213x213
Poly6	15	4	False	2	136x136+30x(16x16)	2.2190E-02	258x258
Poly7	17	4	False	2	171x171+34x(18x18)	1.6233E+00	288x288

Table 8.3: Results of the experiments when using both `gloptipoly3` and our method.

the underlying domain; (c) the degree deg of the polynomial function; (d) wether or not gloptipoly3 has converged (cv column); (e) the relaxation order we have tested for before computational issues (relax column); (f) the dimensions of the PSD constraints for the Lasserre hierarchy (PSD moment matrix + the ones due to the constraints); (g) the gap between our solution and the Lasserre lower bound; (h) the dimension of the PSD constraint in our mehod.

Our method is statistical and therefore does not enjoy the same precision that `gloptipoly3` achieves in the case when it converges. However, our method is clearly more scalable in the sense that it returns an approximate solution for polynomials of high degree and dimension, for any chosen dimension of the PSD matrix, and very small matrices allow to achieve already interesting precisionsm, as it is possible to see in Table 8.3.

8.11 Discussion

In this section, we discuss our results and propose a series of extensions.

Main technical contribution and extensions. We see that from Eq. (8.2), the problem of minimization can be easily written in terms of an infinite set of inequality constraints on $u(x) = f(x) - c$ that must hold for every $x \in \Omega$. While it is well known how to approximate efficiently an infinite set of equality constraints via a finite subset (e.g. via *bounds on functions with scattered zeros* (Wendland, 2004) from the field of approximation theory), leading to optimal rates for the approximation problem, the situation is more difficult in the case of an infinite set of inequality constraints. The main technical contribution of this paper, on which the whole result of the paper is based, is Theorem 8.4, that allows to deal with an infinite set of inequality constraints as efficiently as in the equality case as discussed in Sec. 8.5 .1. In particular, we rewrite the infinite set of inequalities $g(x) \geq 0,\ \forall x \in \Omega$ in terms of a very sparse set of constraints of the form $g(x_i) = \Phi_i B \Phi_i$, for some points $x_1, \ldots, x_n \in \Omega$ and a matrix $B \in \mathbb{S}_+(\mathbb{R}^n)$, with n in the same order of the one required by the equality case. Assume for simplicity that Ω is contained in the unit ball and the points are uniformly distributed in Ω. From Theorem 8.4 we derive that if B exists,

$$g(x) \ \geq -C\ n^{-m/d}\ (|g|_{\Omega,m} + \mathrm{Tr}(B)),$$

modulo logarithmic factors, where m is the order of smoothness of g. This result is particularly useful for two reasons. First, it recovers the same dependence on m, the smoothness of g, and n the number of sample points, as in the case of equality constraints. This is particularly convenient

when $m \gg d$, e.g. with $m \geq d$ the rate becomes $O(n^{-1})$, that is independent from d in the exponent (the dependence of d is still present in the hidden constants and it is exponential in the worst case). Second, if used in an optimization problem, the matrix B can be found via a convex formulation, by requiring $u(x_i) = \Phi_i^\top B \Phi_i$ for $i \in [n]$ and penalizing $\operatorname{Tr}(B)$ in the functional. This technique allows, for example, to deal with more general optimization problems with infinite constraints than the one considered in this paper, as

$$\min_{\theta \in \Theta} F(\theta) \;\text{ such that }\; g(\theta, x) \geq 0, \;\forall x \in \Omega,$$

by translating it as follows

$$\min_{\theta \in \Theta, B \succeq 0} F(\theta) + \lambda \operatorname{Tr}(B) \;\text{ such that }\; g(\theta, x_i) = \Phi_i B \Phi_i \;\forall i \in [n].$$

If F and u are convex in θ and Θ a convex set, then the second is a convex problem that has the potential to approximate very efficiently the first, due to Theorem 8.4. From this viewpoint this paper is an application of this principle to Eq. (8.2).

Duality. Beyond using duality in Sec. 8.6 for algorithmic purposes, there is also a dual for the infinite-dimensional problem, which can be written as,

$$\inf_{p:\Omega \to \mathbb{R}} \int_\Omega p(x) f(x) dx \text{ such that } \int_\Omega p(x) dx = 1 \text{ and } \int_\Omega p(x) \phi(x) \otimes \phi(x) dx \succcurlyeq 0.$$

Replacing the constraint $\int_\Omega p(x)\phi(x)\otimes\phi(x)dx \succcurlyeq 0$ by $\forall x \in \Omega,\ p(x) \geqslant 0$ leads to the usual relaxation of optimization with probability measures. Thus, like for polynomial optimization (Henrion, Korda, and Lasserre, 2020), our formulation corresponds also to a relaxation in the dual formulation to signed measures.

Comparison with algorithms based on SOS polynomials. According to recent results on SOS polynomials (see the work by Slot and Laurent (2020a) and references therein) which apply to polynomial relaxations as described in Sec. 8.9 , when f is a polynomial, such algorithms can achieve the global mininum with a rate $O(1/r^2)$ via an SDP problem based on the representation of SOS polynomials of degree r in terms of positive definite matrices. Since the dimension of the corresponding matrix is $n = \binom{d+r}{r}$ corresponding to $n = O(r^d)$, by expressing the rate with respect to the dimensionality of the matrix, such methods achieve the global minimum with an error that is in the order of $O(n^{-2/d})$. This can be compared with the approach proposed in this paper as algorithm 6. By sampling n points from the domain of interest, we cast an SDP problem in terms of a n-dimensional positive definite matrix, achieving a rate that is $C_{s,d} n^{-s/d+1/2}$ (see Theorem 8.6) modulo logaritmic factors, by using a Sobolev kernel k_{s+3} with $s > d/2$ (see Example 8.1). Since the polynomials are arbitrarily differentiable, we can choose s arbitrarily large at the cost of a larger constant $C_{s,d}$ completely characterized in Theorem 8.6. For example, by choosing $s = 5d/2$ we achieve the global minimum with a rate $O(n^{-2})$ that does not suffer of the curse of dimensionality except in the constants, and that is faster than the one obtained by SOS polynomial methods especially when $d \gg 1$. It must be noted that our result holds under the sufficient assumption Assumption 8.1(b) that can be relaxed according to Remark 26, but that it is not required by SOS polynomial methods. It would be of interest to know if such methods can achieve our rates under the same assumption.

Comparison with simpler algorithms. Similar reasoning can be done with respect to simple algorithms for global optimization. We consider here the algorithm that consists in sampling n points at random in Ω and taking the one with minimum value. A simple analysis, that we report below shows that this method achieves a rate of $O(n^{-2/d})$. So our method is strictly better than taking the minimum $f(x_i)$ for $i \in [n]$ when f is at least 3-times differentiable (see Sec. 8.8 .2). Note that even in the case when the function f is infinitely differentiable, the the algorithm that consists in sampling n points at random in Ω and taking the one with minimum value cannot go faster that $O(n^{-2/d})$. To see this, consider $\Omega = [0,1]^d$ and the points x_i to be chosen as a grid of step τ. This means that $n = O(\tau^{-d})$. Now let $f(x) = \|x - y\|^2$ for some $y \in [0,1]^d$. This function is infintely differentiable. Nevertheless, in general the best approximation of y on the grid will be $\tilde{y} = \tau \lfloor y/\tau + 1/2 \rfloor$ (componentwise). Since, for any τ, there exists always an $y \in [0,1]^d$ such that $\lfloor y/\tau + 1/2 \rfloor - y/\tau = 1/2$, we have that in the worst case

$$f(\tilde{y}) - f(y) = \|\tilde{y} - y\|^2 = \tau^2 \|\lfloor y/\tau + 1/2 \rfloor - y/\tau\|^2 = \tau^2/4.$$

Now if we express τ w.r.t., n, i.e., $\tau = n^{-1/d}$, we see that we obtain an error that is in the order of $n^{-2/d}$. So this simple algorithm cannot be better than $n^{-2/d}$ even if the function is infinitely differentiable. A similar argument can be obtained when the points are a generic covering of Ω.

Obtaining optimal rates. Our current analysis, even for functions f in Sobolev spaces, does not lead to the optimal rate of convergence (we obtain an extra term of $2/d$ in the exponents). We conjecture, that this could be removed by a more refined analysis (in particular in the construction of the operator A_*).

Modelling gradients. Our current framework only used function values. If gradients are observed, it could be possible to use them to reduce the number of sampled points, using tools by Zhou (2008).

The choice of Ω. Since we assume that f has at least one global minimum, then there exists always an open set Ω that contains it and that satisfies the required properties. In this work, we don't discuss how to find Ω. While, in general, this could be not an easy problem. In practice, many non-convex optimization problems come already with a region of interest where to look for the global minimum. Such a region is typically obtained by considering some basic properties of the function of interest. For example, if are minimizing a polynomial of the form $f(x) = B(x) + p(x)$, where $B(x) = x_1^{2r} + \cdots + x_d^{2r}$ for some $r \in \mathbb{N}$ and $p(x)$ is a polynomial of degree $q \leq 2r - 1$. Note that by construction f admits a global minimum, since it goes to $+\infty$ at infinity and has $p(0) < \infty$ (while any polynomial without this structure does not have a minimizer). Now it is possible to easily derive a hypercube that contains the global minimum. Indeed by construction $f(x_*) \leq f(0) = p(0)$. Denote by L the sum of the absolute values of the coefficients of p. Now take the smallest $R \geq 1$ such that $R^m - LR^q \geq p(0) + \varepsilon$ for an $\varepsilon > 0$. For any $x \notin (-R, R)$, we have

$$f(x) = B(x) + p(x) \geq R^m - LR^q \ > \ p(0) \geq f(x_*).$$

Then the region $(-R, R)$ contains all the global minimizers.

Efficient kernel approximations. The current algorithm has a complexity of $O(n^3)$ for n sampled points, partly due to the need to compute inverse of kernel matrices. There is a large literature within machine learning aiming at providing low-rank approximations, either from

approximations of K from a subset of its columns (see, e.g., the work by Bach (2013); Rudi, Camoriano, and Rosasco (2015) and references therein) or using random feature vectors (see, e.g., the work by Rudi and Rosasco (2017); Bach (2017b) and references therein). This requires to relax the equality constraint on the subset $\widehat{X}$ to an mean square deviations, as allowed by Sec. 8.8 .

Constrained optimization. Following the work by Lasserre (2001), we can apply the same algorithmic technique to constrained optimization, by formulating the problem of minimizing $f(x)$ such that $g(x) \geqslant 0$ as maximizing c such that $f(x) = c + p(x) + g(x)q(x)$, and p, q non-negative functions. We can then replace the non-negative constraints by $p(x) = \langle \phi(x), A\phi(x) \rangle$ and $q(x) = \langle \phi(x), B\phi(x) \rangle$ for positive operators A and B. We can then subsample and penalize the traces of A and B to obtain an algorithm. A detailed study of the approximation properties of this algorithm remains to be done.

Acknowledgements. We would like to thank Jean-Bernard Lasserre and Edouard Pauwels for their feedback on an earlier version of the manuscript. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

8.A Additional notations and definitions

We provide here some basic notation that will be used in the rest of the appendices.

Multi-index notation. Let $\alpha \in \mathbb{N}^d$, $x \in \mathbb{R}^d$ and f be an infinitely differentiable function on $\mathbb{R}^d$, we introduce the following notation

$$|\alpha| = \sum_{j\in[d]} \alpha_i, \quad \alpha! = \prod_{j\in[d]} \alpha_j!, \quad x^\alpha = \prod_{j\in[d]} x_j^{\alpha_j}, \quad \partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Some useful space of functions. Let Ω be an open set. In this paper we will denote by $C^s(\Omega)$, $s \in \mathbb{N}$, the set of s-times differentiable functions on Ω and by $C_0^s(\Omega)$ the set of functions that are differentiable at least s times and that are supported on a compact in Ω. Denote by $L^p(\Omega)$ the *Lebesgue space* of p-integrable functions with respect to the Lebesgue measure and denote by $\|\cdot\|_{L^p(\Omega)}$ the associated norm (Adams and Fournier, 2003).

8.A .1 Fourier Transform.

Given two functions $f, g : \Omega \to \mathbb{R}$ on some set Ω, we denote by $f \cdot g$ the function corresponding to *pointwise product* of f, g, i.e.,

$$(f \cdot g)(x) = f(x)g(x), \quad \forall x \in \Omega.$$

Let $f, g \in L^1(\mathbb{R}^d)$ we denote the *convolution* by $f \star g$

$$(f \star g)(x) = \int_{\mathbb{R}^d} f(y)g(x-y)dy.$$

Let $f \in L^1(\mathbb{R}^d)$. The Fourier transform of f is denoted by $\tilde{f}$ and is defined as

$$\tilde{f}(\omega) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-i\,\omega^\top x}\, f(x)\, dx,$$

We now recall some basic properties, that will be used in the rest of the appendix.

Proposition 8.2 (Basic properties of the Fourier transform (Wendland, 2004), Chapter 5.2.)**.**

(a) *Let $f \in L^1(\mathbb{R}^d)$ and let $r > 0$. Denote by $\tilde{f}$ its Fourier transform and by f_r the function $f_r(x) = f(x/r)$ for all $x \in \mathbb{R}^d$, then*

$$\tilde{f}_r(\omega) = r^d \tilde{f}(r\omega).$$

(b) *Let $f, g \in L^1(\mathbb{R}^d)$, then*

$$\widetilde{f \cdot g} = (2\pi)^{d/2} \tilde{f} \star \tilde{g}.$$

(c) *Let $\alpha \in \mathbb{N}_0^d$, $f : \mathbb{R}^d \to \mathbb{R}$ and $f, \partial^\alpha f \in L^1(\mathbb{R}^d)$, then*

$$\widetilde{\partial^\alpha f}\,(\omega) = i^{|\alpha|}\omega^\alpha \tilde{f}(\omega), \quad \forall \omega \in \mathbb{R}^d.$$

(d) *Let $f \in L^1(\mathbb{R}^d)$, then*

$$\|\tilde{f}\|_{L^\infty(\mathbb{R}^d)} \leq (2\pi)^{-d/2} \|f\|_{L^1(\mathbb{R}^d)}.$$

(e) *Let $f \in L^1(\mathbb{R}^d)$ and assume that $\tilde{f} \in L^1(\mathbb{R}^d)$, then*

$$f(x) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{i\omega^\top x}\, \tilde{f}(\omega)\, dx, \quad \text{and} \quad \|f\|_{L^\infty(\mathbb{R}^d)} \leq (2\pi)^{-d/2} \|\tilde{f}\|_{L^1(\mathbb{R}^d)}.$$

(f) *There exists a linear isometry $\mathcal{F} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ satisfying*

$$\mathcal{F}f = \tilde{f}, \quad f \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d).$$

The isometry is uniquely determined by the property in the equation above. For any $f \in L^2(\mathbb{R}^d)$ we denote by $\tilde{f}$ the function $\tilde{f} = \mathcal{F}f$.

8.A .2 Sobolev Spaces

For this section we refer to the work by Adams and Fournier (2003). For any $\alpha \in \mathbb{N}_0^d$ we say that $v_\alpha \in L^1_{loc}(\mathbb{R}^d)$ is the *α-weak derivative* of $u \in L^1_{loc}(\mathbb{R}^d)$ if, for all compactly supported smooth functions $\tau \in C_0^\infty(\mathbb{R}^d)$, we have

$$\int_{\mathbb{R}^d} v_\alpha(x)\tau(x)dx = (-1)^{|\alpha|} \int_{\mathbb{R}^d} u(x)(\partial^\alpha \tau)(x)dx,$$

and we denote v_α by $D^\alpha u$. Let $\Omega \subseteq \mathbb{R}^d$ be an open set. For $s \in \mathbb{N}, p \in [1, \infty]$ the Sobolev spaces $W_p^s(\Omega)$ are defined as

$$W_p^s(\Omega) = \{f \in L^p(\Omega) \mid \|f\|_{W_p^s(\Omega)} < \infty\}, \quad \|f\|_{W_p^s(\Omega)} = \sum_{|\alpha| \leq s} \|D^\alpha f\|_{L^p(\Omega)}.$$

We now recall some basic results about Sobolev spaces that are useful for the proofs in this paper. First we start by recalling the restriction properties of Sobolev spaces. Let $\Omega \subseteq \Omega' \subseteq \mathbb{R}^d$ be two open sets. Let $\beta \in \mathbb{N}$ and $p \in [1, \infty]$. By definition of the Sobolev norm above we have

$$\|g|_\Omega\|_{W_p^s(\Omega)} \leq \|g\|_{W_p^s(\Omega')},$$

and so $g|_\Omega \in W_p^s(\Omega)$ for any $g \in W_p^s(\Omega')$. Now we recall the extension properties of Sobolev spaces.

Proposition 8.3 (Extension operator, 5.24 in the work by Adams and Fournier (2003))**.** *Let Ω be a bounded open subset of $\mathbb{R}^d$ with locally Lipschitz boundary (Adams and Fournier, 2003). Let $\beta \in \mathbb{N}$ and $p \in [1, \infty]$. There exists a bounded operator $E : W_p^\beta(\Omega) \to W_p^\beta(\mathbb{R}^d)$ and a constants C_3 depending only on β, p, Ω such that for any $h \in W_p^\beta(\Omega)$ the following holds (a) $h = (Eh)|_\Omega$ (b) $\|Eh\|_{W_p^\beta(\mathbb{R}^d)} \leq C_3 \|h\|_{W_p^\beta(\Omega)}$ with $C_3 = \|E\|_{op}$.*

Proposition 8.4 (Approximation property of Sobolev spaces, Thm 5.33 in the work by Adams and Fournier (2003))**.** *Let Ω be a bounded open subset of $\mathbb{R}^d$ with locally Lipschitz boundary (Adams and Fournier, 2003), or $\Omega = \mathbb{R}^d$. Let $s, d \in \mathbb{N}, r \geq s$ and $p \in [1, \infty]$. There exists C_1 depending only on s, d, p and C_2 depending only on r, s, d, p such that for any $\varepsilon \in (0, 1]$ and $g \in W_p^s(\Omega)$ there exists $g_\varepsilon \in C^\infty(\Omega)$ satisfying (i) g_ϵ is the restriction to Ω of a certain $\tilde{g}_\epsilon \in C^\infty(\mathbb{R}^d)$ and (ii)*

$$\|g - g_\varepsilon\|_{L^p(\Omega)} \leq C_1 \varepsilon^s \|g\|_{W_p^s(\Omega)}, \qquad \|g_\varepsilon\|_{W_p^r(\Omega)} \leq C_2 \varepsilon^{-(r-s)} \|g\|_{W_p^s(\Omega)}.$$

Proof. The case $\Omega = \mathbb{R}^d$ is covered explicitly by Thm. 5.33 in the work by Adams and Fournier (2003). The result holds also for $W_p^s(\Omega)$ when Ω has Lipschitz boundaries as discussed in the work by Adams and Fournier (2003), above Theorem 5.33. The result is obtained considering that when Ω has Lipschitz boundaries, then there exists a bounded extension operator between $W_p^s(\Omega)$ and $W_p^s(\mathbb{R}^d)$ (Adams and Fournier, 2003). Here we provide the proof for the sake of completeness. Let $g \in W_p^s(\Omega)$ and let $\varepsilon \in (0, 1]$. Then, by proposition 8.3 since Ω has Lipschitz boundary, there exists a bounded extension operator $E : W_p^s(\Omega) \to W_p^s(\mathbb{R}^d)$. Denote by $\tilde{g}$ the function $\tilde{g} = Eg$ and note that $\tilde{g} \in W_p^s(\mathbb{R}^d)$. Then, by applying Thm. 5.33 in the work by Adams and Fournier (2003) we have that there exists $\tilde{g}_\varepsilon \in C^\infty(\mathbb{R}^d)$ such that

$$\|\tilde{g} - \tilde{g}_\varepsilon\|_{L^p(\mathbb{R}^d)} \leq C\varepsilon^s \|\tilde{g}\|_{W_p^s(\mathbb{R}^d)}, \qquad \|\tilde{g}_\varepsilon\|_{W_p^r(\mathbb{R}^d)} \leq C'\varepsilon^{-(r-s)} \|\tilde{g}\|_{W_p^s(\mathbb{R}^d)},$$

for some C depending only on s, p and C' depending on r, s, p. Since by proposition 8.3 we have $\|\tilde{g}\|_{W_p^s(\mathbb{R}^d)} = \|Eg\|_{W_p^s(\mathbb{R}^d)} \leq C_3 \|g\|_{W_p^s(\Omega)}$, so

$$\|g - \tilde{g}_\varepsilon|_\Omega\|_{L^p(\Omega)} \leq \|\tilde{g} - \tilde{g}_\varepsilon\|_{L^p(\mathbb{R}^d)} \leq C\varepsilon^s \|\tilde{g}\|_{W_p^s(\mathbb{R}^d)} \leq CC_3\varepsilon^s \|g\|_{W_p^s(\Omega)},$$

and analogously,

$$\|\tilde{g}_\varepsilon|_\Omega\|_{W_p^r(\Omega)} \leq \|\tilde{g}_\varepsilon\|_{W_p^r(\mathbb{R}^d)} \leq C'\varepsilon^{s-r} \|\tilde{g}\|_{W_p^s(\mathbb{R}^d)} \leq C'C_3\varepsilon^{s-r} \|g\|_{W_p^s(\Omega)}.$$

The proof is concluded by taking $g_\varepsilon = \tilde{g}_\varepsilon|_\Omega$ and $C_1 = CC_3, C_2 = C'C_4$. □

In the next proposition we recall some aspects of the more general *Sobolev embedding theorem* (Adams and Fournier, 2003).

Proposition 8.5. *Let Ω be a bounded open set with Lipschitz continouos boundary. Let $r \in \mathbb{N}$ and $1 \leq p \leq q \leq \infty$. Then $W_q^r(\Omega) \subseteq W_p^r(\Omega)$. In particular there exists a constant C_5 such that*

$$\|\cdot\|_{W_p^r(\Omega)} \leq C_5 \|\cdot\|_{W_q^r(\Omega)}.$$

Finally, note that for any $f \in C^r(\mathbb{R}^d)$, it holds $f|_\Omega \in W_\infty^r(\Omega)$.

Proof. The main statement of the proposition is a subcase of the more general Sobolev embedding theorem (Adams and Fournier, 2003).

Finally, we recall that, since $f \in C^r(\mathbb{R}^d)$ and Ω is bounded, then $\partial^\alpha f$ is uniformly bounded on Ω, for any $\alpha \in \mathbb{N}^d$ satisfying $|\alpha| \leq r$. This implies that $f|_\Omega \in W_\infty^r(\Omega)$. □

Finally, note that the semi-norm $\|\cdot\|_{\Omega,r}$ defined in Eq. (8.31) and the Sobolev norm $\|\cdot\|_{W_\infty^r}$ are equivalent in the following sense.

Proposition 8.6. *Let $\Omega \subset \Omega'$ be two bounded open sets. Let $r \in \mathbb{N}$. For any $u \in C^r(\Omega')$, recall the definition of $\|u\|_{\Omega,r}$ from Eq. (8.31). There exists an explicit constant $C_6 > 0$ such that*

$$\forall u \in C^r(\Omega'), \frac{1}{C_6}\|u|_\Omega\|_{W_\infty^r(\Omega)} \leq \|u\|_{\Omega,r} \leq C_6\|u|_\Omega\|_{W_\infty^r(\Omega)}.$$

Note that this inequality holds also when the norms are unbounded, by using the convention $+\infty \leq +\infty$.

Proof. Since by Eq. (8.31), $\|u\|_{\Omega,r} = \max_{|\alpha|\leq r} \|\partial^\alpha u\|_{L^\infty(\Omega)}$ and $\|u|_\Omega\|_{W_\infty^r(\Omega)} = \sum_{|\alpha|\leq r} \|\partial^\alpha u\|_{L^\infty(\Omega)}$, and $\{|\alpha| \leq r\}$ is of size $1 + d + ... + d^r = \frac{d^{r+1}-1}{d-1}$ (where this is taken to be equal to $k + 1$ in the case where $d = 0$), the result holds for $C_6 = \frac{d^{r+1}-1}{d-1}$.

□

8.A .3 Reproducing Kernel Hilbert spaces

For this section we refer to the work by Aronszajn (1950); Steinwart and Christmann (2008); Paulsen and Raghupathi (2016). Let S be a set and $k : S \times S \to \mathbb{R}$ be a p.d. kernel. We denote by $\mathcal{H}_k(S)$ the reproducing kernel Hilbert space (RKHS) associated to the kernel k, and by $\langle \cdot, \cdot \rangle_k$ the associated inner product. In particular, we will omit the dependence in k from $\mathcal{H}$ and $\langle \cdot, \cdot \rangle$ when the used kernel is clear from the context. We will omit also the dependence on S when $S = \Omega$, the region we are using in this paper. In particular we will use the following shortcuts $\mathcal{H} = \mathcal{H}_k(\Omega)$ and $\mathcal{H}(\mathbb{R}^d) = \mathcal{H}_k(\mathbb{R}^d)$.

Concrete constructions and useful characterizations. In the rest of the section we provide other methods to build RKHS and some interesting characterizations of $\mathcal{H}_k(S)$ and $\langle \cdot, \cdot \rangle_k$ that will be useful int the rest of the appendix.

Proposition 8.7 (Construction of RKHS given S, ϕ, Thm. 4.21 by Steinwart and Christmann (2008))**.** *Let $\phi : S \to V$ be a continuous map, where V is separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$. Let $k(x, x') = \langle \phi(x), \phi(x') \rangle_V$ for any $x, x' \in S$. Then k is a p.d. kernel and the associated RKHS is characterized as follows*

$$\mathcal{H}_k(S) = \{\langle w, \phi(\cdot) \rangle_V \mid w \in V\}, \quad \|f\|_{\mathcal{H}_k(S)} = \inf_{u \in V} \|u\|_V \text{ s.t. } f = \langle u, \phi(\cdot) \rangle_V .$$

Proposition 8.8 (Restriction of a RKHS $\mathcal{H}_{k_1}(S_1)$ on a subset $S_0 \subset S_1$ (Aronszajn, 1950; Paulsen and Raghupathi, 2016))**.** *Let k_0 be the restriction on S_0 of the kernel k_1 defined on S_1. Then the following holds*

(a) *k_0 is a p.d. kernel,*

(b) *the RKHS $\mathcal{H}_{k_0}(S_0)$ is characterized as $\mathcal{H}_{k_0}(S_0) = \{f|_{S_0} \mid f \in \mathcal{H}_{k_1}(S_1)\}$,*

(c) *the norm $\|\cdot\|_{\mathcal{H}_{k_0}(S_0)}$ is characterized by*

$$\|f\|_{\mathcal{H}_{k_0}(S_0)} = \inf_{g \in \mathcal{H}_{k_1}(S_1)} \|g\|_{\mathcal{H}_{k_1}(S_1)}, \text{ s.t. } f(x) = g(x) \ \forall x \in S_0,$$

(d) *there exist a linear bounded* extension *operator $E : \mathcal{H}_{k_0}(S_0) \to \mathcal{H}_{k_1}(S_1)$ such that $(Ef)(x) = f(x)$ for any $x \in S_0$ and $f \in \mathcal{H}_{k_0}(S_0)$ and such that*

$$\|f\|_{\mathcal{H}_{k_0}(S_0)} = \|Ef\|_{\mathcal{H}_{k_1}(S_1)}, \quad \forall f \in \mathcal{H}_{k_0}(S_0),$$

(e) *there exist a linear bounded* restriction *operator $R : \mathcal{H}_{k_1}(S_1) \to \mathcal{H}_{k_0}(S_0)$ such that $(Rf)(x) = f(x)$ for any $x \in S_0$ and $f \in \mathcal{H}_{k_1}(S_1)$,*

(f) *R and E are partial isometries. In particular $E = R^*$ and RE is the identity on $\mathcal{H}_{k_0}(S_0)$, while ER is a projection operator on $\mathcal{H}_{k_1}(S_1)$.*

Proposition 8.9 (Translation invariant kernels on $\mathbb{R}^d$)**.** *Let $v : \mathbb{R}^d \to \mathbb{R}$ such that its Fourier transform $\tilde{v}$ is integrable and satisfies $\tilde{v} \geq 0$ on $\mathbb{R}^d$. Then*

(a) *The function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as $k(x, x') = v(x - x')$ for any $x, x' \in \mathbb{R}^d$ is a kernel and is called* translation invariant kernel.

(b) *The RKHS $\mathcal{H}_k(\mathbb{R}^d)$ and the norm $\|\cdot\|_{\mathcal{H}_k(\mathbb{R}^d)}$ are characterized by*

$$\mathcal{H}_k(\mathbb{R}^d) = \{f \in L^2(\mathbb{R}^d) \mid \|f\|_{\mathcal{H}_k(\mathbb{R}^d)} < \infty\}, \quad \|f\|^2_{\mathcal{H}_k(\mathbb{R}^d)} = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \frac{|(\mathcal{F}f)(\omega)|^2}{\tilde{v}(\omega)} d\omega,$$

where $\mathcal{F}f$ is the Fourier transform of f (see proposition 8.2 for more details on $\mathcal{F}$).

(c) *The inner product $\langle \cdot, \cdot \rangle_k$ is characterized by*

$$\langle f, g \rangle_k = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \frac{(\mathcal{F}f)(\omega)\overline{(\mathcal{F}g)(\omega)}}{\tilde{v}(\omega)} d\omega.$$

8.A .4 Auxiliary results on C^∞ functions

Proposition 8.10. *Let U be an open set of $\mathbb{R}^d$ and $K \subset U$ be a compact set. Let $u \in C^\infty(U)$, then there exists $v \in C_0^\infty(\mathbb{R}^d)$ (with compact support), such that $v(x) = u(x)$ for all $x \in K$.*

Proof. By Thm. 1.4.1, pag. 25 by Hörmander (2015) there exists $z_{K,U} \in C_0^\infty(U)$, i.e., a smooth function with compact support, such that $z_{K,U}(x) \in [0,1]$ for any $x \in U$ and $z(x) = 1$ for any $x \in K$. Consider now the function $v_{K,U}$ defined as $v_{K,U}(x) = z_{K,U}(x)u(x)$ for all $x \in U$. The function $v_{K,U}$ is in $C_0^\infty(U)$, since it is the product of a $C_0^\infty(U)$ and a $C^\infty(U)$ function, moreover $v_{K,U}(x) = u(x)$ for all $x \in K$. The theorem is concluded by defining v as the extension of $v_{K,U}$ to $\mathbb{R}^d$, i.e., the function $v_K(x) = z_{K,U}(x)$ for any $x \in U$ and $v_K(x) = 0$ for any $x \in \mathbb{R}^d \setminus U$. This is always possible since $v_{K,U}$ is supported on a compact set K' which is contained in the open set U, so $v_{K,U}$ is already identically zero in the open set $U \setminus K'$. □

Lemma 8.6. *Given $\zeta \in \mathbb{R}^d$ and $r > 0$, there exists $u \in C_0^\infty(\mathbb{R}^d)$ such that for any $x \in \mathbb{R}^d$, it holds*

(i) $u(x) \in [0,1]$;

(ii) $\|x\| \geq r \implies u(x) = 0$;

(iii) $\|x\| \leq r/2 \implies u(x) = 1$.

Proof. Assume without loss of generality that $\zeta = 0$ and $r = 1$. Consider the following functions :

$$u_1(x) = \begin{cases} \exp\left(-\frac{1}{1-\|x\|^2}\right) & \text{if } \|x\| < 1 \\ 0 & \text{otherwise} \end{cases}, \qquad u_2(x) = \begin{cases} \exp\left(-\frac{1}{\|x\|^2 - 1/4}\right) & \text{if } \|x\| > 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

Both u_1 and u_2 belong to $C^\infty(\mathbb{R}^d)$ with values in $[0,1]$. Moreover, $u_1 > \alpha_1$ on $B_{3/4}(0)$ and $u_2 \geq \alpha_2$ for some $\alpha_1, \alpha_2 > 0$ on $\mathbb{R}^d \setminus B_{3/4}(0)$, which implies that $u_1 + u_2 \in I$ on $\mathbb{R}^d$, where $I = [\min(\alpha_1, \alpha_2), 2]$. Since $(\cdot)^{-1}$ is infinitely differentiable on $(0, \infty)$ we see that $1/(u_1 + u_2)$ is well defined on all $\mathbb{R}^d$ and belongs to $C^\infty(\mathbb{R}^d)$, since $I \subset\subset (0, \infty)$. Consider the function

$$u_0 = \frac{u_1}{u_1 + u_2}.$$

It is non-negative, bounded by 1, and infinitely differentiable as a product. Moreover :

$$\forall x \in B_{1/2}(0),\ u_2(x) = 0 \implies u_0(x) = 1, \qquad \forall x \in \mathbb{R}^d,\ u_1(x) = 0 \Leftrightarrow u_0(x) = 0 \Leftrightarrow x \in \mathbb{R}^d \setminus B_1(0).$$

To conclude the proof, given $r > 0$ and $\zeta \in \mathbb{R}^d$ we will take $u(x) = u_0((x - \zeta)/r)$. □

Lemma 8.7. *Let $N \in \mathbb{N}_+$, $\zeta_1, ..., \zeta_N \in \mathbb{R}^d$ and $r_1, ..., r_N > 0$. For $n \in \{1, \dots, N\}$, let $B_n = B_{r_n}(\zeta_n)$ be the open ball centered in ζ_n of radius r_n and $B'_n = B_{r_n/2}(\zeta_n) \subset B_n$ be the open ball centered in ζ_n of radius $r_n/2$. Then there exists functions $v_0, v_1, ..., v_N \in C^\infty(\mathbb{R}^d)$ such that*

(i) $v_0 = v_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n=1}^N B'_n}$

(ii) $v_n = v_n \cdot \mathbf{1}_{B_n},\ \forall n \in \{1, \dots, N\}$

(iii) $\sum_{n=0}^N v_n^2 = 1$.

Proof. For all $n \in [N]$, take u_n as in Lemma 8.6 with $r = r_n, \zeta = \zeta_n$ and define $u_0 = \prod_{n=1}^N (1 - u_n)$. Since $\forall n \in [N],\ u_n \in [0,1]$, we also have $u_0 \in [0,1]$. Moreover, let $R = \max_{n\in[N]} \|\zeta_n\| + r_n$, then

$$\forall \|x\| \geq R,\ \forall 1 \leq n \leq N,\ u_n(x) = 0 \text{ and } u_0(x) = 1.$$

Step 1. $u_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n\in[N]} B'_n} = u_0$ *and for all* $n \in [N]$, $u_n \cdot \mathbf{1}_{B_n} = u_n$.

By point (iii) of Lemma 8.6, $u_n = 1$ on B'_n for all $n \in [N]$, which shows that $u_0 = 0$ on $\bigcup_{n=1}^N B'_n$ and hence $u_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n\in[N]} B'_n} = u_0$. On the other hand, for all $n \in [N]$, point (ii) of Lemma 8.6 directly implies $u_n \cdot \mathbf{1}_{B_n} = u_n$.

Step 2. *The function* $\frac{1}{\sqrt{\sum_{n=0}^N u_i^2}}$ *is well defined and in* $C^\infty(\mathbb{R}^d)$.

By definition of u_0, if $u_0(x) = 0$, then there exists $n \in [N]$ such that $u_n(x) = 1$. Since all the u_n are non-negative, this shows that $s := \sum_{n=0}^N u_n^2 > 0$. Moreover, consider the closed ball $\bar{B}$ of radius R and centered in 0. Since $\bar{B}$ is compact, s is continuous and $s(x) > 0$ for any $x \in \bar{B}$, then there exists $0 < m_R \leq M_R < \infty$ such that $s(x) \in [m_R, M_R]$ for any $x \in \bar{B}$. Moreover, since for any $\|x\| \geq R,\ u_0(x) = 1$ and $\forall n \in [N],\ u_n(x) = 0$, we see that

$$\forall x \in \mathbb{R}^d \setminus B_R(0),\ \sum_{n=0}^N u_n^2(x) = 1.$$

Then $s \in [m, M]$ for any $x \in \mathbb{R}^d$, where $m = \min(m_R, 1)$ and $M = \max(M_R, 1)$.

Since the interval $I = [m, M]$ is a compact set included in the open set $(0, \infty)$ and $1/\sqrt{\cdot}$ is infinitely differentiable on $(0, \infty)$ then by proposition 8.10 there exists $q_I \in C_0^\infty(\mathbb{R})$ such that $q_I(x) = 1/\sqrt{x}$ for any $x \in I$. Since $s(x) \in I$ for any $x \in \mathbb{R}^d$ we have

$$\frac{1}{\sqrt{\sum_{n=0}^N u_i^2}} = q_I \circ s.$$

Finally $q_I \circ s \in C^\infty(\mathbb{R}^d)$ since it is the composition of $q_I \in C_0^\infty(\mathbb{R})$ and $s = \sum_{n=0}^N u_n^2 \in C^\infty(\mathbb{R}^d)$ (since all the u_n are in $C^\infty(\mathbb{R}^d)$) and $s \in [m, M]$.

Step 3.

Finally, defining $v_n = \frac{u_n}{\sqrt{\sum_{n=0}^N u_n^2}}$ for all $0 \leq n \leq N$, $v_n \in C^\infty(\mathbb{R}^d)$ since it is the product of two infinitely differentiable functions. Moreover, $\sum_{n=0}^N v_i^2 = 1$ by construction and $v_0 = v_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n=1}^N B'_n}$ since u_0 satisfies the same equality and v_0 is the product of u_0 by the strictly positive function $1/\sqrt{s}$. Analogously $v_n = v_n \cdot \mathbf{1}_{B_n},\ \forall n \in \{1, \dots, N\}$, since u_n satisfy the same equality and v_n is the product of u_n by the strictly positive function $1/\sqrt{s}$. □

8.B Fundamental results on scattered data approximation

We recall here some fundamental results about local polynomial approximation. In particular, we report here the proofs to track explicitly the constants. The proof techniques are essentially from the work by Narcowich, Ward, and Wendland (2003); Wendland (2004). Denote by $\pi_k(\mathbb{R}^d)$ the set of multivariate polynomials of degree at most k, with $k \in \mathbb{N}$. In this section $B_r(x) \subset \mathbb{R}^d$ denotes the open ball of radius r and centered in x.

Proposition 8.11 ((Wendland, 2004), Corollary 3.11. Local polynomial reproduction on a ball)**.** *Let $k \in \mathbb{N}$, $d, m \in \mathbb{N}_+$ and $\delta > 0$. Let B_δ be an open ball of radius $\delta > 0$ in $\mathbb{R}^d$. Let $\widehat{Y} = \{y_1, \dots, y_m\} \subset B_\delta$ be a non empty finite subset of B_δ. If either $k = 0$ or $h_{\widehat{Y}, B_\delta} \leq \frac{\delta}{9k^2}$, there exist $u_j : B_\delta \to \mathbb{R}$ with $j \in [m]$ such that*

(a) $\sum_{j\in[m]} p(y_j)u_j(x) = p(x), \quad \forall x \in B_\delta, p \in \pi_k(\mathbb{R}^d)$

(b) $\sum_{j\in[m]} |u_j(x)| \leq 2, \quad \forall x \in B_\delta.$

Lemma 8.8 (Bounds on functions with scattered zeros on a small ball (Narcowich, Ward, and Wendland, 2003; Wendland, 2004))**.** *Let $k \in \mathbb{N}$, $d, m \in \mathbb{N}_+$ and $\delta > 0$. Let $B_\delta \subset \mathbb{R}^d$ be a ball of radius δ in $\mathbb{R}^d$. Let $f \in C^{k+1}(B_\delta)$. Let $\widehat{Y} = \{y_1, \dots, y_m\} \subset B_\delta$ be a non empty finite subset of B_δ. If either $k = 0$ or $h_{\widehat{Y}, B_\delta} \leq \frac{\delta}{9k^2}$, it holds:*

$$\sup_{x\in B_\delta} |f(x)| \;\leq\; 3C\delta^{k+1} \;+\; 2\max_{i\in[m]} |f(y_i)|, \qquad C := \sum_{|\alpha|=k+1} \frac{1}{\alpha!} \|\partial^\alpha f\|_{L^\infty(B_\delta)}.$$

Proof. Note that since either $k = 0$ or $h_{\widehat{Y}, B_\delta} \leq \frac{\delta}{9k^2}$, then we can apply proposition 8.11 obtaining u_j with $j \in [m]$ with the local polynomial reproduction property. Define the function $s_{f,\widehat{Y}} = \sum_{j\in[m]} f(y_j)u_j$ and let $\tau = \max_{i\in[m]} |f(y_i)|$. Now, by using both propositions 8.11(a) and 8.11(b), we have that for any $p \in \pi_k(\mathbb{R}^d)$ and any $x \in B_\delta$,

$$\begin{aligned}
|f(x)| &\leq |f(x) - p(x)| + |p(x) - s_{f,\widehat{Y}}(x)| + |s_{f,\widehat{Y}}(x)| \\
&\leq |f(x) - p(x)| + \sum_{j\in[m]} |p(y_j) - f(y_j)||u_j(x)| + \max_{j\in[m]} |f(y_j)| \sum_{j\in[m]} |u_j(x)| \\
&\leq \|f - p\|_{L^\infty(B_\delta)}(1 + \sum_{j\in[m]} |u_j(x)|) + \tau \sum_{j\in[m]} |u_j(x)| \\
&\leq 3\|f - p\|_{L^\infty(B_\delta)} + 2\tau.
\end{aligned}$$

In particular, consider the Taylor expansion of f at the center x_0 of B_δ up to order k (e.g. the work by Brenner and Scott (2007) Eq. 4.2.5 pag 95). For any $x \in B_\delta$, it holds

$$f(x) = \sum_{|\alpha|\leq k} \frac{1}{\alpha!} \partial^\alpha f(x_0)(x - x_0)^\alpha \;+\; \sum_{|\alpha|=k+1} \frac{k+1}{\alpha!} (x - x_0)^\alpha \int_0^1 (1-t)^k \partial^\alpha f((1-t)x_0 + tx)dt.$$

By choosing $p(x) = \sum_{|\alpha|\leq k} \frac{1}{\alpha!} \partial^\alpha f(x_0)(x - x_0)^\alpha \;\in \pi_k(\mathbb{R}^d)$ it holds:

$$\|f - p\|_{L^\infty(B_\delta)} \leq \sum_{|\alpha|=k+1} \frac{\delta^{k+1}}{\alpha!} \|\partial^\alpha f\|_{L^\infty(B_\delta)} = C\delta^{k+1},$$

where $C = \sum_{|\alpha|=k+1} \frac{1}{\alpha!} \|\partial^\alpha f\|_{L^\infty(B_\delta)}$ is defined in the lemma. Gathering the previous equations,

$$\sup_{x \in B_\delta} |f(x)| \leq 2\tau + 3C\delta^{k+1}.$$

□

Theorem 8.13 (Bounds on functions with scattered zeros (Narcowich, Ward, and Wendland, 2003; Wendland, 2004))**.** *Let $k, m \in \mathbb{N}$ s.t. $k \leq m$ and $n, d \in \mathbb{N}_+$. Let $r > 0$ and Ω an open set of $\mathbb{R}^d$ of the form $\Omega = \bigcup_{x \in S} B_r(x)$ for some subset S of $\mathbb{R}^d$. Let $\widehat{X} = \{x_1, \ldots, x_n\}$ be a non-empty finite subset of Ω. Let $f \in C^{m+1}(\Omega)$. If $h_{\widehat{X},\Omega} \leq r \max(1, \frac{1}{18k^2})$, then*

$$\sup_{x \in \Omega} |f(x)| \;\leq\; CC_f h_{\widehat{X},\Omega}^{k+1} \;+\; 2 \max_{i \in [n]} |f(x_i)|,$$

where $C = 3\max(1, 18\ k^2)^{k+1}$ and $C_f = \sum_{|\alpha|=k+1} \frac{1}{\alpha!} \|\partial^\alpha f\|_{L^\infty(\Omega)}$.

Proof. First, note that the condition that there exists a set S such that $\Omega = \bigcup_{x \in S} B_r(x)$ implies

$$\forall \delta \leq r,\ \Omega = \bigcup_{x_0 \in S_\delta} B_\delta(x_0), \qquad S_\delta = \{x' \in \Omega \ :\ \exists x \in S,\ \|x - x'\| \leq r - \delta\}.$$

We will now prove the theorem for $k \geq 1$ and then the easier case $k = 0$, where we will use essentially only the Lipschitzianity of f.

Proof of the case $k \geq 1$. The idea of the proof is to apply Lemma 8.8 to a collection of balls of radius δ for a well chosen $\delta \leq r$ and centered in $x_0 \in S_\delta$ defined above. Given $\widehat{X}$, to apply Lemma 8.8 on a ball of radius δ we have to restrict the points in $\widehat{X}$ to the subset belonging to that ball, i.e., $\widehat{Y}_{x_0,\delta} = \widehat{X} \cap B_\delta(x_0)$, $x_0 \in S_\delta$ and $\delta > 0$. The set $\widehat{Y}_{x_0,\delta}$ will have a fill distance $h_{x_0,\delta} = h_{\widehat{Y}_{x_0,\delta}, B_\delta(x_0)}$. First we are going to show that $\widehat{Y}_{x_0,\delta}$ is not empty, when $r > \delta > h_{\widehat{X},\Omega}$. To obtain this result we need to study also the ball $B_{\delta'}(x_0)$ with $\delta' = \delta - h_{\widehat{X},\Omega}$.

Step 1. Showing that $\widehat{Y}_{x_0,\delta}$ is not empty and for any $y \in B_{\delta'}(x_0)$ there exists $z \in \widehat{Y}_{x_0,\delta}$ satisfying $\|y - z\| \leq h_{\widehat{X},\Omega}$. Let $x_0 \in S_\delta$ and $\delta \leq r$. This implies that $B_\delta(x_0) \subseteq \Omega$ by the characterization of Ω in terms of S_δ we gave above. Define now $\delta' = \delta - h_{\widehat{X},\Omega}$ and note that $B_{\delta'}(x_0)$ is non empty, since $\delta' > 0$, and that $B_{\delta'}(x_0) \subset B_\delta(x_0) \subseteq \Omega$. Now note that by definition of fill distance, for any $y \in B_{\delta'}(x_0)$ there exists a $z \in \widehat{X}$ such that $\|z - y\| \leq h_{\widehat{X},\Omega}$. Moreover note that $z \in B_\delta(x_0)$, since $\|x_0 - z\| \leq \|x_0 - y\| + \|y - z\| < \delta - h_{\widehat{X},\Omega} + h_{\widehat{X},\Omega} = \delta$. Since $z \in \widehat{X}$ and also in $B_\delta(x_0)$, then $z \in \widehat{Y}_{x_0,\delta}$ by definition of $\widehat{Y}_{x_0,\delta}$.

Step 2. Showing that $h_{x_0,\delta} \leq 2h_{\widehat{X},\Omega}$. Let $x \in B_\delta(x_0)$. We have seen in the previous step that the ball $B_{\delta'}(x_0)$ is well defined and non empty, with $\delta' = \delta - h_{\widehat{X},\Omega}$. Now note that also $B_{h_{\widehat{X},\Omega}}(x) \cap B_{\delta'}(x_0)$ is not empty, indeed the distance between the centers x, x_0 is strictly smaller than the sum of the two radii, indeed $\|x - x_0\| < \delta = \delta' + h_{\widehat{X},\Omega}$, since $x \in B_\delta(x_0)$. Take $w \in B_{h_{\widehat{X},\Omega}}(x) \cap B_{\delta'}(x_0)$. Since $w \in B_{\delta'}(x_0)$ by Step 1 we know that there exists $z \in \widehat{Y}_{x_0,\delta}$ with $\|w - z\| \leq h_{\widehat{X},\Omega}$. Since $w \in B_{h_{\widehat{X},\Omega}}(x)$, then we know that $\|x - w\| < h_{\widehat{X},\Omega}$. So $\|x - z\| \leq \|x - w\| + \|w - z\| < 2h_{\widehat{X},\Omega}$.

Step 3. Applying Lemma 8.8. Since, by assumption $h_{\widehat{X},\Omega} \leq r/(18k^2)$ and $k \geq 1$, then the choice $\delta = 18k^2 h_{\widehat{X},\Omega}$ implies $r \geq \delta > h_{\widehat{X},\Omega}$. So we can use the characterization of Ω in terms of S_δ and the results in the previous two steps, obtaining that for any $x_0 \in S_\delta$ the set $B_\delta(x_0) \subseteq \Omega$

and moreover the set $\widehat{Y}_{x_0,\delta}$ is not empty and covers $B_\delta(x_0)$ with a fill distance $h_{x_0,\delta} \leq 2h_{\widehat{X},\Omega}$. Since, $h_{x_0,\delta} \leq 2h_{\widehat{X},\Omega} \leq \delta/(9k^2)$ then we can apply Lemma 8.8 to each ball $B_\delta(x_0)$ obtaining

$$\sup_{x\in B_\delta(x_0)} |f(x)| \leq 3C_{\delta,x_0}\delta^{k+1} + 2\max_{z\in\widehat{Y}_{x_0,\delta}} |f(z)|, \qquad C_{\delta,x_0} := \sum_{|\alpha|=k+1} \frac{1}{\alpha!}\|\partial^\alpha f\|_{L^\infty(B_\delta(x_0))}.$$

The proof is concluded by noting that $\Omega = \bigcup_{x_0\in S_\delta} B_\delta(x_0)$ and that for any $x_0 \in S_\delta$ we have $C_{\delta,x_0} \leq C_f$, $\delta^{k+1} \leq (18k^2)^{k+1}h_{\widehat{X},\Omega}^{k+1}$ and moreover that $\max_{z\in\widehat{Y}_{x_0,\delta}} |f(z)| \leq \max_{i\in[n]} |f(x_i)|$, since $\widehat{Y}_{x_0,\delta} \subseteq \widehat{X}$ by construction.

Proof of the case $k = 0$ Since $h_{\widehat{X},\Omega} \leq r$, by assumption, then $\delta = h_{\widehat{X},\Omega}$ implies that Ω admits a characterization as $\Omega = \bigcup_{x_0\in S_\delta} B_\delta(x_0)$. Now let $x \in \Omega$ and choose $x_0 \in S_\delta$ such that $x \in B_\delta(x_0)$. One the one hand, since the segment $[x_0, x]$ is included in Ω, by Taylor inequality, $|f(x) - f(x_0)| \leq C_f\|x - x_0\| \leq C_f h_{\widehat{X},\Omega}$ and $C_f = \sum_{|\alpha|=1} \frac{1}{\alpha!}\|\partial^\alpha f\|_{L^\infty(\Omega)}$. One the other hand, by definition of $h_{\widehat{X},\Omega}$, there exists $z \in \widehat{X} \subset \Omega$ such that $\|z - x_0\| \leq h_{\widehat{X},\Omega} = \delta$. Since both the open segment $[x_0, z) \subset B_\delta(x_0) \subset \Omega$ and $z \in \Omega$, then the whole segment $[x_0, z] \subset \Omega$ and hence we can apply Taylor inequality to show $\|f(x_0) - f(z)\| \leq C_f\|z - x_0\| \leq C_f h_{\widehat{X},\Omega}$. Then we have

$$|f(x)| \leq |f(x) - f(x_0)| + |f(x) - f(z)| + |f(z)| \leq 2C_f h_{\widehat{X},\Omega} + \max_{i\in[n]} |f(x_i)|.$$

The proof of the step $k = 0$ is concluded by noting that the previous inequality holds for every $x \in \Omega$. □

8.C Auxiliary results on RKHS

We recall that the *nuclear norm* of a compact linear operator A is defined as $\|A\|_\star = \mathrm{Tr}(\sqrt{A^*A})$ or equivalently $\|A\|_\star = \sum_{j\in\mathbb{N}} \sigma_j$, where $(\sigma_j)_{j\in\mathbb{N}}$ are the singular values of A (Chapter 7 by Weidmann (1980) or Bhatia (2013) for the finite dimensional analogue).

Lemma 8.9. *Let Ω be a set, k be a kernel and $\mathcal{H}$ the associated RKHS. Let $A : \mathcal{H} \to \mathcal{H}$ be a trace class operator. If $\mathcal{H}$ satisfies Assumption 8.2(a), then*

$$\|r_A\|_\mathcal{H} \leq \mathsf{M}\|A\|_\star, \quad \textit{where} \quad r_A(x) := \langle \phi(x), A\phi(x)\rangle\,, \ \forall x \in \Omega,$$

and $\|A\|_\star$ is the nuclear norm *of A. We recall that if $A \in \mathbb{S}_+(\mathcal{H})$ then $\|A\|_\star = \mathrm{Tr}(A)$.*

Proof. Since A is compact, it admits a singular value decomposition $A = \sum_{i\in\mathbb{N}} \sigma_i u_i \otimes v_i$. Here, $(\sigma_j)_{j\in\mathbb{N}}$ is a non-increasing sequence of non-negative eigenvalues converging to zero, and $(u_j)_{j\in\mathbb{N}}$ and $(v_j)_{j\in\mathbb{N}}$ are two orthonormal families of corresponding eigenvectors, (a family (e_j) is said to be orthonormal if for $i, j \in \mathbb{N}$, $\langle e_i, e_j\rangle = 1$ if $i = j$ and $\langle e_i, e_j\rangle = 0$ otherwise) (Weidmann, 1980). Note that we can write r_A using this decomposition as $r_A(x) = \sum_{i\in\mathbb{N}} \sigma_i u_i(x)v_i(x) = \sum_{i\in\mathbb{N}} \sigma_i\,(u_i \cdot v_i)(x)$, for all $x \in \Omega$, where we denote by $\cdot$ the pointwise multiplication between two functions (this equality is justified by the following absolute convergence bound). By Assumption 8.2(a), the fact that A is trace-class (i.e., $\|A\|_\star < \infty$) and the fact that u_j, v_j satisfy $\|u_j\|_\mathcal{H} = \|v_j\|_\mathcal{H} = 1, j \in \mathbb{N}$, the following holds

$$\begin{aligned}\|r_A\|_\mathcal{H} &= \|\sum_{j\in\mathbb{N}} \sigma_j(u_j \cdot v_j)\|_\mathcal{H} \leq \sum_{j\in\mathbb{N}} \sigma_j\|u_j \cdot v_j\|_\mathcal{H} \\ &\leq \mathsf{M}\sum_{j\in\mathbb{N}} \sigma_j\|u_j\|_\mathcal{H}\|v_j\|_\mathcal{H} \leq \mathsf{M}\sum_{j\in\mathbb{N}} \sigma_j = \mathsf{M}\|A\|_\star.\end{aligned}$$

In the case where $A \in \mathbb{S}_+(\mathcal{H})$, we have $\|A\|_\star = \mathrm{Tr}(\sqrt{A^*A}) = \mathrm{Tr}(A)$. □

8.C .1 Proof of Lemma 8.2

Given the kernel k, the associated RKHS $\mathcal{H}$ and the canonical feature map $\phi : \Omega \to \mathcal{H}$ and a set of distinct points $\widehat{X} = \{x_1, \dots, x_n\}$ define the *kernel matrix* $K \in \mathbb{R}^{n\times n}$ as $K_{i,j} = \langle \phi(x_i), \phi(x_j)\rangle = k(x_i, x_j)$ for all $i, j \in [n]$. Note that, since k is a p.d. kernel, then K is positive semidefinite, moreover when k is universal, then $\phi(x_1), \dots, \phi(x_n)$ are linearly independent, so K is full rank and hence invertible. Universality of k is guaranteed since $\mathcal{H}$ contains the $C_0^\infty(\Omega)$ functions, by Assumption 8.1(a), and so can approximate continuous functions over compacts in Ω (Steinwart and Christmann, 2008). Denote by R the upper triangular matrix corresponding to the Cholesky decomposition of K, i.e., R satisfies $K = R^\top R$. We are ready to start the proof of Lemma 8.2.

Proof. Denote by $\widehat{S} : \mathcal{H} \to \mathbb{R}^n$ the linear operator that acts as follows

$$\widehat{S}g \;=\; (\,\langle \phi(x_1), g\rangle \;,\; \dots,\; \langle \phi(x_n), g\rangle\,) \in \mathbb{R}^n, \qquad \forall g \in \mathcal{H}.$$

Define $\widehat{S}^* : \mathbb{R}^n \to \mathcal{H}$, i.e., the adjoint of $\widehat{S}$, as $\widehat{S}^*\beta = \sum_{i=1}^n \beta_i \phi(x_i)$ for $\beta \in \mathbb{R}^n$. Note, in particular, that $K = \widehat{S}\widehat{S}^*$ and that $\widehat{S}^* e_j = \phi(x_i)$, where e_j is the j-th element of the canonical basis of $\mathbb{R}^n$. We define the operator $V = R^{-\top}\widehat{S}$ and its adjoint $V^* = \widehat{S}^* R^{-1}$. By using the definition of V, the fact that $K = R^\top R$ by construction of R, and the fact that $K = \widehat{S}\widehat{S}^*$, we derive two facts.

On the one hand,

$$VV^* = R^{-\top}\widehat{S}\widehat{S}^* R^{-1} = R^{-\top} K R^{-1} = R^{-\top} R^\top R R^{-1} = I.$$

On the other hand, P is a projection operator, i.e., $P^2 = P$, P is positive definite and its range is $\operatorname{range} P = \operatorname{span} \phi(x_i) \mid i \in [n]$, implying $P\phi(x_i) = \phi(x_i)$ for all $i \in [n]$. Indeed, using the equation above, $P^2 = V^*VV^*V = V^*(VV^*)V = V^*V = P$, and the positive-semi-definiteness of P is given by construction since it is the product of an operator and its adjoint. Moreover, the range of P is the same as that of V^* which in turn is the same as that of S^*, since R is invertible : $\operatorname{range} P = \operatorname{span} \phi(x_i) \mid i \in [n]$.

Finally, note that since $k(x, x') = \langle \phi(x), \phi(x')\rangle$, for any $x, x' \in \Omega$, then for any $j \in [n]$, Φ_j is characterized by

$$\begin{aligned}\Phi_j &= R^{-\top}(k(x_1, x_j), \dots, k(x_n, x_j)) \\ &= R^{-\top}(\langle \phi(x_1), \phi(x_j)\rangle, \dots, \langle \phi(x_n), \phi(x_j)\rangle) = R^{-\top}\widehat{S}\phi(x_j) = V\phi(x_j).\end{aligned}$$

□

8.D The constants of translation invariant and Sobolev kernels

8.D .1 Results for translation invariant and Sobolev kernels

Lemma 8.10. *Let Ω be a set and let $k(x, x') = v(x - x')$ for all $x, x' \in \Omega$, be a translation invariant kernel for some function $v : \mathbb{R}^d \to \mathbb{R}$. Denote by $\tilde{v}$ the Fourier transform of v. Let $\mathcal{H}$ be the associated RKHS. For any $f, g \in \mathcal{H}$ we have*

$$\|f \cdot g\|_{\mathcal{H}} \le C\|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}, \qquad C = (2\pi)^{d/4}\left\|\frac{\tilde{v} \star \tilde{v}}{\tilde{v}}\right\|^{1/2}_{L^\infty(\mathbb{R}^d)}.$$

In particular, if there exists a non-increasing $g : [0, \infty] \to (0, \infty]$ *s.t.* $\tilde{v}(\omega) \leq g(\|\omega\|)$*, then*

$$C \leq \sqrt{2}(2\pi)^{d/2} v(0)^{1/2} \sup_{\omega \in \mathbb{R}^d} \sqrt{\frac{g(\frac{1}{2}\|\omega\|)}{\tilde{v}(\omega)}}.$$

Proof. First note that by as recalled in proposition 8.8, there exists an extension operator, i.e., a partial isometry $E : \mathcal{H} \to \mathcal{H}(\mathbb{R}^d)$ such that $r = Eu$ satisfies $r(x) = u(x)$ for all $x \in \Omega$ and $\|u\|_{\mathcal{H}} = \|r\|_{\mathcal{H}}$, for any $u \in \mathcal{H}$. Moreover there exists a restriction operator $R : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}$, as recalled in proposition 8.8, such that $RE : \mathcal{H} \to \mathcal{H}$ is the identity operator and $ER : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}(\mathbb{R}^d)$ is a projection operator whose range is $\mathcal{H}$. Moreover, note that $f \cdot g = R(Ef \cdot Eg)$ since for any $x \in \Omega$, $(R(Ef \cdot Eg))(x) = (Ef)(x)(Eg)(x) = f(x)g(x) = (f \cdot g)(x)$. Since ER is a projection operator, then $\|ER\|_{\text{op}} \leq 1$, hence

$$\begin{aligned}\|f \cdot g\|_{\mathcal{H}} &= \|R(Ef \cdot Eg)\|_{\mathcal{H}} = \|ER(Ef \cdot Eg)\|_{\mathcal{H}(\mathbb{R}^d)} \\ &\leq \|ER\|_{\text{op}} \|Ef \cdot Eg\|_{\mathcal{H}(\mathbb{R}^d)} \leq \|Ef \cdot Eg\|_{\mathcal{H}(\mathbb{R}^d)}.\end{aligned}$$

Let $a = Ef$ and $b = Eg$. Denote by $\tilde{a}, \tilde{b}$ their Fourier transform and by $\widetilde{a \cdot b}$ the Fourier transform of $a \cdot b$ (see proposition 8.2 for more details). By expanding the definition of the Hilbert norm of translation invariant kernel

$$\|Ef \cdot Eg\|^2_{\mathcal{H}(\mathbb{R}^d)} = \|a \cdot b\|^2_{\mathcal{H}(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\widetilde{a \cdot b}(\omega)|^2}{\tilde{v}(\omega)} d\omega.$$

Now we bound $\widetilde{a \cdot b}$. Since $\widetilde{a \cdot b} = (2\pi)^{d/2} \tilde{a} \star \tilde{b}$ (see proposition 8.2) where $\star$ corresponds to the convolution, by expanding it and by applying Cauchy-Schwarz we obtain

$$\begin{aligned}(2\pi)^{-d/2} |\widetilde{a \cdot b}(\omega)|^2 = |(\tilde{a} \star \tilde{b})(\omega)|^2 &= \left(\int_{\mathbb{R}^d} \tilde{a}(\sigma) \tilde{b}(\omega - \sigma) d\sigma \right)^2 \\ &= \left(\int_{\mathbb{R}^d} \frac{\tilde{a}(\sigma)}{\sqrt{\tilde{v}(\sigma)}} \frac{\tilde{b}(\omega - \sigma)}{\sqrt{\tilde{v}(\omega - \sigma)}} \sqrt{\tilde{v}}(\sigma) \sqrt{\tilde{v}}(\omega - \sigma) d\sigma \right)^2 \\ &\leq \int_{\mathbb{R}^d} \frac{\tilde{a}^2}{\tilde{v}}(\sigma) \frac{\tilde{b}^2}{\tilde{v}}(\omega - \sigma) d\sigma \int_{\mathbb{R}^d} \tilde{v}(\sigma) \tilde{v}(\omega - \sigma) d\sigma = \left(\frac{\tilde{a}^2}{\tilde{v}} \star \frac{\tilde{b}^2}{\tilde{v}} \right)(\omega) \; (\tilde{v} \star \tilde{v})(\omega).\end{aligned}$$

By using the bound above together with Hölder inequality and Young inequality for convolutions, we have

$$\begin{aligned}(2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\widetilde{a \cdot b}(\omega)|^2}{\tilde{v}(\omega)} d\omega &\leq \int_{\mathbb{R}^d} \left(\frac{\tilde{a}^2}{\tilde{v}} \star \frac{\tilde{b}^2}{\tilde{v}} \right)(\omega) \frac{(\tilde{v} \star \tilde{v})(\omega)}{\tilde{v}(\omega)} d\omega \leq \left\| \frac{\tilde{a}^2}{\tilde{v}} \star \frac{\tilde{b}^2}{\tilde{v}} \right\|_{L^1(\mathbb{R}^d)} \left\| \frac{\tilde{v} \star \tilde{v}}{\tilde{v}} \right\|_{L^\infty(\mathbb{R}^d)} \\ &\leq \left\| \frac{\tilde{a}^2}{\tilde{v}} \right\|_{L^1(\mathbb{R}^d)} \left\| \frac{\tilde{b}^2}{\tilde{v}} \right\|_{L^1(\mathbb{R}^d)} \left\| \frac{\tilde{v} \star \tilde{v}}{\tilde{v}} \right\|_{L^\infty(\mathbb{R}^d)} \\ &= (2\pi)^{d/2} \left\| \frac{\tilde{v} \star \tilde{v}}{\tilde{v}} \right\|_{L^\infty(\mathbb{R}^d)} \|a\|^2_{\mathcal{H}(\mathbb{R}^d)} \|b\|^2_{\mathcal{H}(\mathbb{R}^d)} = C^2,\end{aligned}$$

where in the last step we used the definitions of inner products for translation invariant kernels. The proof is concluded by noting that $\|a\|_{\mathcal{H}(\mathbb{R}^d)} = \|Ef\|_{\mathcal{H}(\mathbb{R}^d)} = \|f\|_{\mathcal{H}}$ and the same holds for b, i.e., $\|b\|_{\mathcal{H}(\mathbb{R}^d)} = \|g\|_{\mathcal{H}}$. A final consideration is that C can be further bounded by applying proposition 8.12 and noting that $v(0) = (2\pi)^{-d/2} \int \tilde{v}(\omega) d\omega = (2\pi)^{-d/2} \|\tilde{v}\|_{L^1(\mathbb{R}^d)}$, via the characterization of v in terms of $\tilde{v}$ in proposition 8.2(e), since $\tilde{v}(\omega) \geq 0$ and integrable. □

Proposition 8.12. *Let $u \in L^1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ be $u(x) \geq 0$ for $x \in \mathbb{R}^d$ and such that there exists a non-increasing function $g : [0, \infty) \to (0, \infty)$ satisfying $u(x) \leq g(\|x\|)$ for all $x \in \mathbb{R}^d$. Then it holds :*

$$\forall x \in \mathbb{R}^d,\ 0 \leq (u \star u)(x) \leq 2\|u\|_{L^1(\mathbb{R}^d)} g(\tfrac{1}{2}\|x\|).$$

In particular, if $u > 0$, it holds

$$\left\| \frac{u \star u}{u} \right\|_{L^\infty(\mathbb{R}^d)} \leq 2\|u\|_{L^1(\mathbb{R}^d)} \sup_{x \in \mathbb{R}^d} \frac{g(\frac{1}{2}\|x\|)}{u(x)}.$$

Proof. For any $x \in \mathbb{R}^d$,

$$(u \star u)(x) = \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} u(y)u(x-y)dy.$$

Let $S_x = \{y \mid \|x - y\| \leq \frac{1}{2}\|x\|\}$. Note that, when $y \in \mathbb{R}^d \setminus S_x$, then $\|x - y\| > \frac{1}{2}\|x\|$. Instead, when $y \in S_x$, then

$$\tfrac{1}{2}\|x\| \leq \|x\| - \|x - y\| \leq \|y\|.$$

Since g is non-increasing, for any $x \in \mathbb{R}^d$ we have

$$\begin{aligned}
\int_{\mathbb{R}^d} u(y)u(x-y)dy &= \int_{S_x} u(y)u(x-y)dy \ + \int_{\mathbb{R}^d \setminus S_x} u(y)u(x-y)dy \\
&\leq \int_{S_x} g(\|y\|)u(x-y)dy \ + \int_{\mathbb{R}^d \setminus S_x} u(y)g(\|x-y\|)dy \\
&\leq \int_{S_x} g(\tfrac{1}{2}\|x\|)u(x-y)dy \ + \int_{\mathbb{R}^d \setminus S_x} u(y)g(\tfrac{1}{2}\|x\|)dy \\
&\leq \int_{\mathbb{R}^d} g(\tfrac{1}{2}\|x\|)u(x-y)dy \ + \int_{\mathbb{R}^d} u(y)g(\tfrac{1}{2}\|x\|)dy \\
&= \int_{\mathbb{R}^d} g(\tfrac{1}{2}\|x\|)u(y)dy \ + \int_{\mathbb{R}^d} u(y)g(\tfrac{1}{2}\|x\|)dy = 2\ g(\tfrac{1}{2}\|x\|) \int_{\mathbb{R}^d} u(y)dy,
\end{aligned}$$

where: in the first inequality we bounded $u(y)$ with $g(\|y\|)$ and $u(x-y)$ with $g(\|x-y\|)$, in the first and the second integral, respectively; in the second inequality we bounded $g(\|y\|)$ with $g(\frac{1}{2}\|x\|)$, since $\|y\| \geq \frac{1}{2}\|x\|$ when $y \in S_x$ and we bounded $g(\|x-y\|)$ with $g(\frac{1}{2}\|x\|)$, since $\|x-y\| \geq \frac{1}{2}\|x\|$ when $y \in \mathbb{R}^d \setminus S_x$; in the third we extended the integration domains to $\mathbb{R}^d$. □

8.D .2 Proof of proposition 8.1

Proof. We prove here that the Sobolev kernel satisfies Assumption 8.2. Let $k = k_s$ from Eq. (8.7). As we have seen in Example 8.1 $\mathcal{H} = W_2^s(\Omega)$ and $\|\cdot\|_{W_2^s(\Omega)}$ is equivalent to $\|\cdot\|_{\mathcal{H}}$, when $s > d/2$ and Ω satisfies Assumption 8.1(a) since this assumption implies that Ω satisfies the *cone condition* (Wendland, 2004).

Recall that k is translation invariant, i.e., $k(x, x') = v(x - x')$ for any $x, x' \in \mathbb{R}^d$, with v defined in Example 8.1. The Fourier transform of v is $\tilde{v}(\omega) = C_0(1 + \|\omega\|^2)^{-s}$ with $C_0 = \frac{2^{d/2}\Gamma(s)}{\Gamma(s-d/2)}$ (Wendland, 2004). In the rest of the proof, C_0 will always refer to this constant.

We are going to divide the proof in one step per point of Assumption 8.2.

Proof of Assumption 8.2(d) for the Sobolev kernel. Let $\alpha \in \mathbb{N}^d$, $m = |\alpha|$. Assume $m < s - d/2$, i.e., $m \in \{1, \ldots, \lfloor s - (d+1)/2 \rfloor\}$. Since k is translation invariant, then $\partial_x^\alpha \partial_y^\alpha k(x, y) = (-1)^m\ v_{2\alpha}(x - y)$ with $v_{2\alpha}(z) = \partial_z^{2\alpha} v(z)$ for all $z \in \mathbb{R}^d$. So

$$\sup_{x,y \in \Omega} |\partial_x^\alpha \partial_y^\alpha k(x, y)| = \sup_{x,y \in \Omega} |\partial_x^\alpha \partial_y^\alpha v(x - y)| \leq \sup_{z \in \mathbb{R}^d} |\partial_z^{2\alpha} v(z)| \leq (2\pi)^{-d/2} \|\omega^{2\alpha} \tilde{v}(z)\|_{L^1(\mathbb{R}^d)},$$

where in the last step we used elementary properties of the Fourier transform (in particular the ones recalled in propositions 8.2(c) and 8.2(e)). Let $S_{d-1} = 2\frac{\pi^{d/2}}{\Gamma(d/2)}$ be the area of the $d-1$ dimensional sphere. Since $m < s - d/2$ and $\tilde{v} \geq 0$,

$$\begin{aligned}\|\omega^{2\alpha}\tilde{v}(z)\|_{L^1(\mathbb{R}^d)} &\leq \int_{\mathbb{R}^d} \|\omega\|^{2m}\tilde{v}(\omega)d\omega = C_0 S_{d-1}\int_0^\infty \frac{r^{2m+d-1}}{(1+r^2)^s}dr \\ &= C_0 S_{d-1}\int_0^\infty \frac{t^{m+d/2-1}}{2(1+t)^s}dt = C_0 S_{d-1}\frac{\Gamma(m+d/2)\Gamma(s-d/2-m)}{2\Gamma(s)},\end{aligned}$$

where we performed a change of variable $r = \sqrt{t}$ and $dr = \frac{dt}{2\sqrt{t}}$ and applied Eq. 5.12.3 pag. 142 by Olver, Lozier, Boisvert, and Clark (2010) to the resulting integral. Thus, Assumption 8.2(d) holds with

$$\mathsf{D}_m^2 = C_0\frac{\pi^{d/2}\Gamma(m+d/2)\Gamma(s-m-d/2)}{\Gamma(d/2)\Gamma(s)} = \frac{(2\pi)^{d/2}\Gamma(m+d/2)\Gamma(s-d/2-m)}{\Gamma(s-d/2)\Gamma(d/2)}.$$

Proof of Assumption 8.2(a) for the Sobolev kernel. First, note that $C^\infty(\mathbb{R}^d)|_\Omega \subset W_\infty^s(\Omega) \subset W_2^s(\Omega)$. Indeed, since Ω is bounded, for any $f \in C^\infty(\mathbb{R}^d)$, $\|\partial^\alpha f|_\Omega\|_{L^\infty(\Omega)} < \infty$ for any $\alpha \in \mathbb{N}^d$. This shows that $f|_\Omega \in W_\infty^s(\Omega)$. Moreover $W_\infty^s(\Omega) \subset W_2^s(\Omega)$ since $\|\cdot\|_{L^2(\Omega)} \leq \mathrm{vol}(\Omega)^{1/2}\|\cdot\|_{L^\infty(\Omega)}$ because Ω is bounded. Second, since $\tilde{v}(\omega) = g_s(\|\omega\|)$ with $g_s(t) = C_0(1+t^2)^{-s}$, positive and non-increasing, we can apply Lemma 8.10. Therefore, for $C = \sqrt{2}(2\pi)^{d/2}v(0)^{1/2}\sup_{t\geq 0}\left(\frac{g_s(t/2)}{g_s(t)}\right)^{1/2}$ it holds $\|f\cdot g\|_{\mathcal{H}} \leq C\|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}$. In particular we have $\sup_{t\geq 0}\left(\frac{g_s(t/2)}{g_s(t)}\right)^{1/2} \leq 2^s$ and $v(0) = 1$, since $\lim_{t\to 0} t^{s-d/2}\mathcal{K}_{s-d/2}(t) = \Gamma(s-d/2)/2^{1+d/2-s} = 1/C_0$ (Eq. 10.30.2 pag. 252 by Olver, Lozier, Boisvert, and Clark (2010)) and $v(x) = C_0 t^{s-d/2}\mathcal{K}_{s-d/2}(t)$, $t = \|x\|$. Thus, Assumption 8.2(a) holds with constant

$$\mathsf{M} = \pi^{d/2}2^{(2s+d+1)/2}.$$

Proof of Assumption 8.2(b) for the Sobolev kernel. First we recall from the work by Adams and Fournier (2003) that for any $s > d/2$, there exists a constant C_s such that

$$\forall h \in W_2^s(\mathbb{R}^d),\ \|h\|_{L^\infty(\mathbb{R}^d)} \leq C_s\|h\|_{W_2^s(\mathbb{R}^d)}.$$

In particular, this shows that $W_2^s(\mathbb{R}^d) \subset L^\infty(\mathbb{R}^d)$. Fix such a constant C_s in the rest of the proof.

Let $p \in \mathbb{N}$ and $g \in C^\infty(\mathbb{R}^p)$ with $g(0,0,\ldots,0) = 0$. From (i) of Thm. 11 in the work by Sickel (1992), there exists a constant c_g depending only on g, p, s such that for any $h_1, \ldots, h_p \in W_2^s(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, it holds

$$\|g(h_1,\ldots,h_p)\|_{W_2^s(\mathbb{R}^d)} \ \leq\ c_g \sup_{i\in[p]}\ \|h_i\|_{W_2^s(\mathbb{R}^d)}\left(1 + \|h_i\|_{L^\infty(\mathbb{R}^d)}^{\max(0,s-1)}\right).$$

Since $s > d/2$, the bound above shows, in particular, that for any $h_1, \ldots, h_p \in W_2^s(\mathbb{R}^d)$, it holds

$$\|g(h_1,\ldots,h_p)\|_{W_2^s(\mathbb{R}^d)} \ \leq\ c_g' \sup_{i\in[p]}\ \left(\|h_i\| + \|h_i\|_{W_2^s(\mathbb{R}^d)}^{\max(1,s)}\right), \qquad c_g' = c_g\max\left(1, C_s^{\max(0,s-1)}\right).$$

Since $W_2^s(\mathbb{R}^d) = \mathcal{H}(\mathbb{R}^d)$ and $\|\cdot\|_{W_2^s(\mathbb{R}^d)}$ and $\|\cdot\|_{\mathcal{H}(\mathbb{R}^d)}$ are equivalent (see the work by Adams and Fournier (2003)), the previous inequality holds for $\|\cdot\|_{\mathcal{H}(\mathbb{R}^d)}$ with a certain constant c_g' depending only on g, p, s, d. In particular, this implies that $g(h_1,\ldots,h_p) \in \mathcal{H}(\mathbb{R}^d)$ for any $h_1,\ldots,h_p \in \mathcal{H}(\mathbb{R}^d)$. Now we are going to prove the same implication for the restriction on Ω.

First note that any function in $a \in C^\infty(\mathbb{R}^p)$ can be written as $a(z) = q\,1(z) + g(z)$, $z \in \mathbb{R}^p$ where $q = a(0,0,\cdots,0) \in \mathbb{R}$, $g \in C^\infty(\mathbb{R}^p)$ with $g(0,0,\cdots,0) = 0$ and $1(z) = 1$ for all $z \in \mathbb{R}^p$. Recall the definition and basic results on the extension operator $E : \mathcal{H} \to \mathcal{H}(\mathbb{R}^d)$ from proposition 8.8. For any $f_1, \ldots, f_p \in \mathcal{H}$, note that $g((Ef_1)(x), \ldots, (Ef_p)(x)) = g(f_1(x), \ldots, f_p(x))$ for all $x \in \Omega$. We can now apply the results of proposition 8.8 to show that $g(f_1, \ldots, f_p) \in \mathcal{H}$:

$$\begin{aligned}
\|g(f_1,\ldots,f_p)\|_{\mathcal{H}} &= \inf_u \|u\|_{\mathcal{H}(\mathbb{R}^d)} \; s.t. \; u(x) = g(f_1(x),\ldots,f_p(x)) \; \forall x \in \Omega \\
&\leq \|g(Ef_1,\ldots,Ef_p)\|_{\mathcal{H}(\mathbb{R}^d)} \\
&\leq c'_g \sup_{j\in[p]} \|Ef_j\|_{\mathcal{H}(\mathbb{R}^d)} + \|Ef_j\|^{\max(1,s)}_{\mathcal{H}(\mathbb{R}^d)} \\
&= c'_g \sup_{j\in[p]} \|f_j\|_{\mathcal{H}} + \|f_j\|^{\max(1,s)}_{\mathcal{H}} < \infty,
\end{aligned}$$

where in the last step we used the fact that $\|\cdot\|_{\mathcal{H}} = \|E\,\cdot\|_{\mathcal{H}(\mathbb{R}^d)}$. The proof of this point is concluded by noting that, $a(f_1, \ldots, f_p) \in \mathcal{H}$, since $1 \in \mathcal{H}$, due to the Point (a) above, and

$$\|a(f_1,\ldots,f_p)\|_{\mathcal{H}} \leq q\|1\|_{\mathcal{H}} + \|g(f_1,\ldots,f_p)\|_{\mathcal{H}} < \infty.$$

Proof of Assumption 8.2(c) for the Sobolev kernel. This proof is done in Lemma 8.11, right below. □

Before stating Lemma 8.11 we are going to recall some properties. First, recall the Young inequality :

$$\forall f \in L^2(\mathbb{R}^d),\ \forall g \in L^1(\mathbb{R}^d),\ \|f \star g\|_{L^2(\mathbb{R}^d)} \leq \|f\|_{L^2(\mathbb{R}^d)}\ \|g\|_{L^1(\mathbb{R}^d)}.$$

Moreover, by definition of the Sobolev kernel, it is a translation-invariant kernel with v defined in Example 8.1, with Fourier transform $\tilde{v}(\omega) = C_0(1 + \|\omega\|^2)^{-s}$. Let $\mathcal{H}(\mathbb{R}^d)$ be the reproducing kernel Hilbert space on $\mathbb{R}^d$ associated to the Sobolev kernel k_s. As recalled in proposition 8.9, the $\mathcal{H}(\mathbb{R}^d)$-norm is characterized by

$$\forall f \in \mathcal{H}(\mathbb{R}^d),\ \|f\|_{\mathcal{H}(\mathbb{R}^d)} = (2\pi)^{-d/4}\|\tilde{f}/\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)}, \tag{8.38}$$

where $\tilde{f} = \mathcal{F}(f)$ is the Fourier transform of f (see the work by Adams and Fournier (2003)). Then we recall that $\tilde{v} \in L^1(\mathbb{R}^d)$, since $s > d/2$, so for any $f \in \mathcal{H}(\mathbb{R}^d)$

$$\|\tilde{f}\|_{L^1(\mathbb{R}^d)} = \|\sqrt{\tilde{v}}\tilde{f}/\sqrt{\tilde{v}}\|_{L^1(\mathbb{R}^d)} \leq \|\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)}\|\tilde{f}/\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)} = C_1\|f\|_{\mathcal{H}(\mathbb{R}^d)}. \tag{8.39}$$

where $C_1 = (2\pi)^{d/4}\|\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)}$. A useful consequence of the inequality above is obtained by considering that $\|f\|_{L^\infty(\mathbb{R}^d)}$ is bounded by the L^1 norm of $\tilde{f}$ (see proposition 8.2(e)), then

$$\|f\|_{L^\infty} \leq (2\pi)^{-d/2}\|\tilde{f}\|_{L^1(\mathbb{R}^d)} \leq C_2\|f\|_{\mathcal{H}(\mathbb{R}^d)}, \tag{8.40}$$

where $C_2 = (2\pi)^{-d/4}\|\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)}$.

Lemma 8.11 (Assumption 8.2(c) for Sobolev Kernels)**.** *Let $\mathcal{H}$ be the RKHS associated to the translation invariant Sobolev Kernel defined in Example 8.1, with $s > d/2$. Then Assumption 8.2(c) is satisfied.*

Proof. For the rest of the proof we fix $u : \Omega \to \mathbb{R}$ with $u \in \mathcal{H}$, $r > 0$ and $z \in \mathbb{R}^d$ such that $B_r(z) \subset \Omega$. Let $E_\Omega : \mathcal{H} \to \mathcal{H}(\mathbb{R}^d)$ be the extension operator from Ω to $\mathbb{R}^d$ (its properties are recalled in proposition 8.8).Let $\chi \in C_0^\infty(\mathbb{R}^d)$ be given by Lemma 8.6 such that $\chi = 1$ on $B_r(z)$, $\chi = 0$ on $\mathbb{R}^d \setminus B_{2r}(z)$ and $\chi \in [0,1]$. Define for any $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$

$$h_t(x) = \chi(x) w_t(x), \qquad w_t(x) = w((1-t)z + tx), \qquad w = E_\Omega u.$$

In particular we recall that, since E_Ω is a partial isometry (see proposition 8.8) then $\|w\|_{\mathcal{H}(\mathbb{R}^d)} = \|u\|_{\mathcal{H}}$. **Step 1. Fourier transform of w_t.** Denote with $\widetilde{w}$ the Fourier transform of w which is well defined since $w \in \mathcal{H}(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$ (see the work by Adams and Fournier (2003)), with $\tilde{\chi}$ the Fourier transform of χ. Since For any $t \neq 0$, denote with $\widetilde{w_t}$ the Fourier transform of w_t which is well defined using the results of proposition 8.2, and which satisfies

$$\forall t \neq 0,\ \forall \omega \in \mathbb{R}^d,\ \widetilde{w_t}(\omega) = |t|^{-d} e^{i\frac{1-t}{t} z^\top \omega} \tilde{w}(\omega/t).$$

Step 2. Separating low and high order derivatives of h_t, and bounding the low order terms. For $t \neq 0$, denote with $\widetilde{h_t}$ the Fourier transform of h_t which is well defined since χ is bounded and $w_t \in L^2(\mathbb{R}^d)$. We will now bound $\|h_t\|_{\mathcal{H}(\mathbb{R}^d)}$ for all $t \neq 0$, by using the characterization in Eq. (8.38). Since $(x+y)^s \leq 2^{\max(s-1,0)}(x^s + y^s)$ for any $x, y \geq 0$, $s \geq 0$, then $(1 + \|\omega\|^2)^{s/2} \leq c_1(1 + \|\omega\|^s)$ for any $\omega \in \mathbb{R}^d$, with $c_1 = 2^{\max(s/2-1,0)}$ so using Eq. (8.38), we have

$$\sqrt{C_0}(2\pi)^{d/4}\|h_t\|_{\mathcal{H}(\mathbb{R}^d)} = \|(1 + \|\cdot\|^2)^{s/2}\widetilde{h_t}\|_{L^2(\mathbb{R}^d)} \leq c_1 \|\widetilde{h_t}\|_{L^2(\mathbb{R}^d)} + c_1 \| \ |\cdot|^s_{\mathbb{R}^d} \widetilde{h_t}\|_{L^2(\mathbb{R}^d)}.$$

The first term on the right hand side can easily be bounded using the fact that the Fourier transform is an isometry of $L^2(\mathbb{R}^d)$ (see proposition 8.2 for more details), indeed

$$\|\widetilde{h_t}\|_{L^2(\mathbb{R}^d)} = \|h_t\|_{L^2(\mathbb{R}^d)} = \|\chi \cdot w_t\|_{L^2(\mathbb{R}^d)} \leq \|w_t\|_{L^\infty(\mathbb{R}^d)} \|\chi\|_{L^2(\mathbb{R}^d)} < \infty.$$

since $\chi \in C_0^\infty(\mathbb{R}^d)$ by definition, so it it bounded and has compact support, implying that $\|\chi\|_{L^2(\mathbb{R}^d)} < \infty$, moreover $\|w_t\|_{L^\infty(\mathbb{R}^d)} = \|w\|_{L^\infty(\mathbb{R}^d)}$ and $\|w\|_{L^\infty(\mathbb{R}^d)} \leq C_2 \|w\|_{\mathcal{H}(\mathbb{R}^d)}$ as recalled in Eq. (8.40) (the constant C_2 is defined in the same equation).

Step 3. Decomposing the high order derivatives of h_t. Note that since $\widetilde{h_t} = \widetilde{\chi \cdot w_t}$, by property of the Fourier transform (see proposition 8.2(b)), $\widetilde{\chi \cdot w_t} = (2\pi)^{d/2} \widetilde{\chi} \star \widetilde{w_t}$. Moreover, since $\|\omega\|^s \leq (\|\omega - \eta\| + \|\eta\|)^s \leq c_s(\|\omega - \eta\|^s + \|\eta\|^s)$ for any $\omega, \eta \in \mathbb{R}^d$, with $c = 2^{\max(s-1,0)}$, then, for all $\omega \in \mathbb{R}^d$ we have

$$\begin{aligned}
\|\omega\|^s |\widetilde{h_t}(\omega)| &= \|\omega\|^s |\widetilde{\chi \cdot w_t}(\omega)| = \|\omega\|^s (2\pi)^{\frac{d}{2}} |(\tilde{\chi} \star \tilde{w_t})(\omega)| = (2\pi)^{\frac{d}{2}} \Big| \int_{\mathbb{R}^d} \|\omega\|^s \widetilde{\chi}(\eta) \widetilde{w_t}(\omega - \eta) d\eta \Big| \\
&\leq (2\pi)^{\frac{d}{2}} c \int_{\mathbb{R}^d} (|\widetilde{\chi}(\eta)| \ \|\eta\|^s) \ |\widetilde{w_t}(\omega - \eta)| \ d\eta \ + \ (2\pi)^{\frac{d}{2}} c \int_{\mathbb{R}^d} |\widetilde{\chi}(\eta)| \ (|\widetilde{w_t}(\omega - \eta)| \ \|\omega - \eta\|^s) \ d\eta \\
&= c\,((J_s|\widetilde{\chi}|) \star |\widetilde{w_t}|)(\omega) \ + \ c\,(|\tilde{\chi}| \star (J_s|\widetilde{w_t}|))(\omega),
\end{aligned}$$

where we denoted by J_s the function $J_s(\omega) = \|\omega\|^s$ for any $\omega \in \mathbb{R}^d$. Applying Young's inequality, it holds :

$$\begin{aligned}
\|J_s \widetilde{h_t}\|_{L^2(\mathbb{R}^d)} &\leq c\,\|(J_s|\widetilde{\chi}|) \star |\widetilde{w_t}|\|_{L^2(\mathbb{R}^d)} \ + \ c\||\tilde{\chi}| \star (J_s|\widetilde{w_t}|)\|_{L^2(\mathbb{R}^d)} \\
&\leq c\|J_s\widetilde{\chi}\|_{L^2(\mathbb{R}^d)} \|\widetilde{w_t}\|_{L^1(\mathbb{R}^d)} + c\|J_s\widetilde{w_t}\|_{L^2(\mathbb{R}^d)} \ \|\widetilde{\chi}\|_{L^1(\mathbb{R}^d)}.
\end{aligned}$$

Step 4. Bounding the elements of the decomposition. Now we are ready to bound the four terms of the decomposition of $\|J_s\widetilde{h_t}\|_{L^2(\mathbb{R}^d)}$. First term, since $\chi \in C_0^\infty(\mathbb{R}^d) \subset \mathcal{H}(\mathbb{R}^d)$, and $J_s(\omega) \leq$

$\sqrt{C_0/\tilde{v}(\omega)}$ for any $\omega \in \mathbb{R}^d$, then $\|J_s\widetilde{\chi}\|_{L^2(\mathbb{R}^d)} \leq \sqrt{C_0}\|\widetilde{\chi}/\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)} = (2\pi)^{d/4}\sqrt{C_0}\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}$, where we used Eq. (8.38). Second term, $\|\widetilde{\chi}\|_{L^1(\mathbb{R}^d)} < \infty$, since $\|\widetilde{\chi}\|_{L^1(\mathbb{R}^d)} \leq C_1\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}$, via Eq. (8.39) (the constant C_1 is defined in the same equation) and we have seen already that $\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}$ is bounded. Third term, by a change of variable $\tau = \omega/t$,

$$\|\widetilde{w}_t\|_{L^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |\widetilde{w}_t(\omega)|d\omega = \int_{\mathbb{R}^d} |t|^{-d}|\tilde{w}(\omega/t)|d\omega = \int_{\mathbb{R}^d} |\tilde{w}(\tau)|d\tau = \|\tilde{w}\|_{L^1(\mathbb{R}^d)},$$

moreover $\|\tilde{w}\|_{L^1(\mathbb{R}^d)} \leq C_1\|w\|_{\mathcal{H}(\mathbb{R}^d)} = C_1\|u\|_{\mathcal{H}}$ via Eq. (8.39) and the fact that $\|w\|_{\mathcal{H}(\mathbb{R}^d)} = \|u\|_{\mathcal{H}}$ as recalled at the beginning of the proof. Finally, fourth term, for $t \in \mathbb{R} \setminus \{0\}$,

$$\begin{aligned}\|J_s\widetilde{w}_t\|^2_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \|\omega\|^{2s}|\widetilde{w_t}(\omega)|^2 \, d\omega = t^{-2d}\int_{\mathbb{R}^d} \|\omega\|^{2s}|\widetilde{w}(\omega/t)|^2 d\omega \\ &= t^{2s-d}\int_{\mathbb{R}^d} \|\tau\|^{2s}|\widetilde{w}(\tau)|^2 d\tau \leq t^{2s-d}\int_{\mathbb{R}^d} (1+\|\tau\|^2)^s|\widetilde{w}(\tau)|^2 d\tau \\ &= t^{2s-d}(2\pi)^{d/2}C_0\|w\|^2_{\mathcal{H}(\mathbb{R}^d)}.\end{aligned}$$

where we performed a change of variable $\omega = t\,\tau$, $t^d d\tau = d\omega$ and used the definition in Eq. (8.38) and the fact that $\|\tau\|^{2s} \leq (1+\|\tau\|^2)^s$ for any $\tau \in \mathbb{R}^d$. The proof of the bound of the fourth term is concluded by recalling that $\|w\|_{\mathcal{H}(\mathbb{R}^d)} = \|u\|_{\mathcal{H}}$ as discussed in the proof of the bound for the previous term.

Conclusion. Putting all our bounds together, we get :

$$\forall t \in \mathbb{R} \setminus \{0\}, \ \|h_t\|_{\mathcal{H}(\mathbb{R}^d)} \leq (A + B\ t^{s-d/2})\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}},$$

where $A = c_1C_2 + cc_1C_1(2\pi)^{d/4}\sqrt{C_0}$ and $B = cc_1C_1(2\pi)^{d/4}\sqrt{C_0}$, where $c = 2^{\max(s-1,0)}$, $c_1 = 2^{\max(s/2-1,0)}$, while C_1 is defined in Eq. (8.39), C_2 in Eq. (8.40). Now define

$$\forall x \in \mathbb{R}^d, \ \overline{g}_{z,r}(x) = \int_0^1 (1-t)h_t(x)dt,$$

and note that, by construction $\overline{g}_{z,r}(x) = \int_0^1 (1-t)u(tz+(1-t)x)dt$ for any $x \in B$ since u and χw coincide on B. Note that the map $t \in (0,1) \mapsto (1-t)\|h_t\|_{\mathcal{H}(\mathbb{R}^d)}$ is measurable, using the expression in Eq. (8.38). Moreover, since for all $t \in (0,1)$, it holds $\|h_t\|_{\mathcal{H}(\mathbb{R}^d)} \leq (A+Bt^{s-d/2})\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}} \leq (A+B)\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}}$ since $s > d/2$, the map $t \mapsto (1-t)h_t$ is in integrable, and thus

$$\|\overline{g}_{z,r}\|_{\mathcal{H}(\mathbb{R}^d)} = \Big\|\int_0^1 (1-t)h_t dt\Big\|_{\mathcal{H}(\mathbb{R}^d)} \leq \int_0^1 |1-t|\|h_t\|_{\mathcal{H}(\mathbb{R}^d)}dt \leq (A+B)\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}} < \infty,$$

which implies that the function $\overline{g}_{z,r}$ belongs to $\mathcal{H}(\mathbb{R}^d)$. Finally, denote by $R_\Omega : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}$ the restriction operator (see proposition 8.8 for more details). By construction $(R_\Omega g)(x) = g(x)$ for any $g \in \mathcal{H}(\mathbb{R}^d)$ and $x \in \Omega$, defining $g_{z,r} = R_\Omega \overline{g}_{z,r}$ the lemma is proven. □

8.E Proofs for algorithm 6

We start with two technical lemmas that will be used by the proofs in this section.

Lemma 8.12 (Technical result)**.** *Let $\alpha \geq 1$, $\beta \geq 2$ and $n \in \mathbb{N}$. If $n \geq 2\alpha\log(2\beta\alpha)$, then it holds*

$$\frac{\alpha\log(\beta n)}{n} \leq 1.$$

Proof. Note that the function $x \mapsto \frac{\log(\beta x)}{x}$ is strictly decreasing on $[\exp(1)/\beta, +\infty]$.

Moreover, $2\alpha\log(2\beta\alpha) \geq 2\log 4 \geq \exp(1)/2 \geq \exp(1)/\beta$ since $\beta \geq 2$ and $\alpha \geq 1$.

Now assume $n \geq c\alpha$ with $c = 2\log(2\beta\alpha)$. It holds:

$$\frac{\alpha\log(\beta n)}{n} \leq \frac{\log(\beta c\alpha)}{c} \leq \frac{\log(\frac{c}{2}) + \log(2\alpha\beta)}{c} \leq \frac{1}{2} + \frac{1}{2}\,\frac{2\log(2\beta\alpha)}{c} \leq 1,$$

where we used the definition of c and the fact that $\log(c/2) \leq c/2 - 1 \leq c/2$. □

Lemma 8.13. *Let $\overrightarrow{u} \in S_{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$, $\alpha \in [0, \pi/2]$, $x_0 \in \mathbb{R}^d$ and $t > 0$. Define the cone centered at x_0, directed by $\overrightarrow{u}$ of radius t with aperture α:*

$$C^{\alpha}_{x_0,\overrightarrow{u},t} = \left\{x \in B_t(x_0) \mid \tfrac{x-x_0}{\|x-x_0\|} \cdot \overrightarrow{u} \leq \cos(\alpha),\ x \neq x_0\right\},$$

where we denoted by $\cdot$ the scalar product among vectors. Then the volume of this cone is lower bounded as

$$\mathrm{vol}(C^{\alpha}_{x_0,\overrightarrow{u},t}) \geq \frac{(\sqrt{\pi}\sin(\alpha))^{d-1}(t\ \cos\alpha)^d}{d\Gamma((d+1)/2)}.$$

Moreover, let $x_0 \in \mathbb{R}^d$ and $r > 0$. Let $x \in B_r(x_0)$ and $0 < t \leq r$. The intersection $B_t(x) \cap B_r(x_0)$ contains the cone $C^{\pi/3}_{x,\overrightarrow{u},t}$, where $\overrightarrow{u} = \frac{x-x_0}{\|x-x_0\|}$ if $x \neq x_0$ and any unit vector otherwise.

Proof. **1. Bound on the volume of the cone**. Without loss of generality, assume $x_0 = 0$ and $\overrightarrow{u} = e_1$ since the Lebesgue measure is invariant by translations and rotations. A simple change of variable also shows that $\mathrm{vol}(C^{\alpha}_{0,\overrightarrow{u},t}) = t^d\,\mathrm{vol}(C^{\alpha}_{0,\overrightarrow{u},1})$. Now note the following inclusion (the proof is trivial):

$$\widetilde{C} := \left\{x = (x_1, z) \in \mathbb{R}^d = \mathbb{R}\times\mathbb{R}^{d-1}\ :\ z \leq \cos(\alpha),\ \|z\|_{\mathbb{R}^{d-1}} \leq x_1\sin(\alpha)\right\} \subset C^{\alpha}_{0,e_1,1}.$$

It is possible to compute the volume of the left hand term explicitly :

$$\begin{aligned}
\mathrm{vol}(\widetilde{C}) &= \int_{\mathbb{R}} \mathbf{1}_{x_1 \leq \cos(\alpha)} \left(\int_{\mathbb{R}^{d-1}} \mathbf{1}_{\|z\| \leq x_1\sin(\alpha)} dz\right) dx_1 \\
&= \int_0^{\cos(\alpha)} V_{d-1}(\sin\alpha x_1)^{d-1}\ dx \\
&= V_{d-1}\frac{\sin^{d-1}(\alpha)\cos^d(\alpha)}{d},
\end{aligned}$$

where $V_{d-1} = \pi^{(d-1)/2}/\Gamma((d-1)/2+1)$ denotes the volume of the $d-1$ dimensional ball.

2. Proof of the second point The case where $x = x_0$ is trivial since $t \leq r$. Assume therefore $x \neq x_0$ and note that by definition, $C^{\pi/3}_{x,\overrightarrow{u},t} \subset B_t(x)$. We will now show that $C^{\pi/3}_{x,\overrightarrow{u},t} \subset B_r(x_0)$. Let $y \in C^{\pi/3}_{x,\overrightarrow{u},t}$ and assume $y \neq x$ (if $y = x$ then $y \in B_r(x_0)$). Expanding the dot product

$$\begin{aligned}
\|y - x_0\|^2 &= \|y - x\|^2 + 2(y-x)\cdot(x - x_0) + \|x - x_0\|^2 \\
&= \|y - x\|^2 - 2\|y - x\|\ \|x_0 - x\|\ \tfrac{y-x}{\|y-x\|} \cdot \overrightarrow{u} + \|x - x_0\|^2 \\
&\leq \|x - y\|^2 - \|x - y\|\ \|x - x_0\| + \|x - x_0\|^2,
\end{aligned}$$

where the last inequality comes from the definition of the cone and $\cos\pi/3 = \frac{1}{2}$. Let us distinguish two cases:

- if $t > \|x_0 - x\|$, we have $-\|x-y\|\|x_0-x\| \leq -t^2$ and hence $\|y-x_0\|^2 \leq t^2 \leq r^2$;
- otherwise $\|x-y\| \leq t \leq \|x_0-x\|$ and thus $\|y-x_0\|^2 \leq \|x-x_0\|^2 \leq r^2$.

In any case, $y \in B_r(x_0)$, which concludes the proof. □

8.E .1 Proof of Lemma 8.4

Proof of Lemma 8.4. Fix Ω as in Lemma 8.4. Let U be the uniform probability over Ω, i.e., $U(A) = \frac{\mathrm{vol}(A \cap \Omega)}{\mathrm{vol}(\Omega)}$ for any Borel-measurable set A. Let $\mathbb{P} = U^{\otimes n}$ over Ω^n. Throughout this proof, we will use the notation V_d to denote the volume of the d-dimensional unit ball (recall that $V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$).

Step 1. Covering Ω. Let $t > 0$. We say that a subset $\overline{X}$ of Ω is a t (interior) covering of Ω if $\Omega \subset \bigcup_{x \in \overline{X}} B_t(x)$. Denote with N_t the minimal cardinal $|\overline{X}|$ of a t interior covering of Ω and fix $\overline{X}_t$ a t interior covering of Ω whose cardinal is minimum, i.e., $|\overline{X}_t| = N_t$. Since the diameter of Ω is bounded by $2R$, it is known that $N_t \leq (1 + 2R/t)^d$

To prove this fact , one defines a maximal $t/2$-packing of Ω as a maximal set $\overline{Y}_{t/2} \subset \Omega$ such that the balls $B_{t/2}(\overline{y})$ are disjoint. It is then easy to check that if $\overline{Y}_{t/2}$ is a maximal $t/2$-packing, then it is also a t-covering and hence $N_t \leq |\overline{Y}_{t/2}|$. Finally, since Ω is included in a ball of radius $B_{2R}(x_0)$ for some $x_0 \in \mathbb{R}^d$ and since $\overline{Y}_{t/2} \subset \Omega$, it holds $\bigcup_{\overline{y} \in \overline{Y}_{t'}} B_t(\overline{y}) \subset B_{R+t/2}(x_0)$. Since the $B_t(\overline{y})$ are two by two disjoint, the result follows from the following equation:

$$|\overline{Y}|_{t/2}(t/2)^d V_d = \mathrm{vol}\left(\cup_{\overline{y} \in \overline{Y}_{t'}} B_t(\overline{y})\right) \leq \mathrm{vol}(B_{R+t/2}(x_0)) = (R + t/2)^d V_d.$$

Step 2. Probabilistic analysis. Note that for any $(x_1, ..., x_n) \in \Omega^n$, writing $\widehat{X} = \{x_1, .., x_n\}$, it holds:

$$\begin{aligned} h_{\widehat{X},\Omega} &= \max_{x \in \Omega} \min_{i \in [n]} \|x - x_i\| = \max_{\overline{x} \in \overline{X}_t} \max_{x \in B_t(\overline{x}) \cap \Omega} \min_{i \in [n]} \|x - x_i\| \\ &\leq t + \max_{\overline{x} \in \overline{X}_t} \min_{i \in [n]} \|\overline{x} - x_i\|. \end{aligned}$$

Define E to be the following event :

$$E = \{(x_1, \ldots, x_n) \in \Omega^n \mid \max_{j \in [m]} \min_{i \in [n]} \|\overline{x}_j - x_i\| < t\}.$$

The n tuple $(x_1, .., x_n)$ belongs to E if for each $\overline{x} \in \overline{X}_t$ there exists at least one $i \in [n]$ for which $\|\overline{x} - x_i\| < t$. E can therefore be rewritten as follows :

$$E = \bigcap_{\overline{x} \in \overline{X}_t} \bigcup_{i \in [n]} \{(x_1, \ldots, x_n) \in \Omega^n \mid \|\overline{x} - x_i\| < t\}.$$

In particular, note that

$$E^c = \Omega^n \setminus E = \bigcup_{\overline{x} \in \overline{X}_t} \bigcap_{i \in [n]} \{(x_1, \ldots, x_n) \in \Omega^n \mid \|\overline{x} - x_i\| \geq t\} = \bigcup_{\overline{x} \in \overline{X}_t} (\Omega \setminus B_t(\overline{x}))^n.$$

Applying a union bound, we get

$$\begin{aligned} \mathbb{P}(E^c) &= \mathbb{P}\Big(\bigcup_{\overline{x} \in \overline{X}_t} (\Omega \setminus B_t(\overline{x}))^n \Big) \\ &\leq \sum_{\overline{x} \in \overline{X}_t} \mathbb{P}\big((\Omega \setminus B_t(\overline{x}))^n\big) = \sum_{j \in [m]} U(\Omega \setminus B_t(\overline{x})))^n, \end{aligned}$$

where the last step is due to the fact that $\mathbb{P}$ is a product measure and so $\mathbb{P}(A^n) = U^{\otimes n}(A^n) = U(A)^n$. Now we need to evaluate $U(\Omega \setminus B_t(\overline{x})) = 1 - U(B_t(\overline{x}))$ for $\overline{x} \in \overline{X_t}$. Since $\overline{X}_t \subset \Omega$, it holds

$$\forall \overline{x} \in \overline{X}_t,\ U(B_t(\overline{x})) = \tfrac{\mathrm{vol}(B_t(\overline{x}) \cap \Omega)}{\mathrm{vol}(\Omega)} \geq \tfrac{\min_{x \in \Omega} \mathrm{vol}(B_t(x) \cap \Omega)}{\mathrm{vol}(\Omega)}.$$

Step 3. Bounding $\mathrm{vol}(B_t(x) \cap \Omega)$ when $t \leq r$. Let us now find a lower bound for $\min_{x \in \Omega} \mathrm{vol}(B_t(x) \cap \Omega)$. Recall that since Ω satisfies Assumption 8.1(a), Ω can be written $\Omega = \cup_{z \in S} B_r(z)$. .Let $t \leq r$, $x \in \Omega$. By the previous point, there exists $z \in S$ such that $x \in B_r(z) \subset \Omega$ and hence $B_t(x) \cap B_r(z) \subset B_t(x) \cap \Omega$. Let $C_{x,z,t}$ denote the cone centered in x and directed to z with aperture $\pi/3$. It is easy to see geometrically that $B_r(z) \cap B_t(x)$ contains the cone $C_{x,z,t}$ (this fact is proved in Lemma 8.13). Moreover, using the lower bound for the volume of this cone provided in Lemma 8.13, it holds:

$$\begin{aligned}\mathrm{vol}(\Omega \cap B_t(x)) &\geq \mathrm{vol}(B_r(z) \cap B_t(x)) \geq \mathrm{vol}(C_{x,z,t}) \\ &\geq \frac{2V_{d-1}}{\sqrt{3}d} \left(\tfrac{\sqrt{3}}{4}\right)^d t^d.\end{aligned}$$

Step 4. Expressing t with respect to n and δ and guaranteeing that $t \leq r$. To conclude, let $C = \frac{V_{d-1}}{2d\,\mathrm{vol}(\Omega)} \left(\frac{\sqrt{3}}{4}\right)^{d-1}$. Since $N_t \leq (1 + 2R/t)^d$, and $(1-c)^x \leq e^{-cx}$ for any $x \geq 0$ and $c \in [0,1]$, then

$$\mathbb{P}(E) \geq 1 - N_t\big(1 - Ct^d\big)^n \geq 1 - e^{-Ct^d n + d\log(1+2R/t)} \geq 1 - \delta,$$

where the last step is obtained by setting

$$t = (Cn)^{-1/d} \left(\log \frac{(1 + 2R(Cn)^{1/d})^d}{\delta}\right)^{1/d}.$$

Then $h_{\widehat{X},\Omega} \leq 2t$ with probability at least $1 - \delta$, when $t \leq r$. The desired result is obtained by further bounding C and t as follows.

Bounding C. It holds $\frac{2V_{d-1}}{\sqrt{3}dV_d} = \left(\frac{4}{3d^2\pi}\right)^{1/2} \frac{\Gamma(d/2+1)}{\Gamma(d/2+1/2)}$. Using Gautschi's inequality and the fact that $d \geq 1$,

$$\left(\tfrac{2}{3d\pi}\right)^{1/2} \leq \frac{2V_{d-1}}{\sqrt{3}dV_d} \leq \left(\tfrac{2(d+2)}{3d^2\pi}\right)^{1/2} \leq 1.$$

Since $\left(\frac{3d\pi}{2}\right)^{1/2d} \frac{4}{\sqrt{3}} \leq 2\sqrt{2\pi}$ for all $d \geq 1$, and since $V_d r^d \leq \mathrm{vol}(\Omega) \leq V_d R^d$, it holds

$$(2\sqrt{2\pi}R)^{-d} \leq C \leq (4r/\sqrt{3})^{-d} \implies \frac{n^{1/d}}{2\sqrt{2\pi}R} \leq (Cn)^{1/d} \leq \frac{\sqrt{3}n^{1/d}}{4r} \leq \frac{n^{1/d}}{2r}.$$

Bounding t. Since, $(1+x)^d \leq (2x)^d$ for any $x \geq 1$ and $2R(Cn)^{1/d} \leq \frac{R}{r} n^{1/d}$, and $\frac{R}{r} n^{1/d} \geq 1$, it holds

$$t \leq 2\sqrt{2\pi}Rn^{-1/d}(\log \tfrac{n}{\delta} + d \log \tfrac{2R}{r})^{1/d}.$$

Guaranteeing $t \leq r$. Applying Lemma 8.12 to $\alpha = (2\pi)^{d/2}(2R/r)^d$ and $\beta = (2R/r)^d/\delta$, it holds that if

$$n \geq 2\alpha \log(2\alpha\beta) = 2\ (2\pi)^{d/2}(2R/r)^d \left(\log \frac{2}{\delta} + d/2 \log(2\pi) + 2d \log(2R/r)\right),$$

then $\alpha/n \log(\beta n) \leq 1$, so

$$t \leq 2\sqrt{2\pi} R n^{-1/d} (\log \tfrac{n}{\delta} + d \log \tfrac{2R}{r})^{1/d} \leq r(\alpha/n \log(\beta n)^{1/d} \leq r.$$

□

8.E .2 Proof of Theorem 8.6

Proof. Recall that $s > d/2$ and $m < s - \frac{d}{2}$ is a positive integer. Assume that Ω satisfies Assumption 8.1(a) for a certain r and that the diameter of Ω is bounded by $2R$. In particular, if Ω is a ball of radius R, then Ω satisfies Assumption 8.1(a) with $r = R$. In the first step of the proof we guarantee that n is large enough to apply Lemma 8.4 and that $h_{\widehat{X},\Omega}$, controlled by Lemma 8.4, satisfies the assumptions of Theorem 8.5. Then we apply Theorem 8.5.

Step 1. Guaranteeing n large enough and $h_{\widehat{X},\Omega} \leq r/(18(m-1)^2)$. Applying Lemma 8.12 to $\alpha = \left(\frac{2R}{r}\right)^d \max(3, 10(m-1))^{2d}$ and $\beta = \frac{(2R)^d}{r^d\ \delta}$, it holds that if

$$n \geq 2\alpha \log(2\alpha\beta) = \left(\frac{2R}{r}\right)^d \max(3, 10(m-1))^{2d} \left(2 \log \frac{2}{\delta} + 4d \log\left(\tfrac{R}{r} \max(6, 20(m-1))\right)\right),$$

then $\alpha/n \log(\beta n) \leq 1$, which implies

$$n^{-1/d} (\log \frac{n}{\delta} + d \log \beta)^{1/d} \leq \frac{r}{2R \max(3, 10(m-1))^2}.$$

In particular, n satisfying the condition above is large enough to satisfy the requirement of Lemma 8.4 (since $r \leq R$). Therefore, by applying Lemma 8.4 we have that with probability at least $1 - \delta$,

$$h_{\widehat{X},\Omega} \ \leq \ 11R\, n^{-\frac{1}{d}} (\log \tfrac{(2R)^d\ n}{r^d\ \delta})^{1/d} \leq \frac{r}{\max(1, 18(m-1)^2)}.$$

Step 2. Applying Theorem 8.5. In the previous step we provided a condition on n such that $h_{\widehat{X},\Omega}$ satisfies $h_{\widehat{X},\Omega} \leq \frac{r}{\max(1,18(m-1)^2)}$. By proposition 8.1, Assumption 8.2 holds for the Sobolev kernel with smoothness s, for any $m \in \mathbb{N}$ since $m < s - d/2$. Then the conditions to apply Theorem 8.5 are satisfied. Applying Theorem 8.5 with $\lambda \geq 2\eta \max(1, \mathsf{MD}_m)$ and $\eta = \frac{3 \max(1,18(m-1)^2)^m\ d^m}{m!} h^m_{\widehat{X},\Omega}$, we have

$$|\hat{c} - f_*| \leq 2\eta |f|_{\Omega,m} + \lambda \operatorname{Tr}(A_*) \leq 3\lambda(|f|_{\Omega,m} + \operatorname{Tr}(A_*)),$$

Thus, under this condition, we have with probability at least $1 - \delta$,

$$|\hat{c} - f_*| \leq C_{m,s,d} R^m n^{-m/d} (\log \frac{2^d n}{\delta}),$$

where

$$C_{m,s,d} = 6 \times 11^m \times \frac{\max(1, 18(m-1)^2)^m d^m}{m!} \max(1, \mathsf{MD}_m).$$

Step 3. Bounding the constant term $C_{m,s,d}$ in terms of m, s, d. Note that

$$\frac{\Gamma(m + d/2)}{\Gamma(d/2)} = (d/2)...(d/2 + m - 1) \leq (d/2 + m - 1)^{m-1}$$

and

$$\frac{\Gamma(s-d/2-m)}{\Gamma(s-d/2)} = \frac{1}{(s-d/2-m)....(s-d/2-1)} \leq \left(\frac{1}{s-d/2-m}\right)^{m-1},$$

which yields:

$$\mathsf{D}_m \leq (2\pi)^{d/4} \left(\frac{d/2+m-1}{s-d/2-m}\right)^{(m-1)/2}.$$

Moreover, using the bound on M, we get

$$\mathsf{D}_m\mathsf{M} \leq 2^{s+1/2}\ (2\pi)^{3d/4} \left(\frac{d/2+m-1}{s-d/2-m}\right)^{(m-1)/2}.$$

This yields the following bound for $C_{m,s,d}$:

$$C_{m,s,d} \leq \frac{6\max(1,18(m-1)^2)^m(11d)^m}{m!} \max\left(1, 2^{s+1/2}\ (2\pi)^{3d/4} \left(\frac{d/2+m-1}{s-d/2-m}\right)^{(m-1)/2}\right).$$

□

8.F Global minimizer. Proofs.

8.F .1 Proof of Remark 27

Proof. Since f satisfies both Assumptions 8.1(b) and 8.4, denote by ζ the unique minimizer of f in Ω. Since ζ is a strict minimum by Assumption 8.1(b), there exists $\beta_1 > 0$ such that $\nabla^2 f(\zeta) \succeq \beta_1 I$. Thus, since $f \in C^2(\mathbb{R}^d)$, there exists a small radius $t > 0$ such that $\nabla^2 f(x) \succeq \frac{\beta_1}{2} I$ for all $x \in B_t(\zeta)$ and hence

$$\forall x \in \Omega \cap B_t(\zeta),\ f(x) - f_* = f(x) - f(\zeta) - \nabla f(\zeta) \geq \tfrac{\beta_1}{4}\|x-\zeta\|^2. \tag{8.41}$$

Moreover, since f has no minimizer on the boundary of Ω and since ζ is the unique minimizer of f on Ω, f has no minimizer on $K = \overline{\Omega} \setminus B_t(x)$ which is a compact set. Denote by m the minimum of f on K. Since K is compact, this minimum is reached and since f does not reach its global minimum f_* on K, we have $m - f_* > 0$. Let R be a radius such that $\overline{\Omega} \subset B_R(\zeta)$, which exists since Ω is bounded. Then, since for any $x \in \overline{\Omega}$, $\|x-\zeta\| < R$, it holds for any $x \in K$:

$$f(x) - f_* = f(x) - m + m - f_* \geq m - f_* = \frac{2(m-f_*)}{2R^2}R^2 \geq \frac{2(m-f_*)}{2R^2}\|x-\zeta\|^2. \tag{8.42}$$

Thus, taking $\beta = \min(\frac{\beta_1}{2}, \frac{2(m-f_*)}{R^2})$ and combining Eqs. (8.41) and (8.42), it holds

$$\forall x \in \Omega,\ f(x) - f_* \geq \frac{\beta}{2}\|x-\zeta\|^2.$$

□

8.F .2 Proof of Theorem 8.7

Proof. Let us divide the proof into four steps.

Step 1: Extending the parabola outside of Ω Since Ω is an open set containing ζ, there exists $t > 0$ such that the ball $B_t(\zeta) \subset \Omega$. Define $\delta = \frac{\beta-\nu}{2}t^2$. It holds :

$$\forall x \in \mathbb{R}^d \setminus \Omega,\ \frac{\beta}{2}\|x-\zeta\|^2 \geq \frac{\nu}{2}\|x-\zeta\|^2 + \delta. \tag{8.43}$$

Now define the following open set :

$$\widetilde{\Omega} = \left\{ x \in \mathbb{R}^d \ : \ f(x) - f_* - \tfrac{\beta}{2}\|x-\zeta\|^2 > -\delta/2 \right\}.$$

It is open since f is continuous. Moreover, it contains the closure of Ω which we denote with $\overline{\Omega}$ which is compact since it is closed and bounded in $\mathbb{R}^d$. Theorem 1.4.2 in the work by Hörmander (2015) applied to $X = \widetilde{\Omega}$ and $K = \overline{\Omega}$ shows the existence of $\chi : \mathbb{R}^d \to \mathbb{R}$ such that $\chi \in C^\infty(\mathbb{R}^d)$, $\chi(x) \in [0,1]$, $\chi = 1$ on $\overline{\Omega}$ and $\chi = 0$ on $\mathbb{R}^d \setminus \widetilde{\Omega}$. Finally, define $\overline{p}_\nu(x) := \frac{\nu}{2}\|x-\zeta\|^2\chi(x)$. $\overline{p}_\nu$ satisfies the following properties :

- $\overline{p}_\nu \in C^\infty(\mathbb{R}^d)$;
- for all $x \in \overline{\Omega}$, $\overline{p}_\nu(x) = \frac{\nu}{2}\|x-\zeta\|^2 \leq \frac{\beta}{2}\|x-\zeta\|^2$;
- for all $x \in \mathbb{R}^d \setminus \widetilde{\Omega}$, $\overline{p}_\nu(x) = 0$;
- for all $x \in \widetilde{\Omega} \setminus \Omega$, $f(x) - f_* - \overline{p}_\nu(x) \geq \delta/2$.

The first, second and third properties are direct consequences of the properties of χ and the fact that $\nu < \beta$. The last property comes from combining Eq. (8.43) with the definition of $\widetilde{\Omega}$ and the fact that $\chi \in [0,1]$:

$$\begin{aligned}
\forall x \in \widetilde{\Omega} \setminus \Omega, \ f(x) - f_* - \overline{p}_\nu(x) &= f(x) - f_* - \chi(x)\tfrac{\nu}{2}\|x-\zeta\|^2 \\
&\geq f(x) - f_* - \tfrac{\nu}{2}\|x-\zeta\|^2 \\
&= \left(f(x) - f_* - \tfrac{\beta}{2}\|x-\zeta\|^2\right) + \left(\tfrac{\beta}{2}\|x-\zeta\|^2 - \tfrac{\nu}{2}\|x-\zeta\|^2\right) \\
&\geq -\delta/2 + \delta = \delta/2.
\end{aligned}$$

Step 2: Extending $x \mapsto f(x) - \frac{\nu}{2}\|x-\zeta\|^2$ outside of Ω. Define $g(x) = f(x) - \overline{p}_\nu(x)$ on $\mathbb{R}^d$. Then g satisfies Assumption 8.1(b), g has exactly one minimizer in Ω which is ζ, and its minimum is $g(\zeta) = f_*$. Indeed, the fact that $g \in C^2(\mathbb{R}^d)$ comes from the fact that $f \in C^2(\mathbb{R}^d)$ by Assumption 8.1(b) on f and the fact that $\overline{p}_\nu \in C^\infty(\mathbb{R}^d)$. Moreover, $g \geq f_*$ on $\mathbb{R}^d$ and $g - f_* \geq \delta/2$ on $\partial\Omega$. Indeed, first note that since $\nu < \beta$, it holds

$$\forall x \in \Omega, \ g(x) = f(x) - \overline{p}_\nu(x) = f(x) - \tfrac{\nu}{2}\|x-\zeta\|^2 \geq f(x) - \tfrac{\beta}{2}\|x-\zeta\|^2 \geq f_*,$$

where the last inequality comes from Eq. (8.21). Second, since $\overline{p}_\nu = 0$ on $\mathbb{R}^d \setminus \widetilde{\Omega}$ and since f_* is the minimum of f, for any $x \in \mathbb{R}^d \setminus \widetilde{\Omega}$, $g(x) - f_* = f(x) - f_* \geq 0$. Finally, by the last point of the previous step, we see that $g(x) \geq f_* + \delta/2 > f_*$ for any $x \in \widetilde{\Omega} \setminus \Omega$. In particular, $g(x) \geq f_* + \delta/2$ for any $x \in \partial\Omega$. Since $g(\zeta) = f(\zeta) = f_*$, we see that f_* is the minimum of g on $\mathbb{R}^d$ and that this minimum is reached at ζ and is not reached on the boundary of Ω. The fact that ζ is the unique minimum on Ω comes from the fact that since $\nu < \beta$ and by Eq. (8.21) we have that for any $x \in \Omega \setminus \{\zeta\}$ the following holds

$$\begin{aligned}
g(x) = f(x) - \overline{p}_\nu(x) &= f(x) - \tfrac{\nu}{2}\|x-\zeta\|^2 \\
&> f(x) - \tfrac{\beta}{2}\|x-\zeta\|^2 \geq f_*.
\end{aligned} \tag{8.44}$$

The fact that this minimum is not reached on the boundary of Ω comes from the fact stated above that $g(x) \geq f_* + \delta/2$ for any $x \in \partial\Omega$. Finally, the fact that ζ is a strict minimum of g also comes from Eq. (8.44) which implies that $\nabla^2 g(\zeta) \succeq (\beta - \nu)I$ since g reaches a minimum in ζ, g is C^2 and $\nu < \beta$.

Note that g also satisfies Assumption 8.3 since f satisfies Assumption 8.3 and $\overline{p}_\mu \in C^\infty(\mathbb{R}^d) \subset C^2(\mathbb{R}^d) \cap \mathcal{H}$ by Assumption 8.2(a).

Step 3: Applying Cor. 8.1 to g. The previous point shows that g satisfies Assumptions 8.1(b) and 8.3and that g has a unique minimum in Ω. Moreover, $\mathcal{H}$ satisfies Assumption 8.2. Hence, Cor. 8.1 to g and $\mathcal{H}$, the following holds : there exists $A_* \in \mathbb{S}_+(\mathcal{H})$ with $\operatorname{rank}(A_*) \leq d+1$ such that $g(x) - f^* = \langle \phi(x), A_* \phi(x) \rangle$ for all $x \in \Omega$.

Step 4. Let p_0 be the maximum of Eq. (8.20). In Lemma 8.5 we have seen that the solution of Eq. (8.20) is $p_0 = f_*$. Since $A \succeq 0$ implies $\langle \phi(x), A\phi(x) \rangle \geq 0$ for all $x \in \Omega$, the problem in Eq. (8.20) is a relaxation of Eq. (8.22), where the constraint $f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top z - c = \langle \phi(x), A\phi(x) \rangle$ is substituted by $f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top z - c \geq 0, \forall x \in \Omega$. Then $p_0 \geq p^*$ if a maximum p^* exists for Eq. (8.22). Thus, if there exists A that satisfies the constraints in Eq. (8.22) for the value $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$ and $z_* = \zeta$, then $p_0 = p^*$ and (c_*, ζ, A) is a minimizer of Eq. (8.22).

The proof is concluded by noting that indeed there exists A that satisfies the constraints in Eq. (8.22) for the value $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$ and $z_* = \zeta$ and it is obtained by the previous step.

□

8.F .3 Proof of Theorem 8.8

Proof. The proof is a variation of the the one for Theorem 8.5, the main difference is that we take care of the additional term $z - \zeta$.

Step 0. The SDP problem in Eq. (8.23) admits a solution

(a) Under the constraints of Eq. (8.23), $c - \frac{\nu}{2}\|z\|^2$ cannot be larger than $\min_{i\in[n]} f(x_i)$. Indeed, for any $i \in [n]$, since $B \succeq 0$, the i-th constraint implies

$$f(x_i) - \tfrac{\nu}{2}\|x_i - z\|^2 - c + \frac{\nu}{2}\|z\|^2 = f(x_i) - \tfrac{\nu}{2}\|x_i\|^2 + \nu x_i^\top z - c = \Phi_i B \Phi_i \geq 0.$$

Hence, $f(x_i) \geq f(x_i) - \frac{\nu}{2}\|x_i - z\|^2 \geq c + \frac{\nu}{2}\|z\|^2$. Thus, since $B \succeq 0$, for any B, z, c satisfying the constraint, $c - \frac{\nu}{2}\|z\|^2 - \lambda \operatorname{Tr}(B) \leq \max_{i\in[i]} f(x_i)$.

(b) There exists an admissible point. Indeed let (c_*, z_*, A_*) be the solution of Eq. (8.22) such that A_* has minimum trace norm (by Theorem 8.7, we know that this solution exists with $c_* = f_*$ and $z_* = \zeta$, under Assumptions 8.1 to 8.4). Then, by Lemma 8.3 applied to $g(x) = f(x) - \frac{\nu}{2}\|x\|^2 - \nu x^\top z_* - c_*$ and $A = A_*$, given $\widehat{X} = \{x_1, \ldots, x_n\}$ we know that there exists $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$ satisfying $\operatorname{Tr}(\overline{B}) \leq \operatorname{Tr}(A_*)$ s.t. the constraints of Eq. (8.23) are satisfied for $c = c_*$ and $z = z_*$. Then $(c_*, z_*, \overline{B})$ is admissible for the problem in Eq. (8.23). Since there exists an admissible point for the constraints of Eq. (8.23) and its functional cannot be larger than $\max_{i\in[n]} f(x_i)$, then the SDP problem in Eq. (8.23) admits a solution (Boyd and Vandenberghe, 2004).

Step 1. Consequences of existence of A_*. Let $(\hat{c}, \hat{z}, \hat{B})$ one minimizer of Eq. (8.23). The existence of the admissible point $(c_*, z_*, \overline{B})$ implies that

$$\hat{c} - \tfrac{\nu}{2}\|\hat{z}\|^2 - \lambda \operatorname{Tr}(\hat{B}) \geq c_* - \tfrac{\nu}{2}\|z_*\|^2 - \lambda \operatorname{Tr}(\overline{B}) \geq f_* - \lambda \operatorname{Tr}(A_*). \tag{8.45}$$

From which we derive,

$$\lambda \operatorname{Tr}(\hat{B}) - \lambda \operatorname{Tr}(A_*) \;\leq\; \Delta, \quad \Delta := \hat{c} - \tfrac{\nu}{2}\|\hat{z}\|^2 - f_*. \tag{8.46}$$

Step 2. L^∞ bound due to the scattered zeros. Note that the solution $(\hat{c}, \hat{z}, \hat{B})$ satisfies $\hat{g}(x_i) = \Phi_i^\top \hat{B} \Phi_i$ for $i \in [n]$, where the function $\hat{g}$ is defined as $\hat{g}(x) = f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top \hat{z} - \hat{c}$ for $x \in \Omega$, moreover $h_{\widehat{X},\Omega} \leq \frac{r}{\max(1,18(m-1)^2)} = \frac{r}{18(m-1)^2}$ by assumption, since $m \geq 2$. Then we can apply Theorem 8.4 with $g = \hat{g}$, $\tau = 0$ and $B = \hat{B}$ obtaining for all $x \in \Omega$

$$f(x) - \tfrac{\nu}{2}\|x\|^2 + \nu x^\top \hat{z} - \hat{c} = \hat{g}(x) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \operatorname{Tr}(\hat{B})), \quad \eta = C_0 h^m_{\widehat{X},\Omega},$$

where C_0 is defined in Theorem 8.4 and $C_0 = 3\frac{(18d)^m (m-1)^{2m}}{m!}$ since $m \geq 2$. Since the inequality above holds for any $x \in \Omega$, by evaluating it in the global minimizer $\zeta \in \Omega$, we have $f(\zeta) = f_*$ and so

$$-\Delta - \tfrac{\nu}{2}\|\hat{z} - \zeta\|^2 = \hat{g}(\zeta) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \operatorname{Tr}(\hat{B})).$$

Now we bound $|\hat{g}|_{\Omega,m}$. Since $\hat{g}(x) = f(x) - p_{\hat{z},\hat{c}}(x)$, where $p_{\hat{z},\hat{c}}$ is a second degree polynomials defined as $p_{\hat{z},\hat{c}} = \frac{\nu}{2}\|x\|^2 - \nu x^\top \hat{z} + \hat{c}$, we have

$$|\hat{g}|_{\Omega,m} \leq |f|_{\Omega,m} + |p_{\hat{z},\hat{c}}|_{\Omega,m} \leq |f|_{\Omega,m} + \nu, \tag{8.47}$$

since for $m = 2$, we have $|p_{\hat{z},\hat{c}}|_{\Omega,2} = \sup_{i,j\in[d],x\in\Omega} |\frac{\partial^2 p_{\hat{z},\hat{c}}(x)}{\partial x_i \partial x_j}| = \nu$ and also $|p_{\hat{z},\hat{c}}|_{\Omega,m} = 0$ for $m > 2$. Then

$$\Delta \leq \Delta + \tfrac{\nu}{2}\|\hat{z} - \zeta\|^2 \leq \eta |f|_{\Omega,m} + \eta \mathsf{MD}_m \operatorname{Tr}(\hat{B}) + \eta\nu. \tag{8.48}$$

Conclusion. Combining Eq. (8.48) with Eq. (8.46), since $\frac{\nu}{2}\|\hat{z} - \zeta\|^2 \geq 0$ and since $\lambda \geq 2\mathsf{MD}_m\eta$ by assumption, we have

$$\tfrac{\lambda}{2} \operatorname{Tr}(\hat{B}) \leq (\lambda - \mathsf{MD}_m\eta) \operatorname{Tr}(\hat{B}) \leq \eta |f|_{\Omega,m} + \eta\nu + \lambda \operatorname{Tr}(A_*),$$

from which we obtain Eq. (8.26). Moreover, the inequality Eq. (8.25) is derived by bounding Δ from below as $\Delta \geq -\lambda \operatorname{Tr}(A_*)$ by Eq. (8.46), since $\operatorname{Tr}(\hat{B}) \geq 0$ by construction, and bounding it from above as

$$\Delta \leq 2\eta |f|_{\Omega,m} + 2\eta\nu + \lambda \operatorname{Tr}(A_*),$$

that is obtained by combining Eq. (8.48) with Eq. (8.26) and with the assumption $\mathsf{MD}_m\eta \leq \lambda/2$. Finally from Eq. (8.48) we obtain

$$\tfrac{\nu}{2}\|\hat{z} - \zeta\|^2 \leq |\Delta| + \eta |f|_{\Omega,m} + \eta \mathsf{MD}_m \operatorname{Tr}(\hat{B}) + \eta\nu,$$

from which we derive the bound $\frac{\nu}{2}\|\hat{z} - \zeta\|^2$ in Eq. (8.24), by bounding $|\Delta|$ and $\operatorname{Tr}(\hat{B})$ via Eq. (8.25) and Eq. (8.26). □

8.G Proofs for the extensions

8.G .1 Proof of Theorem 8.9

Proof. Let $(\hat{c}, \hat{B})$ be a minimum trace-norm solution of Eq. (8.5). The minimum $p_{\lambda,n}$ of Eq. (8.5) then corresponds to $p_{\lambda,n} = \hat{c} - \lambda \operatorname{Tr}(\hat{B})$. Combining Eq. (8.27) with Eq. (8.17) from the proof of Theorem 8.5 and the fact that $\theta_2 \leq \lambda/8$, we have that

$$\tfrac{7}{8}\lambda \operatorname{Tr}(\tilde{B}) - \lambda \operatorname{Tr}(A_*) - \theta_1 \leq \tilde{\Delta}, \quad \tilde{\Delta} := \tilde{c} - f_*. \tag{8.49}$$

Analogously to Step 3 of the proof of Theorem 8.5, by applying Theorem 8.4 to Eq. (8.28) with $g(x) = f(x) - \tilde{c}, B = \tilde{B}$ and $\tau = \tau_1 + \tau_2 \operatorname{Tr}(\tilde{B})$, we obtain for any $x \in \Omega$

$$f(x) - \tilde{c} \;\geq\; -2\tau_1 - 2\tau_2 \operatorname{Tr}(\tilde{B}) \;-\; \eta(|g|_{\Omega,m} + \mathsf{MD}_m \operatorname{Tr}(\tilde{B})), \qquad \eta = C_0 h^m_{\widehat{X},\Omega},$$

with C_0 defined in Theorem 8.4. Now evaluating the inequality above for $x = \zeta$, noting that $|g|_{\Omega,m} = |f|_{\Omega,m}$ since $m \geq 1$, and considering that by assumption $\tau_2 \leq \lambda/8$ and $\mathsf{MD}_m\eta \leq \lambda/2$ we derive

$$\tilde{\Delta} = -(f(\zeta) - \tilde{c}) \leq 2\tau_1 + \tfrac{3}{4}\lambda \operatorname{Tr}(\tilde{B}) + \eta|f|_{\Omega,m}. \tag{8.50}$$

The desired result is obtained by combining Eq. (8.50) and Eq. (8.49) as we did in Step 3 of Theorem 8.5. □

8.G .2 Proof of Cor. 8.2

Proof. Define $\mathcal{H} = \{g \in C^s(\Omega) \ : \ \exists f \in C^s(\mathbb{R}^d),\ f|_\Omega = g\}$, endowed with the following norm :

$$\forall g \in \mathcal{H},\ \|g\|_{\mathcal{H}} = \sup_{|\alpha|\leq s} \sup_{x\in\Omega} \|\partial^\alpha g(x)\|.$$

Note that this norm is well defined since for any $g \in \mathcal{H}$, since there exists $f \in C^s(\mathbb{R}^d)$ such that $g = f|_\Omega$, since all the derivatives of f are continuous hence bounded on Ω which is bounded, so are all the derivatives of g.

Now note that $\mathcal{H}$ satisfies Assumptions 8.2(a) to 8.2(c). Indeed, given $u, v \in \mathcal{H}$ the first assumption is satisfied as a simple consequence of the Leibniz formula, since for any $x \in \Omega$, $\partial^\alpha(u \cdot v)(x) = \sum_{\beta\leq\alpha} \binom{\alpha}{\beta} \partial^\beta u(x)\partial^{\alpha-\beta}v(x)$ which in turn implies that for any $|\alpha| \leq s$ and $x \in \Omega$, $\|\partial^\alpha(u \cdot v)(x)\| \leq 2^{|\alpha|}\ \|u\|_{\mathcal{H}}\ \|v\|_{\mathcal{H}}$ and hence $\|u \cdot v\|_{\mathcal{H}} \leq 2^s\|u\|_{\mathcal{H}}\ \|v\|_{\mathcal{H}}$. Assumption 8.2(b) is trivially satisfied and Assumption 8.2(c) is a simple consequence of the dominated convergence theorem. Indeed, if $u \in \mathcal{H}$ and $\overline{u} \in C^s(\mathbb{R}^d)$ such that $\overline{u}|_\Omega = u$, define

$$\forall x, z \in \mathbb{R}^d,\ \overline{v}_z(x) = \int_0^1 (1-t)\overline{u}(z + t(x-z))dt.$$

$\overline{v}_z$ is in $C^s(\mathbb{R}^d)$ by dominated convergence, and $v_z = \overline{v}|_\Omega$ satisfies the desired property (in this case, there is no need to depend on r and one can simply take $g_{r,z} = v_z$).

Moreover, if $f \in C^{s+2}(\mathbb{R}^d)$, then in particular, for any $i, j \in [d]$, $\frac{\partial f}{\partial x_i \partial x_j} \in C^s(\mathbb{R}^d)$ and hence its restriction to Ω is in $\mathcal{H}$. Moreover, in that case, it is obvious that since $s \geq 0$, $f|_\Omega \in \mathcal{H}$. This shows that f satisfies Assumptions 8.1(b) and 8.3.

Therefore, Theorem 8.2 can be applied, and there exist $\tilde{w}_1, \ldots, \tilde{w}_p\mathcal{H}$, $p \in \mathbb{N}_+$, such that

$$\forall x \in \Omega,\ f(x) - f_* = \sum_{j\in[p]} w_j^2(x).$$

By definition of $\mathcal{H}$, taking $w_1, ..., w_p$ such that $w_j|_\Omega = \tilde{w}_j$, the corollary holds.

□

8.G .3 Proof of Theorem 8.10

Proof. In this proof we will use the results recalled in Sec. 8.A .2 about Sobolev spaces. By Cor. 8.2 we have that there exists $\overline{w}_1, \ldots, \overline{w}_p \in C^s(\mathbb{R}^d)$ such that $f(x) = \sum_{j\in[p]} \overline{w}_j^2(x)$ for any $x \in \Omega$. Define $w_j = \overline{w}_j|_\Omega$. Note that by proposition 8.5, $w_j \in W^s_\infty(\Omega)$ for $j \in [p]$. Now let $\varepsilon \in (0, 1]$, for any $j \in [p]$, let $w_j^\varepsilon \in C^\infty(\Omega)$ be the ε approximation of w_j as defined in proposition 8.4, i.e., $w_j = \tilde{w}_j^\epsilon|_\Omega$ where $\tilde{w}_j^\epsilon \in C^\infty(\mathbb{R}^d)$ and

$$\|w_j - w_j^\varepsilon\|_{L^\infty(\Omega)} \leq C_1\varepsilon^s\|w_j\|_{W^s_\infty(\Omega)}, \qquad \|w_j^\varepsilon\|_{W^r_\infty(\Omega)} \leq C_2\varepsilon^{s-r}\|w_j\|_{W^s_\infty(\Omega)}. \tag{8.51}$$

with C_1, C_2 depending only on r, s, d, Ω. Now, since $f(x) - f_* = \sum_{j\in[p]} w_j^2(x)$, for any $x \in \Omega$, we have

$$\begin{aligned}\Big\|f - f_* - \sum_{j\in[p]} w_j^{\varepsilon 2}\Big\|_{L^\infty(\Omega)} &= \Big\| \sum_{j\in[p]} (w_j - w_j^\varepsilon)(2w_j \ - \ (w_j - w_j^\varepsilon))\Big\|_{L^\infty(\Omega)} \\ &\le \sum_{j\in[p]} \|w_j - w_j^\varepsilon\|_{L^\infty(\Omega)}(2\|w_j\|_{L^\infty(\Omega)} + \|w_j - w_j^\varepsilon\|_{L^\infty(\Omega)}) \\ &\le \sum_{j\in[p]} C_1\varepsilon^s\|w_j\|_{W^s_\infty(\Omega)}(2\|w_j\|_{W^s_\infty(\Omega)} + C_1\varepsilon^s\|w_j\|_{W^s_\infty(\Omega)}) \\ &\le \varepsilon^s \ p \ C_1(2 + C_1) \max_{j\in[p]} \|w_j\|^2_{W^s_\infty(\Omega)},\end{aligned}$$

where we use the first equation of Eq. (8.51) to go from the second to the third line.

Recall that $\mathcal{H}$ is defined to be a RKHS associated to the Sobolev kernel k_r defined in Example 8.1 for a given $r > \max(s, d/2)$. As mentioned in Example 8.1, we have $\mathcal{H} = W_2^r(\Omega)$, and $\|\cdot\|_{\mathcal{H}}$ is equivalent to $\|\cdot\|_{W_2^r(\Omega)}$, i.e., there exists C_4 depending on Ω, r, d such that $\frac{1}{C_4}\|\cdot\|_{W_2^r(\Omega)} \le \|\cdot\|_{\mathcal{H}} \le C_4\|\cdot\|_{W_2^r(\Omega)}$.

Since $w_j^\varepsilon \in W_2^r(\Omega) = \mathcal{H}$ by Eq. (8.51) for all $j \in [p]$, we can define :

$$A_\varepsilon = \sum_{j\in[p]} w_j^\varepsilon \otimes_{\mathcal{H}} w_j^\varepsilon.$$

It holds :

$$\begin{aligned}\mathrm{Tr}(A_\varepsilon) &\le p \max_{j\in[p]} \|w_j^\varepsilon\|^2_{\mathcal{H}} \le pC_4^2 \max_{j\in[p]} \|w_j^\varepsilon\|^2_{W_2^r(\Omega)} \\ &\le pC_4^2C_5^2 \max_{j\in[p]} \|w_j^\varepsilon\|^2_{W^r_\infty(\Omega)} \\ &\le \varepsilon^{2(s-r)} \ p(C_2C_4C_5)^2 \max_{j\in[p]} \|w_j\|^2_{W^s_\infty(\Omega)}.\end{aligned}$$

where we used Eq. (8.51) and the fact that there exists C_5 such that $\|\cdot\|_{W^r_\infty(\Omega)} \le C_5\|\cdot\|_{W_2^r(\Omega)}$ (see proposition 8.5). To conclude, we use proposition 8.6 to bound $\|\cdot\|_{W^s_\infty(\Omega)} \le C_6\|\cdot\|_{\Omega,s}$. □

8.G .4 Proof of Theorem 8.11

Proof. The proof of the existence of a minimizer corresponds essentially to the first part of the proof of Theorem 8.5 and we skip it. Let $\varepsilon \in (0, 1]$, by applying Theorem 8.10 to f we know that there exists $A_\varepsilon \in \mathbb{S}_+(\mathcal{H})$ satisfying Eq. (8.33). Define $f_\varepsilon = \langle\phi(x), A_\varepsilon\phi(x)\rangle$ for all $x \in \Omega$, by Theorem 8.10 we have

$$\mathrm{Tr}(A_\varepsilon) \ \le \ C_f\epsilon^{-2(r-s)}, \qquad \sup_{x\in\Omega} |f(x) - f_\varepsilon(x)| \ \le \ C'_f\varepsilon^s.$$

Now consider the problem in Eq. (8.34) and denote by $p^\varepsilon_{\lambda,n}$ its optimum. Since $f_\varepsilon(x_i) - c = \Phi_i^\top B\Phi_i$ implies $|f(x_i) - c - \Phi_i^\top B\Phi_i| \le \tau$, since we required $\tau \ge \sup_{x\in\Omega} |f(x) - f_\varepsilon(x)|$. Then in this case Eq. (8.32) is a relaxation of Eq. (8.34) and we have that $p^\varepsilon_{\lambda,n} - \tilde{c} - \lambda\,\mathrm{Tr}(\tilde{B}) \le 0$. So, we can apply Theorem 8.9 to f_ε with $\theta_1, \theta_2, \tau_2 = 0$ and $\tau_1 = \tau$, obtaining for any $m \in \mathbb{N}$ and $m < s - d/2$

$$\begin{aligned}|\tilde{c} - f^\varepsilon_*| \ &\le \ 14\tau + 7\eta|f_\epsilon|_{\Omega,m} + 6\lambda\,\mathrm{Tr}(A_\epsilon), \\ \mathrm{Tr}(\tilde{B}) \ &\le \ 8\ \mathrm{Tr}(A_\epsilon) \ + \ 8\tfrac{\eta}{\lambda}\,|f_\epsilon|_{\Omega,m} \ + \ 16\tfrac{\tau}{\lambda}.\end{aligned}$$

where f_*^ε is the infimum of f_ε (see Remark 26), and satisfies

$$|f_* - f_*^\varepsilon| = |\min_{x\in\Omega} f(x) - \inf_{x\in\Omega} f_\varepsilon(x)| = |\inf_{x\in\Omega} f(x) - \inf_{x\in\Omega} f_\varepsilon(x)| \le \sup_{x\in\Omega} |f(x) - f_\varepsilon(x)| \le \tau.$$

By the same reasoning in the proof of Theorem 8.4 used to bound $|g|_{\Omega,m}$, we have that

$$|f_\varepsilon|_{\Omega,m} \le \mathsf{MD}_m \operatorname{Tr}(A_\varepsilon) \le C_f \mathsf{MD}_m \varepsilon^{-2r+2s}.$$

Combining together the inequalities above, with the fact that $\lambda \ge 2\mathsf{MD}_m\eta$, we have

$$|\tilde{c} - f_*| \;\le\; 10\lambda C_f \epsilon^{-2(r-s)} \;+\; 15\tau, \qquad \operatorname{Tr}(\tilde{B}) \;\le\; 12\, C_f \epsilon^{-2(r-s)} \;+\; 16\tfrac{\tau}{\lambda}.$$

Now we set ε as large as possible such that $\tau \ge \sup_{x\in\Omega} |f(x) - f_\varepsilon(x)|$ holds. In particular we know that requiring $\tau \ge C'_f \varepsilon^s$ guarantees $\tau \ge \sup_{x\in\Omega} |f(x) - f_\varepsilon(x)|$. Then by setting $\varepsilon = 1$ when $\tau \ge C'_f$, we have

$$|\tilde{c} - f_*| \;\le\; 10\,\lambda\, C_f \;+\; 15\tau\,, \qquad \operatorname{Tr}(\tilde{B}) \;\le\; 12\, C_f \;+\; 16\tfrac{\tau}{\lambda}.$$

By setting $\varepsilon = (\tau/C'_f)^{1/s}$ when $\tau \le C'_f$, we have

$$|\tilde{c} - f_*| \;\le\; 10\,\lambda\, C_f (C'_f)^{2\frac{r-s}{s}} \tau^{-2\frac{r-s}{s}} \;+\; 15\tau\,, \qquad \operatorname{Tr}(\tilde{B}) \;\le\; 12\, C_f (C'_f)^{2\frac{r-s}{s}} \tau^{-2\frac{r-s}{s}} \;+\; 16\tfrac{\tau}{\lambda}.$$

Selecting $\tau = \lambda^{\frac{s}{2r-s}}$ and combining the inequality for the two cases above, leads to

$$|\tilde{c} - f_*| \;\le\; \widetilde{C}_f(\lambda + \lambda^{\frac{s}{2r-s}}),\; \operatorname{Tr}(\tilde{B}) \;\le\; 12\, C_f \;+\; \widetilde{C}'_f \lambda^{-(1-\frac{s}{2r-s})}.$$

where

$$\widetilde{C}_f = \max\left(10 C_f (C'_f)^{2\frac{r-s}{s}} + 15, 10 C_f\right), \qquad \widetilde{C}'_f = 12\, C_f (C'_f)^{2\frac{r-s}{s}} + 16$$

□

8.G .5 Certificate of optimality for the global minimizer candidate of Eq. (8.23)

Theorem 8.14 (Certificate of optimality for Eq. (8.23))**.** *Let Ω satisfy Assumption 8.1(a) for some $r > 0$. Let k be a kernel satisfying Assumptions 8.2(a) and 8.2(d) for some $m \ge 2$. Let $\widehat{X} = \{x_1, \dots, x_n\} \subset \Omega$ with $n \in \mathbb{N}$ such that $h_{\widehat{X},\Omega} \le \frac{r}{18(m-1)^2}$. Let $f \in C^m(\Omega)$ and let $\hat{c} \in \mathbb{R}, \hat{z} \in \mathbb{R}^d, \hat{B} \in \mathbb{S}_+(\mathbb{R}^n)$ and $\tau \ge 0$ satisfying*

$$|f(x_i) - \tfrac{\nu}{2}\|x_i\|^2 + \nu x_i^\top \hat{z} - \hat{c} \;-\; \Phi_i^\top \hat{B} \Phi_i| \le \tau, \quad i \in [n] \tag{8.52}$$

where Φ_i are defined in Sec. 8.2 . Let $f_* = \min_{x\in\Omega} f(x)$ and $\hat{f} = \hat{c} - \frac{\nu}{2}\|\hat{z}\|^2$. Then,

$$|f(\hat{z}) - f_*| \le f(\hat{z}) - \hat{f} + 2\tau + C_1 h^m_{\widehat{X},\Omega}, \tag{8.53}$$

$$\tfrac{\nu}{2}\|\zeta - \hat{z}\|^2 \le f(\hat{z}) - \hat{f} + 2\tau + C_2 h^m_{\widehat{X},\Omega}. \tag{8.54}$$

and $C_1 = C_0(|f|_{\Omega,m} + \mathsf{MD}_m \operatorname{Tr}(\hat{B}) + \mathsf{MD}_m \hat{C})$, $C_2 = C_0(|f|_{\Omega,m} + \nu + \mathsf{MD}_m \operatorname{Tr}(\hat{B}))$, where $\hat{C} = \frac{\nu}{2}\|R^{-\top}(X - 1_n\hat{\zeta}^\top)\|^2$, with $X \in \mathbb{R}^{n\times d}$ the matrix whose i-th row corresponds to the point x_i and $1_n \in \mathbb{R}^n$ the vector where each element is 1. The constants C_0, defined in Theorem 8.4, and $m, \mathsf{M}, \mathsf{D}_m$, defined in Assumptions 8.2(a) and 8.2(d), do not depend on $n, \widehat{X}, h_{\widehat{X},\Omega}, \hat{c}, \hat{B}$ or f.

Proof. We divide the proof in two steps

Step 1. First note that

$$\hat{g}(x) := f(x) - \tfrac{\nu}{2}\|x\|^2 + \nu x^\top \hat{z} - \hat{c} = f(x) - \tfrac{\nu}{2}\|x - \hat{z}\|^2 - \hat{f}.$$

By applying Theorem 8.4 with $g = \hat{g}$ and $B = \hat{B}$ we have that for any $x \in \Omega$ $f(x) - \frac{\nu}{2}\|x - \hat{z}\|^2 - \hat{f} = \hat{g}(x) \geq -\varepsilon - 2\tau$, where $\varepsilon = C_0(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \operatorname{Tr}(\hat{B}))h^m_{\hat{X},\Omega}$ and C_0 is defined in Theorem 8.4. In particular this implies that

$$f(\zeta) - \hat{f} - \tfrac{\nu}{2}\|x - \hat{z}\|^2 \geq -\varepsilon - 2\tau,$$

from which Eq. (8.54) is obtained by considering that $f(\hat{z}) \geq f(\zeta)$ since ζ is a minimizer of f. To conclude the proof of Eq. (8.54) note that $|\hat{g}|_{\Omega,m} \leq |f|_{\Omega,m} + \nu$ since $m \geq 2$.

Step 2. Now to obtain Eq. (8.53) we need to do a slightly different construction. Let $u_j(x) = e_j^\top(x - \hat{z})$ for any $x \in \Omega$. Note that since u_j is the restriction to Ω of a C^∞ function on $\mathbb{R}^d$, by Assumption 8.2(a), $u_j \in \mathcal{H}$. Moreover, note that $\frac{\nu}{2}\|x - \hat{z}\|^2 = \frac{\nu}{2}\sum_{j=1}^d u_j(x)^2$. Take $\hat{u}_j \in \mathbb{R}^n$ defined as $\hat{u}_j = V^* u_j$ and note that

$$\Phi_i^\top \hat{u}_j = \langle V\phi(x_i), V^* u_j \rangle = \langle V^* V \phi(x_i), u_j \rangle = \langle P\phi(x_i), u_j \rangle = u_j(x_i).$$

Then, defining $\hat{G} = \frac{\nu}{2}\sum_{i=1}^d \hat{u}_j \hat{u}_j^\top \in \mathbb{S}_+(\mathbb{R}^n)$ we have

$$\tfrac{\nu}{2}\|x_i - \hat{z}\|^2 = \Phi_i^\top \hat{G} \Phi_i, \qquad \forall i \in [n].$$

Substituting $-\frac{\nu}{2}\|x_i\|^2 + \nu x_i^\top \hat{z}$ with $\frac{\nu}{2}\|\hat{z}\|^2 - \Phi_i^\top \hat{G}\Phi_i$ in the inequality in Eq. (8.52), we obtain

$$|f(x_i) - \hat{f} \; - \; \Phi_i^\top(\hat{B} + \hat{G})\Phi_i^\top| \leq \tau, \quad \forall i \in [n].$$

By applying Theorem 8.4 with $g(x) = f(x) - \hat{f}$ and $B = \hat{B} + \hat{G}$ we have that $f(x) - \hat{f} \geq -\varepsilon - 2\tau$ for all $x \in \Omega$, where $\varepsilon = C' h^m_{\hat{X},\Omega}$ with $C' = C_0(|g|_{\Omega,m} + \mathsf{MD}_m \operatorname{Tr}(\hat{B} + \hat{G}))$. In particular, $f(\zeta) - \hat{f} \geq -\varepsilon - 2\tau$, from which Eq. (8.53) is obtained considering that $f(\hat{z}) \geq f_*$ since ζ is a minimizer of f.

Finally, note that $|g|_{\Omega,m} \leq |f|_{\Omega,m}$ since $m \geq 1$. The proof is concluded by noting that using the definition of V we have $\hat{u}_j = R^{-\top}\hat{v}_j$ with $\hat{v}_j \in \mathbb{R}^n$ corresponding to $\hat{v}_j = (u_j(x_1), \ldots, u_j(x_n))$ for $j \in [d]$ and that $\operatorname{Tr}(\hat{G}) = \frac{\nu}{2}\sum_{j\in[d]}\|\hat{u}_j\|^2$. In particular, some basic linear algebra leads to $\operatorname{Tr}(\hat{G}) = \frac{\nu}{2}\|R^{-\top}(X - 1_n\hat{z}^\top)\|^2$. □

8.H Details on the algorithmic setup used in the benchmark experiments

In this section, we explain exactly the algorithmic setup which we used to perform the experiments in Section Sec. 8.10 .1. In all the following problems, the set Ω on which we will minimize the function will be a hyper-rectangle. Given a hyper-rectangle R, we will identify it with its center $c_R \in \mathbb{R}^d$ and its width $w_R \in \mathbb{R}^d$, such that $R = \prod_{i=1}^d ((c_R)_i - (w_R)_i/2, (c_R)_i + (w_R)_i/2)$.

We start by defining algorithm 7 whose main goal is to find a global minimizer as described in the previous sections given sample points $(x_1, ..., x_n)$.

Algorithm 7 Finding a minimizer given points X

function FINDMINIMIZER($f, X, k(\cdot,\cdot), \lambda_{\min}, \lambda_{\max}, \varepsilon$)
  $K = (k(x_i, x_j))_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$
  Φ such that $\Phi^\top\Phi = K$ (cholesky decomposition)
  $\boldsymbol{f}_X = (f(x_i))_{1\leq i\leq n} \in \mathbb{R}^n$

  function SCALARFUNCTION(t)
    $\lambda = e^t$
    $\widehat{\alpha}$ solution to Eq. (8.55) with $\lambda, \varepsilon, \Phi, \boldsymbol{f}_X$
    $\widehat{x} = \sum_{i=1}^n \widehat{\alpha}_i x_i$
    $\widehat{f} = f(\widehat{x})$
    return $\widehat{f}, \widehat{x}$
  end function
  $\widehat{f}, \widehat{x}$ = MINIMIZESCALAR(SCALARFUNCTION ,$t_{\min} = \log(\lambda_{\min}), t_{\max} = \log(\lambda_{\max})$)
  return $\widehat{f}, \widehat{x}$
end function

Recall that the algorithm introduced in sections 6 and 7.1 computes a minimizer by solving problem :

$$\widehat{\alpha} = \arg\min_{\substack{\alpha\in\mathbb{R}^n \\ \alpha^\top 1_n = 1}} \sum_{i=1}^n \alpha_i f(x_i) - \frac{\varepsilon}{n}\log\det\left(\Phi^\top \mathrm{Diag}(\alpha)\Phi + \lambda I\right) + \frac{\varepsilon}{n}\log\frac{\varepsilon}{n} - \varepsilon, \tag{8.55}$$

where Φ satisfies $\Phi^\top\Phi = K$ for $K = (k(x_i, x_j))_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$, and choosing $\widehat{x}$ as the approximation of the minimizer, defined by

$$\widehat{x} = \sum_{i=1}^n \widehat{\alpha}_i x_i. \tag{8.56}$$

However, the kernel k, and the hyper-parameters λ also have to be chosen.

Therefore, algorithm 7 will use as inputs :

- The function f to minimize;
- the evaluation points $x_i,\ 1 \leq i \leq n$, summarized in a matrix $X \in \mathbb{R}^{n\times d}$;
- the kernel k;
- two parameters $\lambda_{\min}$ and $\lambda_{\max}$ such that we can choose λ in $[\lambda_{\min}, \lambda_{\max}]$;
- The paramter ε, which controls the log barrier.

For simplicity, we hide the hyperparameters linked to the solving of Eq. (8.55) through a Newton method, as explained in Sec. 8.6 .

algorithm 7 automatically selects the hyperparameter λ by minimizing the function wich to λ associates the function value of the resulting $\widehat{x}$ on a log scale (SCALARFUNCTION). This function is minimized in the range $[\lambda_{\min}, \lambda_{\max}]$ through the function MINIMIZESCALAR. Hence, the number of function evaluations inherent to running this algorithm is $n + n_{\min}$ where $n_{\min}$ is the number of

In our experiments, we use $\varepsilon = 10^{-3}$, $\lambda_{\min} = 10^{-12}, \lambda_{\max} = 1$ and we use either the Brent method or simply a grid search with a maximum number of 100 points. This minimization does not have to be very precise.

The full algorithm we use is an iterative scheme and is written down in algorithm 8, computing a sequence (x_k) of approximations of a minimizer of f by iteratively reducing the size of the hyper-rectangle from which the points used in algorithm 7 are sampled.

More precisely, we start from points sampled from a hyper-rectangle with center $x_0 = c_\Omega$ and with width $w_0 = w_\Omega$ (that is the hyper-rectangle Ω) to form $m-1$ samples which, together with x_0, form the points $\widetilde{X}_0 \in \mathbb{R}^{m\times d}$ used to compute the first approximation of the minimizer using FindMinimizer : x_1. Then at each step k, we use the last approximation of the minimizer x_k as the new center of the hyper-rectangle, with width w_k which is set through the predefined function Contraction as $w_k = \text{Contraction}(k)w_0$. As for the first iteration, we then form the concatenation $\widetilde{X}_k \in \mathbb{R}^{m\times d}$ of $m-1$ samples from this hyper-rectangle plus x_k. In order to keep track of the previous points (as a kind of momentum), we apply FindMinimizer with $X_k = [\widetilde{X}_k, \widetilde{X}_{k-1}, \widetilde{X}_{k-2}]$, that is keeping the two last set of points as well as the ones sampled for the k-th step.

Algorithm 8 Converging to the minimum

function FindMinimizerIter($f, \Omega, m, N, k_{(\cdot)}(\cdot,\cdot)$, Contraction)
  $\varepsilon = 10^{-3}, \lambda_{\min} = 10^{-12}, \lambda_{\max} = 1$
  $\widetilde{X}_{-2}, \widetilde{X}_{-1} = [], []$
  $x_0, w_0 = c_\Omega, w_\Omega$
  for $k = 0$ to $N-1$ **do**
    $w_k = \text{Contraction}(k)\times w_0$
    $\sigma_k = \|w_k\|/2$
    $\widetilde{X}_k = [x_k^\top, \text{Uniform}(x_k, w_k, m-1)^\top]^\top$
    $X_k = [\widetilde{X}_{k-2}^\top, \widetilde{X}_{k-1}^\top, \widetilde{X}_k^\top]^\top$
    $f_{k+1}, x_{k+1} = \text{FindMinimizer}(f, X_k, k_{\sigma_k}, \lambda_{\min}, \lambda_{\max}, \varepsilon)$
  end for
  return f_N, x_N
end function

The function FindMinimizerIter uses the following parameters:

- a kernel function $x, x', \sigma \mapsto k_\sigma(x, x')$ such that σ is a parameter to adapt to the the typical width of the data;
- the initial hyper-rectangle Ω;
- the function f;
- the contraction function Contraction to set the width of the successive hyper-rectangles;
- the number m of new points sampled and used at each iteration;
- the number N of iterations.

In our implementation, we use the following parameters.

- For $\sigma > 0$ and $x, y \in \mathbb{R}^d$, we will use the following kernel, which is a mix between the Gaussian (very regular functions) and the Abel kernel (Sobolev functions of order

$s = (d+1)/2$ functions), plus a small term 0.01 which allows to handle the constant component of a function more easily.

$$k_\sigma(x,y) = 0.01 + \exp(-\|x-y\|^2/(2\sigma^2)) + \exp(-\|x-y\|/\sigma). \tag{8.57}$$

- We will use the following contraction function, which depends on the dimension as well as the number of iterations :

$$\textsc{Contraction}(k) = \max\left(\left(1+\tfrac{1}{d}\right)^{-k}, \tfrac{1}{1+k^{0.6}}\right). \tag{8.58}$$

- The number N of iterations will be set to $N = 200$ unless stated otherwise.
- m will be specified in the experiments : indeed, the higher the dimension, the larger m has to be in order to get meaningful results. Note that one actually uses $n = 3m$ points (from the third iteration onwards) to form the optimization problem, hence the dimension of the SDP solved with the Newton method will be $3m$.

Remark 29. *It is equivalent to minimize a function f and minimize the function $\frac{f}{f+c}$ for a positive constant c. This allows to minimize a function in $[0,1]$ instead of minimizing a real-valued function : however, this also makes higher derivatives behave differently than those of the original function. In practice, instead of minimizing f directly, we minimize $\frac{f}{f+c}$, where c is chosen such that $\frac{f}{f+c}$ will be spread evenly over $[0,1]$, typically by selecting c as a quantile of the $(f(x_i))_{1\leq i\leq n}$ (we choose the 0.25 quantile). We performed experiments by comparing this renormalization method with simply minimizing f, and this yields slightly better results.*

8.H .1 Additional experiments for global optimization

	d	iters thresh	final absolute error	fevs/iter
Griewank	2	4	8.54E-13	21
CrossInTray	2	10	0.00E+00	21
Bukin04	2	16	1.58E-10	21
Matyas	2	1	1.80E-16	21
BartelsConn	2	6	0.00E+00	21
RotatedEllipse01	2	3	2.63E-12	21
Branin01	2	5	0.00E+00	21
OddSquare	2	7	2.38E-07	21
Ursem04	2	4	0.00E+00	21
Ripple01	2	NaN	9.52E-02	21
Brent	2	2	2.77E-09	21
Schaffer02	2	1	2.64E-14	21
DropWave	2	27	0.00E+00	21
NeedleEye	2	1	0.00E+00	21
Schwefel22	2	5	1.17E-08	21
XinSheYang01	2	2	1.05E-08	21
Pinter	2	1	5.40E-16	21
Penalty01	2	3	5.10E-17	21
Langermann	2	5	0.00E+00	21
Salomon	2	7	1.81E-08	21
VenterSobiezcczanskiSobieski	2	1	0.00E+00	21

Schaffer03	2	14	8.38E-07	21
Shubert04	2	7	-1.91E-06	21
Price01	2	5	2.40E-05	21
Giunta	2	3	0.00E+00	21
Cigar	2	2	1.94E-08	21
Bukin02	2	4	0.00E+00	21
YaoLiu09	2	2	0.00E+00	21
Vincent	2	8	0.00E+00	21
Qing	2	8	3.04E-05	21
WayburnSeader01	2	2	7.58E-11	21
Levy13	2	5	7.61E-13	21
Schaffer04	2	7	-3.58E-07	21
Brown	2	3	4.37E-18	21
Ackley01	2	8	6.50E-09	21
CrossLegTable	2	NaN	9.98E-01	21
Schwefel36	2	4	0.00E+00	21
CosineMixture	2	9	0.00E+00	21
Quadratic	2	2	0.00E+00	21
Exponential	2	1	0.00E+00	21
NewFunction01	2	NaN	1.18E-01	21
HolderTable	2	5	0.00E+00	21
TestTubeHolder	2	NaN	1.98E-02	21
Ursem03	2	5	0.00E+00	21
Sphere	2	1	7.73E-16	21
Levy03	2	3	1.19E-18	21
Schaffer01	2	1	4.44E-15	21
Rastrigin	2	7	1.07E-14	21
McCormick	2	2	0.00E+00	21
SixHumpCamel	2	5	-4.77E-07	21
RotatedEllipse02	2	3	4.94E-15	21
Branin02	2	NaN	1.25E+00	21
Alpine01	2	28	7.44E-09	21
Quintic	2	10	2.25E-05	21
Schwefel26	2	9	-5.45E-07	21
SineEnvelope	2	3	7.83E-15	21
Stochastic	2	5	3.99E-07	21
ZeroSum	2	NaN	1.00E+00	21
UrsemWaves	2	NaN	9.08E-01	21
RosenbrockModified	2	11	0.00E+00	21
Penalty02	2	68	1.68E-06	21
XinSheYang02	2	7	1.21E-11	21
Ackley03	2	8	0.00E+00	21
YaoLiu04	2	7	5.22E-09	21
Schwefel21	2	3	9.83E-09	21
Zimmerman	2	5	7.46E-03	21
Decanomial	2	2	3.95E-09	21
Alpine02	2	3	-3.81E-06	21
Ursem01	2	3	0.00E+00	21
NewFunction02	2	NaN	2.01E-01	21

Levy05	2	6	0.00E+00	21
Chichinadze	2	2	7.63E-06	21
Trefethen	2	55	6.91E-06	21
Hansen	2	28	0.00E+00	21
DeckkersAarts	2	13	0.00E+00	21
Michalewicz	2	4	-3.34E-06	21
Treccani	2	1	1.87E-15	21
Bukin06	2	2	3.42E-04	21
XinSheYang03	2	2	0.00E+00	21
Ackley02	2	23	0.00E+00	21
Csendes	2	NaN	NAN	21
Schwefel20	2	8	4.63E-08	21
Whitley	2	3	2.95E-11	21
XinSheYang04	2	49	0.00E+00	21
Step	2	1	0.00E+00	21
WayburnSeader02	2	4	6.14E-09	21
Mishra02	2	1	0.00E+00	21
StyblinskiTang	2	6	0.00E+00	21
ElAttarVidyasagarDutta	2	47	0.00E+00	21
Price03	2	31	3.10E-16	21
HimmelBlau	2	2	1.17E-14	21
DixonPrice	2	6	6.07E-14	21
Parsopoulos	2	6	3.35E-16	21
Price02	2	17	0.00E+00	21
Mishra08	2	2	3.99E-16	21
Beale	2	3	5.06E-19	21
Sodp	2	1	1.62E-23	21
Schwefel06	2	15	3.02E-08	21
Zirilli	2	6	-5.96E-08	21
Mishra01	2	1	0.00E+00	21
Damavandi	2	NaN	NAN	21
ThreeHumpCamel	2	1	1.68E-19	21
Mishra07	2	1	1.97E-29	21
Pathological	2	2	2.22E-14	21
Ripple25	2	NaN	1.66E-01	21
Hosaki	2	2	0.00E+00	21
Exp2	2	2	2.62E-15	21
Trigonometric02	2	158	0.00E+00	21
Judge	2	2	0.00E+00	21
AMGM	2	1	0.00E+00	21
CrownedCross	2	NaN	3.35E-02	21
JennrichSampson	2	6	7.63E-06	21
Schwefel01	2	1	3.67E-16	21
Adjiman	2	5	0.00E+00	21
Katsuura	2	3	0.00E+00	21
Bohachevsky1	2	6	1.24E-14	21
Shubert01	2	7	-1.53E-05	21
Zacharov	2	2	4.89E-16	21
Mishra06	2	5	2.38E-07	21

Deb01	2	2	0.00E+00	21
Infinity	2	1	2.51E-34	21
Sargan	2	4	4.92E-13	21
Mishra03	2	NaN	1.33E-01	21
Bird	2	3	0.00E+00	21
Mishra05	2	11	4.66E-04	21
GoldsteinPrice	2	3	0.00E+00	21
Mishra10	2	1	0.00E+00	21
EggCrate	2	5	1.43E-16	21
Schwefel02	2	2	1.47E-13	21
Cube	2	21	5.98E-09	21
Rosenbrock	2	35	5.15E-05	21
Bohachevsky2	2	2	4.82E-14	21
DeflectedCorrugatedSpring	2	8	0.00E+00	21
Step2	2	1	0.00E+00	21
Deb03	2	4	0.00E+00	21
Wavy	2	1	0.00E+00	21
Trigonometric01	2	NaN	5.83E-01	21
Tripod	2	20	8.92E-08	21
Weierstrass	2	16	3.52E-06	21
Shubert03	2	28	0.00E+00	21
MultiModal	2	1	6.00E-26	21
Price04	2	3	1.06E-20	21
Deceptive	2	7	2.97E-04	21
Schwefel04	2	2	1.28E-17	21
Leon	2	12	8.62E-15	21
CarromTable	2	10	0.00E+00	21
Rana	2	NaN	1.86E+00	21
Bohachevsky3	2	6	1.67E-15	21
Plateau	2	1	0.00E+00	21
PenHolder	2	6	0.00E+00	21
Zettl	2	2	0.00E+00	21
Keane	2	1	2.25E-36	21
Mishra11	2	6	3.16E-30	21
Mishra04	2	NaN	1.78E-01	21
FreudensteinRoth	2	5	1.55E-10	21
BoxBetts	3	10	0.00E+00	26
Gulf	3	5	1.22E-03	26
Wolfe	3	1	0.00E+00	26
Mishra09	3	1	2.47E-25	26
Ratkowsky02	3	2	2.86E-06	26
Hartmann3	3	5	0.00E+00	26
Meyer	3	NaN	3.72E+09	26
HelicalValley	3	5	7.67E-09	26
Colville	4	32	1.87E-03	31
Corana	4	1	0.00E+00	31
Shekel07	4	20	9.54E-07	31
PowerSum	4	2	3.26E-04	31
Ratkowsky01	4	90	3.69E+02	31

MieleCantrell	4	3	9.03E-13	31
Powell	4	6	2.85E-07	31
Shekel10	4	18	0.00E+00	31
Shekel05	4	18	1.91E-06	31
BiggsExp04	4	12	7.88E-05	31
Gear	4	2	1.18E-09	31
Kowalik	4	15	4.87E-05	31
DeVilliersGlasser01	4	4	1.06E+03	31
DeVilliersGlasser02	5	NaN	2.28E+03	36
Dolan	5	2	3.78E-13	36
BiggsExp05	5	3	2.64E-03	36
Trid	6	10	0.00E+00	41
Watson	6	11	1.09E-03	41
Hartmann6	6	8	0.00E+00	41
LennardJones	6	2	0.00E+00	41
Thurber	7	125	9.70E+03	46
Xor	9	NaN	6.99E-03	56
Paviani	10	23	1.03E-04	61
Cola	17	68	3.35E-01	96

Table 8.4: Complete results of our algorithm on functions on $\mathbb{R}^d$ for $d \geq 2$

Chapter 9

Second order condition to decompose smooth functions as sums of squares

This chapter is a verbatim of the work :

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Second order conditions to decompose smooth functions as sums of squares, 2022b. URL https://arxiv.org/abs/2202.13729.

Contents

9.1 Introduction 419
9.2 Decomposition as sums of squares given second order conditions (Euclidean case) 424
9.3 Global decomposition as a sum of squares for functions on manifolds 431
9.4 Proof of the local decomposition as a sum of squares 437
9.5 Discussion and possible extensions 439
9.A Around partitions of unity and gluing functions 441
9.B Morse lemma 443

9.1 Introduction

The relationship between non-negative functions and functions decomposable as sums of squares is a fundamental question in both theoretical and applied mathematics. From a theoretical viewpoint, the decomposability of a non-negative function in terms of sum of squares is the basis of important theoretical objects and properties: quadratic modules (Marshall, 2006) in algebraic geometry, regularizing operators such as Laplacians or sub-Laplacians in (sub-)Riemannian geometry (Hörmander, 1967; Bony, 1998-1999), non-negative symbols in pseudo-differential calculus (Hörmander, 2007; Tataru, 2002). From an applicative viewpoint, representing a non-negative function in terms of sum of squares allows to simplify the analysis of probability representations and optimization problems (Lasserre, 2010; Marteau-Ferey, Bach, and Rudi, 2020). Restricting to the case of non-negative polynomials this has been applied to global optimization and generalized methods of moments (Lasserre, 2010; Henrion, Korda, and Lasserre, 2020). More generally, the decomposition of non-negative p-times differentiable functions allowed

to derive simple and fast optimization algorithms in the context of global optimization (Rudi, Marteau-Ferey, and Bach, 2020), the Kantorovich problem in optimal transport (Vacher, Muzellec, Rudi, Bach, and Vialard, 2021), some formulations of optimal control (Berthier, Carpentier, Rudi, and Bach, 2021). Moreover, it allowed to obtain an effective and concise representation for probability densities, with applications in probabilistic inference, sampling, machine learning (Rudi and Ciliberto, 2021; Marteau-Ferey, Bach, and Rudi, 2022a).

The importance of preserving regularity

In this work, we state sufficient conditions for a non-negative function f to be written as a sum of squares of functions f_i. Of course, if no other constraints are added, this is a trivial problem as writing $f = (\sqrt{f})^2$ would offer an immediate solution. What we want to understand here are sufficient conditions which allow to inherit a form of regularity of the function f in the sum of squares decomposition. This is not necessarily the case when taking the square root: for example, the map $(x, y) \mapsto x^2 + y^2$ is smooth and a sum of smooth squares, but its square root is not differentiable at $(0, 0)$.

In general, being able to decompose the function with a certain regularity is important. Of course, there is a complex interaction between the structural constraint of being a sum of squares and the original regularity of the function, and the two may not work very well together (see Theorem 9.4). However, for certain theoretical and applied problems, it is crucial to maintain some regularity. For example, in the setting introduced by Rudi, Marteau-Ferey, and Bach (2020), the speed of convergence of the presented algorithm of global optimization depends on the regularity of the sum of squares representation of the function $f - f_*$ (where f_* is the global minimum of f). In polynomial sum-of-squares (SoS) optimization, the running time depends on the degree in the sum of squares decomposition of $P - P_*$.

Abstractly, we can formulate the following generic question. If $f \in \mathcal{C}_1$ where $\mathcal{C}_1$ describes a form of regularity, can we write f as a sum of squares of functions of class $\mathcal{C}_2$, where $\mathcal{C}_2$ inherits the regularity properties $\mathcal{C}_1$ as much as possible?

Problem setting

In this work, we will concentrate on the class C^p of p times differentiable functions with continuous p-th derivatives on $\mathbb{R}^d$ (or any d-dimensional manifolds M). For simplicity, in this introduction, we will state the main results for functions on $\mathbb{R}^d$. We will show that under a certain condition on the set of zeros $\mathcal{Z}$ of f, if f is a C^p non-negative function on $\mathbb{R}^d$, it can be decomposed as

$$f = \sum_{i \in I} f_i^2, \qquad f_i \in C^{p-2}(\mathbb{R}^d), \tag{9.1}$$

where (f_i) is an at most countable family and has locally finite support. Two elements are important in Eq. (9.1): the locally finite aspect and the regularity of the functions f_i, i.e., $p - 2$. This is a consequence of the fact that we will consider *second order sufficient conditions*, hence the loss of two derivatives.

9.1 .1 Intuition and previous results

Let us give an intuition as to how we obtain decompositions in the form Eq. (9.1). First, we start by proving that this decomposition holds locally in a neighborhood of any $x_0 \in \mathbb{R}^d$. It is then possible to invoke a result to "glue" the local decompositions together; we develop the tools to do so in Sec. 9.3 .2 (note that this is one of the key differences between results for

polynomials and results for functions). For any fixed $x_0 \in \mathbb{R}^d$, if $f(x_0) > 0$, then $f_1 := \sqrt{f}$ is well defined and of class C^p around x_0, and so Eq. (9.1) holds locally around x_0 since $f = f_1^2$. The crux of the problem is to determine whether f can be decomposed as a sum of squares around a point in the set of zeros $\mathcal{Z}$ of f, i.e., the set of points x such that $f(x) = 0$. Since f is non-negative, all such points are necessarily minimizers of f, hence the following *necessary second-order condition*:

$$\forall x_0 \in \mathcal{Z},\ \nabla f(x_0) = 0,\ \nabla^2 f(x_0) \succeq 0. \tag{9.2}$$

Around any $x_0 \in \mathcal{Z}$, f can be approximated by a parabola since the eigenvalues of $\nabla^2 f(x_0)$ are non-negative: $f(x) = x^\top \nabla^2 f(x_0) x + o(\|x\|^2)$ using a Taylor expansion. Since any parabola can be written as the sum of at most d squares of linear functions (just write the eigen-decomposition of $\nabla^2 f(x_0)$), we see that up to the $o(\|x\|^2)$ factor, we can indeed write f as a sum of at most d squares around x_0. The whole difficulty of the following results is to go beyond this $o(\|x\|^2)$ approximation and have an exact decomposition, using the Taylor expansion with integral remainder.

It turns out that in the case where $\nabla^2 f(x_0) \succ 0$, that is when the Hessian has strictly positive eigenvalues, this decomposition can be made exact. We will call this condition the *strict Hessian condition* (SHC) at x_0. This result exists in recent work: it is a particular case of Theorem 2 by Rudi, Marteau-Ferey, and Bach (2020), applied to the set $\mathcal{H} = C^{p-2}$. Precisely, it states

Theorem 9.1 (Rudi, Marteau-Ferey, and Bach (2020, Theorem 2))**.** *Let f be a non-negative function of class C^p for $p \geq 2$, and assume that the zeros $\mathcal{Z}$ of f satisfy the strict Hessian condition:*

$$\forall x_0 \in \mathcal{Z},\ \nabla^2 f(x_0) \succ 0. \tag{9.3}$$

If f has a finite number $m = |\mathcal{Z}|$ of zeros, then f satisfies Eq. (9.1) with $dm + 1$ functions f_i.

This situation is illustrated on the left hand side of Sec. 9.1 .1, where the Hessian is positive definite at all four zeros of f and hence satisfies the SHC: by Theorem 9.1, it can be decomposed as a sum of squares. It is not the case on the right hand side, where there is a continuous subspace of zeros: in that case, f does not satisfy the SHC.

Contribution

While the SHC condition Eq. (9.3) already offers a nice result in Theorem 9.1, we see that there is a big difference with the necessary condition Eq. (9.2). Previous results in the literature show that Eq. (9.2) is not sufficient to be decomposed as a sum of squares of C^{p-2} functions as soon as the dimension d is greater than 3 (see Theorem 9.4 in the background section form more details). On the other hand, Eq. (9.3) is very restrictive. In particular, it implies that the set $\mathcal{Z}$ of zeros is discrete. However, in some situations (Vacher, Muzellec, Rudi, Bach, and Vialard, 2021), the set of zeros has a natural structure, which can be a sub-manifold of $\mathbb{R}^d$ (consider for instance the extreme case where $f = 0$). In this paper, we show that if the set $\mathcal{Z}$ of zeros is a sub-manifold of $\mathbb{R}^d$ such that the Hessian of f along this manifold is positive along all directions which are not tangent to $\mathcal{Z}$, then Eq. (9.1) still holds. This is the case for the function depicted in the right hand side of Sec. 9.1 .1, and illustrates the difference between previous works and our contributions. More formally, we prove the following result.

Theorem 9.2. *Let f be a non-negative function of class C^p for $p \geq 2$ and let $\mathcal{Z}$ denote the set of zeros of f. If $\mathcal{Z}$ is a sub-manifold of $\mathbb{R}^d$ of class C^1 such that*

$$\forall x_0 \in \mathcal{Z},\ \forall h \in \mathbb{R}^d \setminus T_{x_0}\mathcal{Z},\ h^\top \nabla^2 f(x_0) h > 0, \tag{9.4}$$

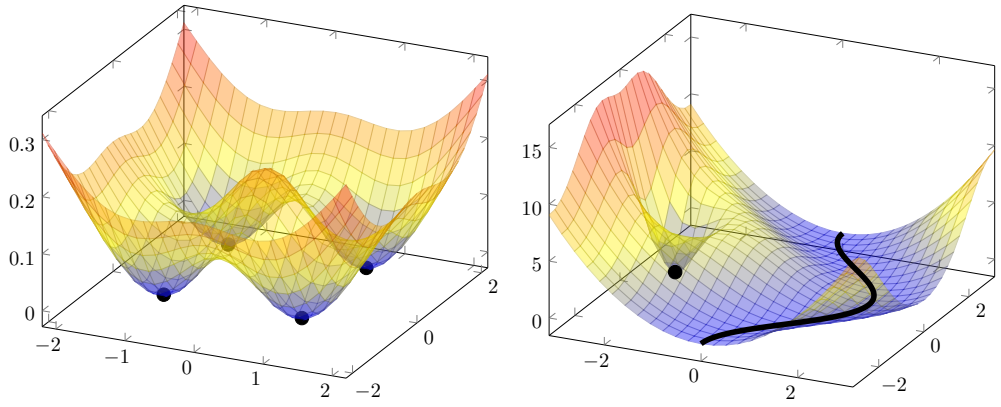

Figure 9.1: Plots of functions $z = f(x, y)$, where the zeros of f are highlighted in black. **left**: f satisfies the SHC, **right**: f satisfies the NHC but not the SHC.

then f satisfies Eq. (9.1), and $\mathcal{Z}$ is of class C^{p-1}. Here, $T_{x_0}\mathcal{Z}$ denotes the tangent space to $\mathcal{Z}$ at x_0, which is a vector sub-space of $\mathbb{R}^d$.

This theorem is proved as Theorem 9.6 in Sec. 9.2 , and the assumption Eq. (9.4) will be referred to as the normal Hessian condition (or NHC). Note that the NHC assumption encompasses that of the SHC assumption of Theorem 9.1; in that case, the results presented in this paper make the result tighter by removing the assumption that $\mathcal{Z}$ be finite and by needing only $d+1$ squares to represent the function, and not $d|\mathcal{Z}|+1$ (see the full version of Theorem 9.6).

The proof techniques used to prove this theorem differ from the proof by Rudi, Marteau-Ferey, and Bach (2020) and use tools from differential geometry and Morse theory. In particular, the proof extends naturally to functions defined on d-dimensional manifolds, which is the object of Sec. 9.3 and Theorem 9.8. This opens the way to new problems, wich are more naturally defined on standard manifolds like the d-dimensional sphere S^d or the d-dimensional torus $\mathbb{T}^d \approx (S^1)^d$.

Background

The problem of decomposing C^p functions as sums of squares has appeared in the context of symbolic calculus, in the proof of the Fefferman-Phong inequality, which is an important regularity result for partial differential operators (see the original article by Fefferman and Phong (1978), and the article by Bony (1998-1999) for the link with sum of squares decompositions, as well as the monograph by Tataru (2002)). In this context, the following result is proved (with $C^{k,1}_{loc}$ denoting the set of k times differentiable functions with locally Lispchitz k-th derivative):

Theorem 9.3 (Fefferman and Phong (1978), Bony, Broglia, Colombini, and Pernazza (2006, Theorem 1.1))**.** *Let Ω be an open set of $\mathbb{R}^d$, $d \geq 1$ and $f \in C^{3,1}_{loc}(\Omega)$ be a non-negative function. Then f can be written as a finite sum of squares of $C^{1,1}_{loc}(\Omega)$ functions.*

In the context of preserving regularity, a natural question which arises is whether increasing the regularity of f can increase the regularity of the functions in a sum of square decomposition. Bony, Broglia, Colombini, and Pernazza (2006); Bony (2005) show that the general answer (under no further assumptions) is negative. More precisely, if f is a function defined on a neighborhood of 0, a local decomposition of f around 0 of class $\mathcal{C}$ is a finite family $(f_i)_{i\in I}$ of functions of class $\mathcal{C}$ defined on an open neighborhood U of 0 such that $\sum_{i\in I} f_i^2 = f$ on U.

Theorem 9.4 (Bony, Broglia, Colombini, and Pernazza (2006, Theorem 2.1))**.** *In all the following cases, there exists $f \in C^\infty$ defined on an open neighborhood of 0 in $\mathbb{R}^d$ such that the following holds:*

- *if $d \geq 4$, f has no local decomposition of class C^2;*
- *if $d = 3$, f has no local decomposition of class C^3.*

The case $d = 1$ is explored by Bony (2005): it is shown in Theorem 1 that if f is of class C^{2m} for m finite, then f can be written as the sum of squares of two functions of class C^m. Moreover, this is shown to be tight: there exists a function $f \in C^{2m}$ with no local decomposition as a sum of squares of functions of class C^{m+k}, for $k \geq 1$. The case $d = 2$ has been explored less in the literature (some results exist when dealing with flat minima, see for example Theorem 2 by Bony (2005)).

To summarize, these results show that without additional assumptions, as soon as the dimension is greater than 3, inheriting the C^p regularity properties of the function f in the sum of squares decomposition is not possible in a satisfactory way, and motivates the introduction of additional geometric assumptions.

Polynomials Decomposing non-negative polynomials as sums of squares has been related to important problems in algebraic geometry during the 20th century. In 1927, on his way to the resolution of Hilbert's 17th problem, Artin (1927) proved that any non-negative polynomial is a sum of squares of rational functions (that is formal fractions of polynomials $P(x)/Q(x)$). Moreover, Hilbert had earlier proved that there exist non-negative polynomials which cannot be written as sum of squares of polynomials by Hilbert (1888) (for more than 3 variables and with degree at least 6 for example). In algebraic geometry, the set of SoS polynomials has very interesting properties, and finding sufficient conditions for a non-negative polynomial or even a positive polynomial to be a sum of squares is an important question. More generally, one usually wishes to understand under which sufficient conditions a polynomial P which is non-negative (or positive) on an algebraic set, i.e., defined by polynomial inequalities of the form $Q_i \geq 0$ for polynomials Q_i, can be written in the form $P = P_0 + \sum_{i=1}^N P_i Q_i$ where the P_i are SoS. The theoretical literature regroups these results under the name "Positivstellensatz". The most often seen in the SoS optmization literature are the Stengle (1974); Schmügden (1991); Putinar (1993) Positivstellensätzen.

If these algebraic geometry considerations seem far from applications and from decomposing smooth functions as sums of squares (indeed, polynomials are much more rigid than smooth functions) at first glance, they are actually related in two ways.

First, as smooth functions can be locally approximated by polynomials, results on polynomials give a good intuition of the difficulties one can encounter at the local level when decomposing a function as a sum of squares. Indeed, on the one hand, the general impossibility results proved by Bony (2005); Bony, Broglia, Colombini, and Pernazza (2006) (see Theorem 9.4) are obtained using Hilbert's theorem on the existence of non-negative polynomials which are not sum of squares. On the other hand, the fact that there is hope using our second-order assumptions is also due to the fact that second order non-negative polynomials can always be written as sums of squares.

Second, the certificates given by Positivstellensatz on the decomposability of certain non-negative polynomials can be algorithmically checked in some cases, using semi-definite programming. This has paved the way to so-called SoS hierarchies, and optimization of polynomial objective functions with polynomial constraints. These have been developed by Lasserre (Lasserre, 2010) (based on

the Postivstellensatz by Putinar (1993)) and Parrilo (2003) (based on the Positivstellensatz by Stengle (1974); Schmügden (1991)). Using these theoretical results, they can provide certificates of lower bounds for certain optimization problems (or upper bound in the dual "moment problem", see the work by Lasserre (2010)). Moreover, to have more interpretable results for these more applied settings, theses works have motivated more practical Positivstellensatz, like that by Marshall (2006), which provides a condition for writing a polynomial with a finite set of zeros as as sum of squares (this condition is actually a second order condition which greatly resembles ours in the polynomial setting, although it deals more with the constraints Q_i).

p-times differentiable functions In the same spirit as the polynomial hierarchies, recent works by Marteau-Ferey, Bach, and Rudi (2020); Rudi, Marteau-Ferey, and Bach (2020); Rudi and Ciliberto (2021) have developed models and methods based on sum of squares of regular functions. The computational properties of these methods are based on the fact that regular functions can be well-approximated by functions of the form $\sum_i \alpha_i k(\cdot, x_i)$ where k is a so-called positive definite kernel (Aronszajn, 1950) and can be adapted to the regularity. In order to obtain guarantees on these methods, it is crucial to have the equivalent of Positivstellensatz in the case of regular functions. Contrary to the case of algebraic geometry, where such results existed for other purposes, there is a need to build such results for regular functions from scratch. Certain results like Theorem 9.1 have been presented. However, the aim of the present paper is to provide more general results, to be used in most situations.

Organisation of the work

In Sec. 9.2 , we formalize the different notions needed to state Theorem 9.2 in the case of non-negative functions defined on open sets of $\mathbb{R}^d$. In particular, we start by presenting a local decomposition in Theorem 9.5, which will be the cornerstone of the work. In Sec. 9.3 , we extend Theorem 9.2 to the manifold setting, and detail the procedure in which we glue local decompositions into a global one, using traditional tools from differential geometry. In Sec. 9.4 , we formally prove Theorem 9.5. We finish by a discussion on the result presented in this paper, as well as possible extensions in Sec. 9.5 .

9.2 Decomposition as sums of squares given second order conditions (Euclidean case)

In this section, we present our results on decomposing a C^p function f as a sum of squares of C^{p-2} functions on open sets of $\mathbb{R}^d$.We start with a brief presentation of the notion of sub-manifold of $\mathbb{R}^d$ in Sec. 9.2 .1. It is the key geometric object we use to represent the set of zeros $\mathcal{Z}$ of the function f. In Sec. 9.2 .2, we present the cornerstone result of this paper in Theorem 9.5, as well as a sketch of its proof, which is done extensively in Sec. 9.4 . It shows that as soon as a non negative function has positive Hessian in the orthogonal direction to its zeros at a given point, then it can be decomposed as a sum of squares around that point. Finally, in Sec. 9.2 .3, we present Theorem 9.6, which shows that given a function defined on an open subset Ω of the Euclidean space $\mathbb{R}^d$, and under conditions on the Hessian of f at its zeros, f can be decomposed as a locally finite sum of squares of functions defined on Ω.

Definitions and notations

In general, given two topological sets M and N as well as $x_0 \in M$ and $y_0 \in N$, we will say that $\phi : (x_0, M) \to (y_0, N)$ is a local map satisfying a property (P) if there exists an open neighborhood

U of x_0 in M such that $\phi : U \to N$ is well defined, satisfies $\phi(x_0) = y_0$ and property (P). We will say that $\phi : U \subset \mathbb{R}^d \to \mathbb{R}^e$ defined on an open set U is of class C^k if it is k times differentiable, and its derivatives of order k are continuous. For any function $\phi : (x, \mathbb{R}^d) \to \mathbb{R}^e$ of class C^1, we denote with $d\phi(x)$ its differential at x. It is an element of $\mathrm{Hom}(\mathbb{R}^d, \mathbb{R}^e)$ the set of linear maps from $\mathbb{R}^d$ to $\mathbb{R}^e$. We will write $d\phi(x)\xi$ or $d\phi(x)[\xi]$ the evaluation of $d\phi(x)$ at ξ. The Jacobian of ϕ at x is the matrix $J_\phi(x) \in \mathbb{R}^{e\times d}$ which is the matrix of $d\phi(x)$ in the canonical bases. Writing the coordinates of ϕ: $\phi = (\phi^1, ..., \phi^e)$, we have $[J_\phi]_{ij} = \frac{\partial \phi^i}{\partial x^j}(x)$.

9.2 .1 Sub-manifolds of $\mathbb{R}^d$

One of the main assumptions in order to achieve our results will be that the set of zeros of the non-negative function f is a sub-manifold of $\mathbb{R}^d$. In this section, we restrict ourselves to introducing the definitions and results needed to state and prove those in this paper. For a more comprehensive introduction, see chapter 1 by Lafontaine (2015), section 2.2 by Paulin (2006) (in French) or Spivak (1999). The notion of sub-manifold generalizes the notion of a curve in $\mathbb{R}^d$ (a one dimensional manifold) or a surface in $\mathbb{R}^d$ (a two dimensional manifold). Intuitively, a sub-manifold N is a subset of $\mathbb{R}^d$ such that at each point $x \in N$, N "looks like" $\mathbb{R}^{d_0}$ where d_0 is the dimension of the sub-manifold at x (one for a line, two for a surface,etc.). Another way to put this is that N can be locally parametrized by $\mathbb{R}^{d_0}$. To formalize this, we need the following definitions. We fix a subset $N \subset \mathbb{R}^d$.

A map $\phi : U \to \mathbb{R}^d$ defined on an open neighborhood U of 0 in $\mathbb{R}^{d_0}$ is said to be a local parameterization of N around x_0 of class C^k for $k \geq 1$ if ϕ is of class C^k, and if there exists an open set $V \subset \mathbb{R}^d$ such that the following conditions are satisfied:

(i) $\phi(0) = x_0$, $\phi(U) = N \cap V$, and $\phi : U \to \phi(U)$ is a homeomorphism, i.e., it is bijective and has continuous inverse;

(ii) its differential at 0 is injective (one to one), i.e., $d\phi(t_0) \in \mathrm{Hom}(\mathbb{R}^{d_0}, \mathbb{R}^d)$ is injective.

The second condition guarantees that the local dimension of N is indeed d_0, that ϕ is not an over-parameterization. N is said to be a sub-manifold of $\mathbb{R}^d$ and of class C^k if there exists a local parameterization ϕ of class C^k around each point $x \in N$. Given a point $x \in N$, the dimension d_x of the local parametrization is independent of the parametrization (two local parametrizations will necessarily be of same dimension); it is called the dimension of N at x. Similarly, the subspace $T_xN := d\phi(x)\mathbb{R}^{d_x}$, which is a subspace of $\mathbb{R}^d$ of dimension d_x is independent of the local parametrization: it is the linear approximation of N at x and is called the tangent space to N at x (see Sec. 9.2 .1 and Sec. 9.2 .2 for more visual representations).

A sub-manifold N of $\mathbb{R}^d$ is said to be connected if it cannot be written as a union of disjoint open sets. Equivalently, it is connected if any two points in N can be connected by a continuous path $\gamma : [0, 1] \to N$. On a connected sub-manifold N, the dimension d_x is the same at every point x, it is called the dimension of the connected sub-manifold N. This implies that all the tangent spaces T_xN have the same dimension.

Example 9.1. *All open sets of $\mathbb{R}^d$ are sub-manifolds of $\mathbb{R}^d$. The d-dimensional sphere S^d is a sub-manifold of $\mathbb{R}^{d+1}$. S^1 is represented in the left hand side (l.h.s.) of Sec. 9.2 .1 and S^2 in the l.h.s. of Sec. 9.2 .2. Given a sub-manifold N of R^d, the intersection of N with any open set of $\mathbb{R}^d$ is a sub-manifold of $\mathbb{R}^d$.*

If U_i is a family of disjoint open sets each containing a connected sub-manifold N_i of $\mathbb{R}^d$, it is clear the the disjoint union $\sqcup_{i\in I} N_i$ is also a sub-manifold. Conversely, any sub-manifold can be

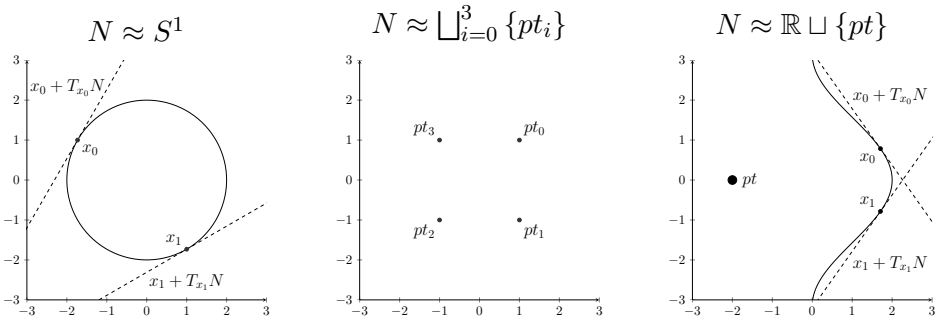

Figure 9.2: Examples of sub-manifolds of $\mathbb{R}^2$; points are denoted with pt. **Left:** connected sub-manifold of dimension 1 (a circle). **Center:** a sub-manifold of 4 connected components which are all points, i.e., of dimension 0 (their tangent space is not represented since it is reduced to $\{0\}$). **Right:** a sub-manifold of two connected components, one point pt of dimension 0 and one of dimension 1.

decomposed into its connected components N_i; moreover, one can find a family of disjoint open sets U_i such that $N_i \subset U_i$ (see Lemma 9.9).

These results, their proof and their broader context can be found in chapter 1.5 by Lafontaine (2015). In particular, Theorem 1.21 presents equivalent definitions of a sub-manifold. Section 2.2 by Paulin (2006) is also a good reference (in French).

9.2 .2 Local decomposition as a sum of squares

In this section, f will always denote a non-negative function defined on an open set of $\mathbb{R}^d$. We will assume that f is of class C^p for $p \geq 2$. We will also denote with $\mathcal{Z}$ the set of zeros of f, i.e., the set of zeros of f. In this section, we will make local assumptions on the Hessian of f at points $x \in \mathcal{Z}$ such that the function f can be decomposed as a sum of squares locally around x.

We will denote with $d^2 f(x)$ the second differential of f (which we will sometimes call abusively its Hessian), which is a symmetric bilinear form on $\mathbb{R}^d$. We denote with $d^2 f(x)[\xi, \eta]$ its evaluation on vectors ξ, η. We denote with $\nabla^2 f(x) \in \mathbb{R}^{d \times d}$ the Hessian matrix of f at x, which is the matrix of $d^2 f(x)$ in the canonical basis of $\mathbb{R}^d$, and we have $d^2 f(x)[\xi, \eta] = \eta^\top \nabla^2 f(x) \xi$. For any vector sub-space space $S \subset \mathbb{R}^d$, and any bilinear form H on $\mathbb{R}^d$, we denote with $H|_S$ the restriction of H to S, which is a bilinear form on F. We say that a bilinear form H is positive semi-definite if $H[\xi, \xi] \geq 0$ for any $\xi \in \mathbb{R}^d$, and is positive definite if $H[\xi, \xi] > 0$ for all $\xi \in \mathbb{R}^d \setminus \{0\}$. We use the same terminology for matrices.

We are now ready to state Theorem 9.5, which is the cornerstone of this work. For the rest of this section (Sec. 9.2 .2), let $x_0 \in \mathbb{R}^d$ and $f : (x_0, \mathbb{R}^d) \to \mathbb{R}$ be a non-negative C^p function, for $p \geq 2$, such that $f(x_0) = 0$. We claim that if there is a sub-manifold of class C^1 and of dimension d_0 around x_0 of zeros of f, and if the Hessian of f at x_0 has rank $d - d_0$ (which we will call the normal Hessian condition), then it can be decomposed as a sum of squares as in Eq. (9.1).

Definition 9.1 (normal Hessian condition)**.** *Let $\mathcal{Z}$ denote the set of zeros of f. We say that f satisfies the* normal Hessian condition *(NHC) at x_0 if there exists a dimension $0 \leq d_0 \leq d$ and a sub-manifold N of class C^k with $k \geq 1$ and of dimension d_0 such that $x_0 \in N \subset \mathcal{Z}$, and on one of the following equivalent conditions is satisfied:*

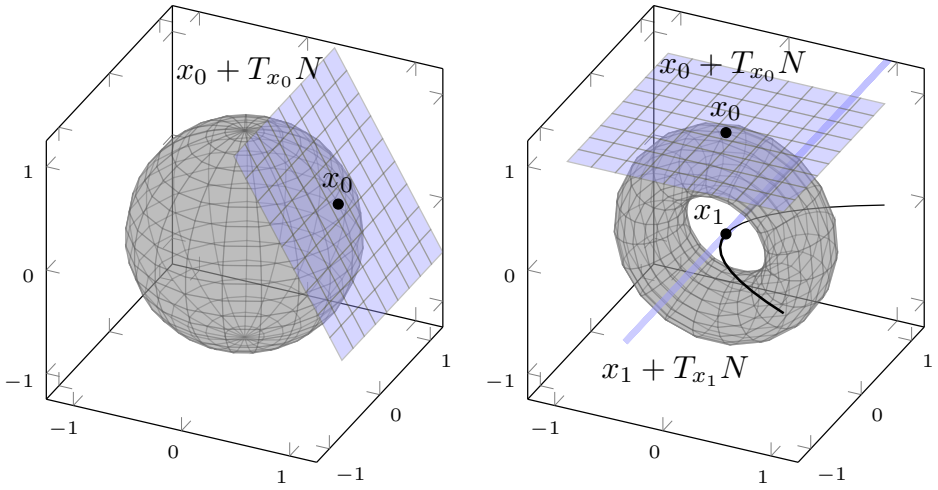

Figure 9.3: Two examples of sub-manifolds of $\mathbb{R}^3$. The blue affine spaces represent tangent spaces. **Left:** connected sub-manifold of dimension 2 (the sphere S^1). **Right:** a sub-manifold of two connected components, one of dimension 2 (homeomorphic to the torus $\mathbb{T}^2$ on which lies x_0), and one of dimension 1 on which lies x_1.

(i) the rank of $\nabla^2 f(x_0)$ *at* x_0 *is* $d - d_0$*;*

(ii) the restriction of $d^2 f(x_0)$ *to* $T_{x_0}N^\perp$ *is positive definite.*

The complete proof of the equivalence of these conditions as well as the proof of Theorem 9.5 can be found in Sec. 9.4 . To illustrate the definition of the normal Hessian condition, we refer to Sec. 9.2 .2 which represents the local behavior of functions f defined locally around a point $x_0 \in \mathbb{R}^2$ in the set of zeros and which satisfies the NHC for $d_0 = 1$.

Theorem 9.5. *If* f *satisfies the NHC at* x_0 *(definition 9.1) with regularity* k *and dimension* d_0*, there exists an open neighborhood* U *of* x_0 *in* $\mathbb{R}^d$ *on which* f *is defined and such that* $U \cap \mathcal{Z}$ *is a sub-manifold of* $\mathbb{R}^d$ *of dimension* d_0 *and of class* $C^{\max(k,p-1)}$*, and there exist functions* $f_i \in C^{p-2}(U)$ *where* $1 \leq i \leq d - d_0$ *such that*

$$\forall x \in U,\ f(x) = \sum_{i=1}^{d-d_0} f_i^2(x). \tag{9.5}$$

Main steps of the proof. The main steps of this proof are represented geometrically in Sec. 9.2 .2.

Step 1. We show that under the NHC at x_0, we have $T_{x_0}N = \ker(\nabla^2 f(x_0))$ and hence that $d^2 f(x_0)|_{T_{x_0}N^\perp}$ is positive definite.

Step 2. Re-parametrizing f on a basis adapted to $T_{x_0}N^\perp \oplus T_{x_0}N$ as $f(x_\perp, x_{\shortparallel})$, we apply the Morse lemma (see Lemma 9.10), which decomposes the function f in the form

$$f(x_\perp, x_{\shortparallel}) = f(\varphi(x_{\shortparallel}), x_{\shortparallel}) + \tfrac{1}{2} d^2 f(x_0)|_{T_{x_0}N^\perp}[\xi(x_\perp, x_{\shortparallel}), \xi(x_\perp, x_{\shortparallel})], \tag{9.6}$$

for a certain function φ of class C^{p-1} and ξ of class C^{p-2} in a certain open set around x_0 (for an easy visualization, see Sec. 9.2 .2).

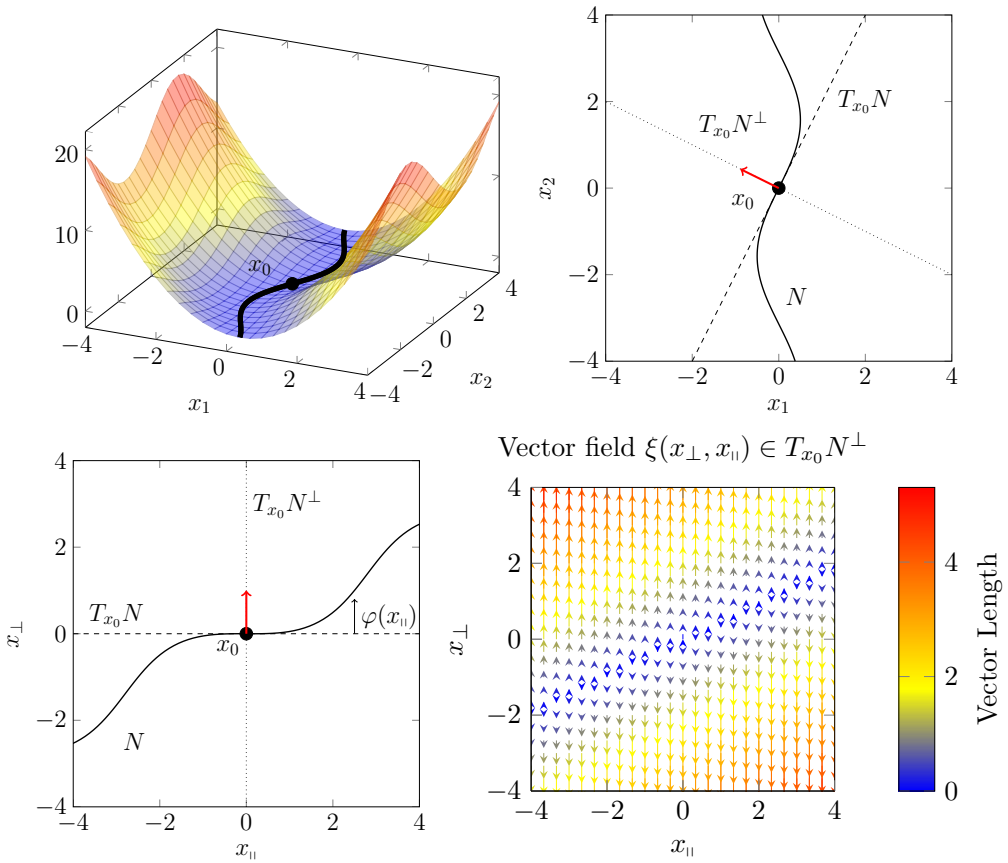

Figure 9.4: Local view of the function around a minimum lying on a 1-dimensional manifold. **Top left:** function around the minimum x_0. **Top right:** decomposition of $\mathbb{R}^2$ at x_0 between tangent space and normal tangent space $T_{x_0}N + T_{x_0}N^\perp$, and positive eigen-vector of the Hessian in red. **Bottom left:** reparametrization in the right coordinate system, and representation of the map $\varphi(x)$ given by the Morse Lemma. **Bottom right:** vector field ξ given by the Morse lemma.

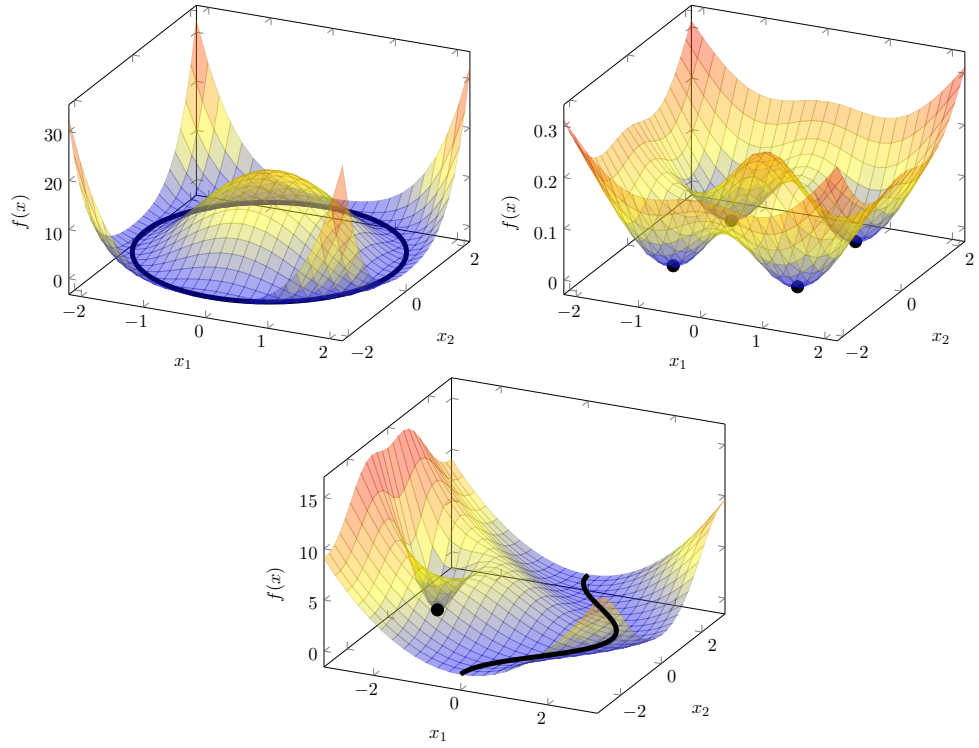

Figure 9.5: Example of functions f which satisfy the global normal Hessian condition, with sub-manifolds $\mathcal{Z}$ of zeros corresponding to the sub-manifolds presented in Sec. 9.2 .1 in the same order

Step 3. We show that the second term of the right hand side of Eq. (9.6) can actually be seen as a sum of squares of $d - d_0$ functions of class C^{p-2}.

Step 4. We characterize the manifold of zeros around x_0.

Step 5. We show that the first term of the result of the Morse lemma is equal to zero using the previous characterization, which shows Eq. (9.5). □

Example 9.2 (case where $d_0 = 0$)**.** *When $d_0 = 0$, the NHC at x_0 is simply the SHC Eq. (9.3), that is the condition that x_0 be a strict minimum. In that case, Theorem 9.5 simply states that there exists an open neighborhood U of x_0 such that $U \cap \mathcal{Z} = \{x_0\}$ and on which f can be decomposed as the sum of d squares.*

Remark 30 (Smoothing effect)**.** *Note that Theorem 9.5 induces a smoothing effect: indeed, if we simply assume that there exists a d_0 dimensional manifold of class C^1 of zeros satisfying the NHC, one sees that this manifold is actually of class C^{p-1} in a neighborhood of x_0.*

9.2 .3 Global decomposition as a sum of squares for functions on $\mathbb{R}^d$

In this section, we fix f to be a non-negative C^p function defined on an open subset Ω of $\mathbb{R}^d$. Once again, we assume $p \geq 2$. The goal is to find conditions on f to be written as a sum of squares of functions defined on Ω. These conditions will be that the NHC holds at every $x_0 \in \mathcal{Z}$. We will start by reformulating this assumption in a more global and geometric way. We introduce the following definition of a manifold to which f is positively normal.

Definition 9.2 (positive normally to a sub-manifold)**.** *Let N be a sub-manifold of $\mathbb{R}^d$ of class C^k for $k \geq 1$ and included in Ω. We say that f is positive normally to N if:*

a) N is included in the set of critical points of f ($df(x) = 0$ for all $x \in N$);

b) for any $x_0 \in N \cap \Omega$, if d_0 is the local dimension of N at x_0, there exists a subspace $S \subset \mathbb{R}^d$ of dimension $d - d_0$ such that $d^2 f(x_0)|_S$ positive definite.

The intuition of this definition is that if f is positively normal to N, then f grows quadratically normally to N, which is a local minimum valley on which f is constant. Note that there can be more than one connected component in N: this will correspond to multiple local minima valleys (see second and third examples in Sec. 9.2 .3). We now reformulate the fact that the NHC holds at every point in $\mathcal{Z}$ as a more geometric global assumption, using definition 9.2.

Lemma 9.1 (Global normal Hessian condition)**.** *The following statements are equivalent and define the global NHC condition:*

(i) for all $x_0 \in \mathcal{Z}$, f satisfies the NHC;

(ii) $\mathcal{Z}$ is a sub-manifold of $\mathbb{R}^d$ (not necessarily connected) of class C^1 such that the Hessian of f is positive normally to $\mathcal{Z}$;

(iii) $\mathcal{Z}$ is a sub-manifold of $\mathbb{R}^d$ (not necessarily connected) of class C^{p-1} such that the Hessian of f is positive normally to $\mathcal{Z}$.

This equivalence is a direct consequence of the local description of $\mathcal{Z}$ obtained in Theorem 9.5 under the local NHC. Examples of functions satisfying the global normal Hessian condition can be found in Sec. 9.2 .3. They have manifolds of zeros which are depicted in the same order in Sec. 9.2 .1. Under this geometric condition, we will show in Theorem 9.6 that f can be written as a sum of squares of C^{p-2} functions with locally finite support, defined below.

Locally finite support. Let X be a topological space (see the book by Jänich (1980) for full definitions). We say that a family (S_i) of subsets of X is locally finite if for every $x \in X$, there exists an open set U_x containing x which intersects a finite number of the S_i, i.e., $|\{i \in I \ : \ U_x \cap S_i \neq \emptyset\}| < \infty$. A family (f_i) of functions on a topological space X has locally finite support if the family of supports $(\mathrm{supp}(f_i))_{i \in I}$ is locally finite (recall that $\mathrm{supp}(f_i) = \overline{\{x \ : \ f(x) \neq 0\}}$). In particular, if (f_i) has locally finite support, the function $\sum_{i \in I} f_i^2$ is well defined and it is also of class C^q if the functions are of class C^q. Using this terminology, the global result can be stated as follows.

Theorem 9.6. *If f satisfies the global normal Hessian condition in Lemma 9.1, there exists an at most countable family $(f_i)_{i \in I} \in (C^{p-2}(\Omega))^I$ with locally finite support such that*

$$\forall x \in \Omega, \ \ f(x) = \sum_{i \in I} f_i(x)^2. \tag{9.7}$$

Moreover:

- *if f satisfies the strict Hessian condition, $\mathcal{Z}$ is discrete and we can find such a decomposition such that $|I| \leq d + 1$.*
- *if $\mathcal{Z}$ is compact, then $|I|$ can be taken to be finite.*

For the formal proof of this result, we refer to the next section, where this result will be proved more generally for functions defined on manifolds (see Theorem 9.8 and Sec. 9.3).

Main steps of the proof. For subtleties pertaining to the SHC case, we refer to Sec. 9.3 . The gluing done in that section is slightly more elaborate.

Step 1. Since the local NHC holds at any point in $\mathcal{Z}$, using Theorem 9.5 shows that at any point x, there exists an open neighborhood U_x of x , an integer n_x and functions $(f_{x,j})_{1\leq j\leq n_x}$ of class C^{p-2} on U_x such that $f = \sum_{j=1}^{n_x} f_{x,j}^2$ on U_x. The collection of sets U_x is then an open covering of $\mathcal{Z}$. Since $\mathbb{R}^d$ is Hausdorff and second-countable (see Sec. 9.3 for precise definitions), only an at most countable subsets of them are necessary to cover $\mathcal{Z}$ (even a finite number if $\mathcal{Z}$ is included in a compact, since it is then itself a compact). Denote with $(U_i)_{i\in I}$ this open covering, and replace x by i to denote the associated $f_{i,j}$ and n_i.

Step 2. Since $\mathcal{Z}$ is closed, as the set of zeros of a continuous function, the set $U_{>0} := \{x \in \Omega \ : \ f(x) > 0\}$ is open and the map $f_1 := \sqrt{f} : U_{>0} \to \mathbb{R}$ is of class C^p and satisfies $f_1^2 = f$. We can therefore add $U_{>0}$ to the collection (U_i) and still guarantee the following property: for all $i \in I$, there exists $n_i \in \mathbb{N}$ and $f_{i,j} \in C^{p-2}(U_i)$ such that $f = \sum_{j=1}^{n_i} f_{i,j}^2$. Moreover, (U_i) becomes an open covering of Ω; in particular, if U_i was a finite covering of $\mathcal{Z}$, it now becomes a finite covering of Ω.

Step 3. Using Lemma 9.4, we can take a partition of unity (χ_i) adapted to the open covering $\bigcup_i U_i$ such that $\sum_i \chi_i^2 = 1$ and which is locally finite. Define $\tilde{f}_{i,j} = f_{i,j}\chi_i$ which is now defined on the whole of Ω (indeed, it can be extended as zero to $\Omega \setminus U_i$ since the support ot χ_i is included in U_i). The $\tilde{f}_{i,j}$ satisfy $\sum_{i,j} \tilde{f}_{i,j}^2 = f$ on the whole of Ω, and is a finite family if the covering U_i is finite (if $\mathcal{Z}$ is assumed to be compact for example). □

9.3 Global decomposition as a sum of squares for functions on manifolds

In this section, we present results analogous to those of Sec. 9.2 but in the more general context of manifolds. After a brief recap on the terminology and definitions related to manifolds, in Sec. 9.3 .1, we will adapt the definitions of the local and global normal Hessian conditions, as well as state the equivalent result to Theorem 9.5 in the context manifolds. In Sec. 9.3 .2, we will introduce the tools to glue local decompositions as sum of squares together. Finally, in Sec. 9.3 .3, we prove Theorem 9.8, the equivalent of Theorem 9.6 in the broader context of manifolds.

Additional definitions and notations for manifolds

In this section, we introduce the basic definitions we will need concerning manifold. For more formal introductions to manifolds, we refer to the works by Lafontaine (2015); Paulin (2006); Spivak (1999). Informally, a manifold of dimension d is a set which "looks like $\mathbb{R}^d$" locally. This means that at every point $x \in M$, we can find a chart ϕ which topologically maps a neighborhood U of x to an open set of $\mathbb{R}^d$.

More generally, we define a chart on a topological space M as a map $\phi : U \to \mathbb{R}^d$ for some $d \in \mathbb{N}$, defined on an open set U of M, and which is a homeomorphism onto its image. We define a manifold M as a second-countable[1], Hausdorff[2] topological space equipped with a collection $\mathcal{A} = (\phi_i)_{i\in I}$ of charts such that

[1] A topological space is said to be second countable if there exists a countable sequence of open sets U_n such that any open set U in the topology is a reunion of a part of the U_n.

[2] A topological space is Hausdorff if for any two points $x \neq x'$, there exists two open sets U, V such that $x \in U$ and $x' \in V$ and $U \cap V = \emptyset$

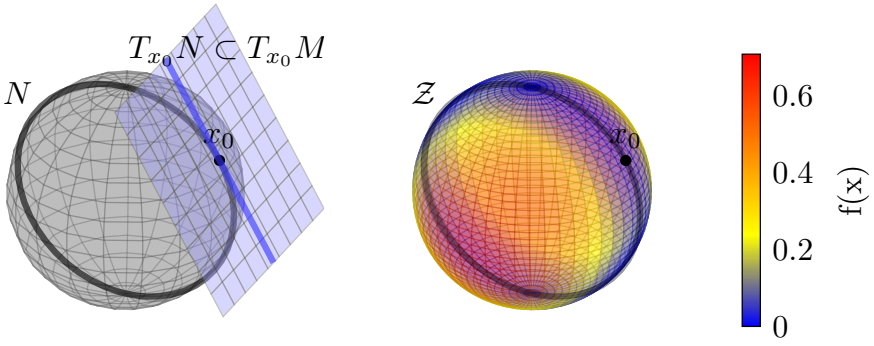

Figure 9.6: **Left:** Representation of the manifold $M = S^2$ as well as a sub-manifold N homeomorphic to a circle. The tangent spaces at a given point $x_0 \in N \subset M$ are represented as well. **Right:** Representation of a non-negative function on the sphere as a color map; it satisfies the NHC, and its null space $\mathcal{Z}$ is represented in black.

(i) all transition maps $\phi_i \circ \phi_j^{-1} : \phi_j(U_j \cap U_i) \to \phi_i(U_i \cap U_j)$ are homeomorphisms;

(ii) the charts cover M entirely, i.e. $M = \bigcup_{i \in I} U_i$.

The set $\mathcal{A}$ is called an atlas. The manifold M is said to be of class C^k for $k \geq 0$ if all the transition maps $\phi_i \circ \phi_j^{-1}$ are of class C^k. It is said to be of dimension d if all its charts are in $\mathbb{R}^d$. As for sub-manifolds of $\mathbb{R}^n$, a manifold can always be decomposed as the union of its connected components, and the dimension is the same on each connected component. Note that with this definition, the restriction of a manifold to any open set is still a manifold (just by restricting the charts).

If M is a manifold of class at least C^1, we can define at each point its tangent space T_xM. Informally, this set T_xM is all the possible derivatives $\gamma'(0)$ of curves $\gamma : I \to M$ defined on an open interval I around 0 such that $\gamma(0) = x$. Of course $\gamma'(0)$ is not yet formally defined. Formally, T_xM can be defined as the classes of C^1 curves defined on an open interval I around 0 such that $\gamma(0) = x$, where we identify two curves γ and $\tilde{\gamma}$ if $(\phi \circ \gamma)'(0) = (\phi \circ \tilde{\gamma})'(0)$ for a (or equivalently any) chart ϕ of M around x. We denote with $[\gamma] \in T_xM$ the equivalence class of a curve γ. It can be shown that T_xM is a vector space (with the natural definition $\lambda[\gamma] + \mu[\tilde{\gamma}] = [\lambda\gamma + \mu\tilde{\gamma}]$) and that it is of dimension d where d is the dimension of M at x.

A map $g : M \to \mathbb{R}^p$ defined on a manifold M is said to be of class C^q if M is of class at least C^q and if for any chart ϕ, the map $g \circ \phi^{-1}$ is of class C^q. If $q \geq 1$, then we can define the differential of g at any point $x \in M$ as $dg(x)[\gamma] = (f \circ \gamma)'(0)$. Hence, $dg(x) \in \mathrm{Hom}(T_xM, \mathbb{R}^p)$.

Example 9.3. *All sub-manifolds of $\mathbb{R}^d$ are manifolds. The notions of regularity, dimension, and tangent space coincide.*

For more precise definitions of topological spaces, atlases, charts, and details on the "Hausdorff second-countable" condition, see for instance the works by Lafontaine (2015); Paulin (2006); Jänich (1980); Spivak (1999). The main idea behind the introduction of manifolds as opposed to sub-manifolds of $\mathbb{R}^d$ is to consider the intrinsic geometric object, and not its relation to the euclidean space it is embedded in (as such an embedding is not unique). An example of manifold as well as a representation of the tangent space is provided in the left of Sec. 9.3 .

9.3 .1 Assumptions in the manifold case

In this section, we formulate the local NHC in the case of manifolds, and rewrite Theorem 9.5 in this setting. We also extend the definitions of being positively normal and the global NHC (Lemma 9.1).

Fix $p \in \mathbb{N},\ p \geq 2$, and a manifold M of regularity at least C^p and of dimension $d \in \mathbb{N}$. To start with, let $f : \Omega \to \mathbb{R}$ be a non-negative function defined on an open set of M and of class C^p. As before, define $\mathcal{Z}$ to be the set of zeros of f.

Contrary to the $\mathbb{R}^d$ case, the second differential of the function f cannot be identified to a symmetric bi-linear form everywhere. However, it is the case at so-called critical points, i.e., points $x \in \Omega$ such that $df(x) = 0$. In particular, since all the zeros of a C^1 non-negative function are critical points, this Hessian will be defined at all points in the set of zeros $\mathcal{Z}$ of f.

Lemma 9.2 (definition of the Hessian)**.** *Let x be a critical point of f. Then there exists a unique symmetric bi-linear form $H_f(x) : T_xM \times T_xM \to \mathbb{R}$ such that for any local chart $\phi : (M, x) \to (\mathbb{R}^d, 0)$ it holds:*

$$\forall \xi, \eta \in T_xM \times T_xM,\ H_f(x)[\xi, \eta] = d^2(f \circ \phi^{-1})(0)[d\phi(x)\xi, d\phi(x)\eta].$$

In order to prove this lemma, we simply define the bilinear form as such for a given chart ϕ around x, and then show that this definition does not depend on the chart ϕ using the fact that x is a critical point. This is completely proved in section 2 of Milnor (1963). In order to formulate the definition of the normal Hessian condition in the setting of manifolds, we further need a definition of what a sub-manifold of M is. A subset $N \subset M$ is said to be a sub-manifold of M of class C^k if M is of class C^k and if, for any $x \in N$ and any local chart $\phi : U \to \mathbb{R}^d$ defined on a neighborhood of x, $\phi(U \cap N)$ is a sub-manifold of $\mathbb{R}^d$ of class C^k. In the literature, this is also called a proper sub-manifold (see on the left of Sec. 9.3 for an example).

Definition 9.3 (Normal Hessian condition for a manifold)**.** *Let $x \in \Omega$ be a point in the domain of f. We say that f satisfies the normal Hessian condition (NHC) at x if there exists a dimension $0 \leq d_0 \leq d$ and a sub-manifold N of M of class C^k with $k \geq 1$ and of dimension d_0 such that $x \in M \subset \mathcal{Z}$, and one of the following equivalent conditions is satisfied:*

(i) the rank of $H_f(x)$ is $d - d_0$;

(ii) $H_f(x)[\xi, \xi] > 0$ for any vector $\xi \in T_xM \setminus T_xN$.

Using any local chart around the point x in the domain of f, one can apply Theorem 9.5 to obtain the following theorem as a corollary.

Theorem 9.7. *If f satisfies the NHC at x_0 with regularity k and dimension d_0, there exists an open neighborhood U of x_0 in M on which f is defined and such that $U \cap \mathcal{Z}$ is a sub-manifold of M of dimension d_0 and of class $C^{\max(k,p-1)}$; and there exist functions $f_i \in C^{p-2}(U)$ where $1 \leq i \leq d - d_0$ such that*

$$\forall x \in U,\ f(x) = \sum_{i=1}^{d-d_0} f_i^2(x). \tag{9.8}$$

Exactly in the same way as for the definition of the NHC for manifolds, we can similarly extend the definition of a function being positively normal to a sub-manifold in definition 9.2 as well as the global NHC in Lemma 9.1. We will therefore say that $f : M \to \mathbb{R}$ which is non-negative satisfies the global NHC if it satisfies the local NHC at every point in its set of zeros $\mathcal{Z}$, or

equivalently if it is positive normally to $\mathcal{Z}$ which is a sub-manifold of M of class C^1 (or C^{p-1}). On the right hand side of Sec. 9.3 , we represent a function which satisfies the NHC on the sphere S^2 through a colormap, with a continuous set of zeros. The goal is to prove that such a function can be decomposed as a sum of squares on S^2.

9.3 .2 Gluing local decompositions to form a global one

In this section, we present and develop the tools to glue local decompositions such as Theorem 9.7 into a global one, which will lead to Theorem 9.8.

The first result we need is a simple result to "extend" a function defined on an open set U of M to M by multiplying it by a function defined on M whose support lies in U (Lemma 9.3). The second one is a variant of the fundamental result of existence of partitions of unity on a manifold, adapted to our sum of squares setting (Lemma 9.4). Recall that the support of a function has been defined in Sec. 9.2 .3. The proof of these results can be found in Sec. 9.A .1.

Lemma 9.3 (Extension lemma)**.** *Let $q \in \mathbb{N}$, M be a manifold of class at least C^q. Let U be an open set of M, $g : U \to \mathbb{R}$ be a C^q function defined on U, and $\chi : M \to \mathbb{R}$ be a C^q function defined on the whole of M but with support included in U. Then the function $\chi g : U \to \mathbb{R}$ extended as 0 on $M \setminus U$, is of class C^q on the whole of M and has support included in $\mathrm{supp}(\chi) \subset U$. We still denote with χg its extension to M.*

Lemma 9.4 (Gluing lemma)**.** *Let $(U_i)_{i\in I}$ be an open covering of a manifold M of class C^k (i.e. $\bigcup_{i\in I} U_i = M$). There exists a family of functions $\chi_i : M \to [0, 1]$ of class C^k with locally finite support, such that $\mathrm{supp}(\chi_i) \subset U_i$ for all $i \in I$ and satisfying:*

$$\sum_{i\in I} \chi_i^2 = 1.$$

We can now proceed from local to global in two steps. First, we use the gluing lemma to glue decompositions in a single connected component of the manifold of zeros (Lemma 9.5). We then glue these different decompositions into a single global one (Lemma 9.6).

Lemma 9.5. *Assume f satisfies the global NHC. Let N be a connected component of its manifold of zeros $\mathcal{Z}$. There exists an open neighborhood U of N as well as a locally finite, at most countable family $(f_j)_{j\in J}$ of functions of class C^{p-2} such that*

$$\forall x \in U,\ f(x) = \sum_{j\in J} f_j(x)^2. \tag{9.9}$$

Moreover, we can find J such that a) $|J| = d$ if $N = \{x_0\}$ is a single point and b) J is finite if N is compact.

Proof. The case where $N = \{x_0\}$ is simply Theorem 9.7 applied to x_0. In the other cases, note that for all $x \in N$, by Theorem 9.5 since the NHC is satisfied at x, there exists an open neighborhood U_x of x as well as functions $(f_{x,i})_{1\le i\le d}$ of class C^{p-2} such that $f = \sum_{i=1}^d f_{x,i}^2$ on U_x. Since $(U_x)_{x\in N}$ covers N, we can extract a covering $(U_{x_j})_{j\in J}$ of N such that *a) J* is finite if N is compact and *b) J* is at most countable otherwise, since N is second-countable and Hausdorff. Denote with $(U_j)_{j\in J}$ this open covering, and replace x by j to denote the associated $f_{j,i}$. Denote with U the open set $\bigcup_j U_j$.

Applying Lemma 9.4 to the manifold U, we can find a family of functions $(\chi_j)_{j\in J}$ with locally finite support, such that $\mathrm{supp}(\chi_j) \subset U_j$ and $\sum_j \chi_j^2 = 1$ on U. By the extension Lemma 9.3, we

can therefore define the functions $\widetilde{f}_{j,i} := \chi_j \ f_{j,i}$ for $i \in \{1, ..., d\}$ and $j \in J$ which are defined on the whole of M. Note that since $\text{supp}(\tilde{f}_{j,i}) \subset \text{supp}(\chi_j)$ and since $1 \leq i \leq d$, the support of $(\tilde{f}_{j,i})$ is also locally finite. To conclude, we use the property that $\sum_j \chi_j^2 = 1$ on U as well as the fact that $\sum_i f_{j,i}^2 = f$ on $\text{supp}(\chi_j) \subset U_j$ to show that $\sum_{i,j} \tilde{f}_{j,i}^2 = f$ on U. The number of functions $\tilde{f}_{j,i}$ is finite if N is compact since J is finite, and is at most countable else since J is at most countable. □

Lemma 9.6. *Let $\mathcal{Z} = \sqcup_{i\in I} N_i$ be the manifold of zeros decomposed along its connected components. Assume that there exists an index set J, such that for all $i \in I$, there exists an open neighborhood U_i of N_i on which f can be decomposed as a sum of squares indexed by J:*

$$\forall i \in I, \ \exists (f_{i,j})_{j\in J} \in (C^{p-2}(U_i))^J, \ \forall x \in U_i, \ f(x) = \sum_{j\in J} f_{i,j}(x)^2, \tag{9.10}$$

and such that the families $(f_{i,j})_{j\in J}$ are all locally finite. Then there exists a locally finite family $(g_j)_{j\in J\cup\{\star\}}$ of C^{p-2} functions on M (we add an extra element $\star$ to J), such that

$$\forall x \in M, \ f(x) = \sum_{j\in J\cup\{\star\}} g_j(x)^2. \tag{9.11}$$

Proof. By Lemma 9.9, there exist disjoint open sets $V_i \subset M$ such that $N_i \subset V_i$, since $\mathcal{Z}$ is a proper sub-manifold of M by Lemma 9.1 (directly adapted to the manifold case). Hence, we can assume that the U_i are disjoint (consider instead $U_i \cap V_i$, the property still holds). Define $U_\star = \{f > 0\}$. Note that since the U_i's cover $\mathcal{Z}$, $U_\star \cup \bigcup_{i\in I} U_i$ covers M since f is non-negative: take $\chi_\star, (\chi_i)_{i\in I}$ to be a gluing family adapted to that covering given by Lemma 9.4. Note that for any $i, i' \in I$, we have $\chi_i \chi_{i'} = 0$ since χ_i is supported on U_i and the U_i are disjoint. Consider the function $g_j = \sum_{i\in I} \chi_i f_{i,j}$, which is well defined on M and C^{p-2} by Lemma 9.3. We have $g_j^2 = \sum_{i\in I} \chi_i^2 f_{i,j}^2$ since $\chi_i \chi_{i'} = 0$ when $i \neq i'$.

Assertion : the family $(g_j)_{j\in J}$ has locally finite support. Let $x \in \mathbb{R}^d$ and assume $g_j(x) \neq 0$. Then there exists $i \in I$ such that $\chi_i(x) > 0$, and hence there exists an open set U_x around x such that $U_x \subset U_i$. But in that case, $\chi_{i'}(x') = 0$ for all other i' and for all $x' \in U_x$ since the U_i are disjoint and $\chi_{i'}$ is supported on $U_{i'}$. Moreover, since $(f_{i,j})_{j\in J}$ is locally finite, there exists an open set V_x around x as well as a finite $J_0 \subset J$ such that $f_{i,j} = 0$ on V_x for all $j \in J \setminus J_0$. Hence, for any $j \in J \setminus J_0$ and any $x' \in U_x \cap V_x$, we have $f_{i,j}(x') = 0$ and $\chi_{i'}(x') = 0$ thus $g_j(x') = 0$. Thus, $U_x \cap V_x \subset M \setminus \text{supp}(g_j)$ for all $j \notin J_0$: the family g_j is locally finite.

Conclusion. Define $g_\star = \chi_\star \sqrt{f}$, which is of class C^p since $\chi_\star$ is supported on $\{f > 0\}$. Since the addition of one function changes nothing to the locally finite property of a family of functions, the family $(g_j)_{j\in J} \cup g_\star$ is still locally finite. Using the fact that $g_j^2 = \sum_{i\in I} \chi_i^2 f_{i,j}^2$, that $\sum_{j\in I\cup\{\star\}} \chi_i^2 = 1$ and Eq. (9.10), it holds

$$\begin{aligned}\sum_{j\in J\cup\{\star\}} g_j^2 &= \chi_\star^2 f + \sum_{j\in J} g_j^2 = \chi_\star^2 f + \sum_{j\in J}\sum_{i\in I} \chi_i^2 f_{i,j}^2 \\ &= \chi_\star^2 f + \sum_{i\in I}\sum_{j\in J} \chi_i^2 f_{i,j}^2 = \chi_\star^2 f + \sum_{i\in I} \chi_i^2 f = f.\end{aligned}$$

□

9.3 .3 Main results

We are now ready to state our main result on manifolds. On the right hand side of Sec. 9.3 , we represent a case where this theorem applies for a non-negative function defined on S^2.

Theorem 9.8. *Let M be a manifold and $f : M \to \mathbb{R}$ be a non-negative map of class C^p. Assume f satisfies the global normal Hessian condition. Then there exists I which is at most countable and functions $f_i \in C^{p-2}(M)$ for $i \in I$ such that the family (f_i) has locally finite support and*

$$\forall x \in M,\ f(x) = \sum_{i \in I} f_i(x)^2. \tag{9.12}$$

Moreover:

- *if f satisfies the strict Hessian condition, $\mathcal{Z}$ is discrete and we can find such a decomposition such that $|I| \leq d + 1$.*
- *if $\mathcal{Z}$ is compact, then $|I|$ can be taken to be finite.*

Proof. The proof of this theorem is a simple consequence of Lemma 9.5 and Lemma 9.6. Note that the global NHC Lemma 9.1 shows that $\mathcal{Z}$ is a sub-manifold of M. Let N_i denote the connected components of $\mathcal{Z}$. By Lemma 9.9, we can find disjoints open sets U_i such that $N_i \subset U_i$.

General case. Without any more assumptions, we know from Lemma 9.5 that on any connected component N_i, we can have a decomposition of the form $f = \sum_{j \in J} f_{i,j}^2$ with $f_{i,j} \in C^{p-2}$ a family with locally finite support on an open neighborhood V_i of N_i. Moreover, we know that J is at most countable. Adding zeros when necessary, and reindexing, we can assume that $J = \mathbb{N}$. Now applying Lemma 9.6, we prove the general case.

Compact case. If we assume that N is compact, since the U_i cover N, necessarily the number of connected components is finite (just extract a finite covering of N from the U_i). We know from Lemma 9.5 that on any connected component N_i, we can have a decomposition of the form $f = \sum_{j=1}^{n_i} f_{i,j}^2$ with $f_{i,j} \in C^{p-2}$ and $n_i \in \mathbb{N}$ on an open neighborhood V_i of N_i, since N_i is compact. Hence, up to adding $f_{i,j} = 0$, we can assume that $n_i = n = \max_i(n_i)$ since there are a finite number of connected components. Now applying Lemma 9.6 with $J = \{1, ..., n\}$, the result is proven in the compact case with $n + 1$ functions.

SHC case. If we assume that the SHC holds, every connected component N_i is a singleton $\{x_i\}$: we know from Lemma 9.5 we can have a decomposition of the form $f = \sum_{j=1}^{d} f_{i,j}^2$ with $f_{i,j} \in C^{p-2}$ on an open neighborhood V_i of N_i, since N_i is compact. Now applying Lemma 9.6 with $J = \{1, ..., d\}$, the result is proven with $d + 1$ functions.

□

Remark 31. *Note that the difference between the number of functions in the SHC case is better than the one obtained by Rudi, Marteau-Ferey, and Bach (2020). This is because of the two step procedure in the gluing: first in a connected component, and then between connected components. The long term goal is to be able to prove that we need only a finite number $N(d)$ of functions per connected component (in the compact case), and hence to have an explicit bound after gluing the connected components together, rather than just relying on a compact extraction argument, which is not as precise.*

9.4 Proof of the local decomposition as a sum of squares

In this section, we formally prove the key result of the paper, Theorem 9.5.

Proof. Note that the existence of the sub-manifold N of dimension d_0 around x_0 implies the existence of a local parametrization around x_0 (see Lafontaine, 2015, Theorem 1.21): there exists an open neighborhood $\widetilde{W}_0$ of 0 in $\mathbb{R}^{d_0}$, an open neighborhood U_{x_0} of x_0 in $\mathbb{R}^d$ and a C^k immersion $\phi : \widetilde{W}_0 \to U_{x_0}$ of class C^k such that ϕ is a homeomorphism from $\widetilde{W}_0$ onto $U_{x_0} \cap N$. Since restricting N to $N \cap U_{x_0}$ does not change the assumptions of the theorem, we will will assume that $N = \mathrm{Im}(\phi)$ for a C^k immersion $\phi : (0, \mathbb{R}^{d_0}) \to (x_0, \mathbb{R}^d)$. We will denote with $T_{x_0} := d\phi_0(\mathbb{R}^{d_0}) = T_{x_0}N$ the tangent space to N at x_0.

Before starting the proof, recall that for any $x \in \mathcal{Z}$, it holds $df(x) = 0$ and $d^2f(x) \succeq 0$ (or equivalently $\nabla^2 f(x) \succeq 0$). Moreover, note that if $A \in \mathbb{S}_+(\mathbb{R}^d)$ is a symmetric positive semi-definite matrix, if a vector $k \in \mathbb{R}^d$ satisfies $k^\top Ak = 0$, then $Ak = 0$ (this is a trivial consequence of the spectral theorem by decomposing k along an orthonormal basis of eigenvectors).

Step 1: characterizing the null-space of the Hessian. We will prove that under the assumptions of the theorem, *a)* T_{x_0} is equal to the null-space $\ker(\nabla^2 f(x_0))$ of the Hessian of f at x_0 and *b)* that for any supplementary S to T_{x_0}, the restricted Hessian $\nabla^2 f(x_0)|_S$ is positive definite.

To prove *a)*, assume that there exists an element in $k \in T_{x_0}$ such that $\nabla^2 f(x_0)k \neq 0$. Since $\nabla^2 f(x_0)$ is positive semi-definite, this implies that $k^\top \nabla^2 f(x_0)k > 0$. Let $h \in \mathbb{R}^{d_0}$ such that $d\phi_0 h = k$, and let $x_t = \phi(th)$ which is defined for t in an open neighborhood of 0. Using the Taylor expansion of f around x_0:

$$f(x) - f(x_0) - df(x_0)[x - x_0] = \tfrac{1}{2}(x - x_0)^\top \nabla^2 f(x_0)(x - x_0) + \epsilon(x - x_0)\|x - x_0\|^2,$$

where $\epsilon(x) \underset{\|x\| \to 0}{\to} 0$. Now applying this for x_t, since $f(x_t) = f(x_0) = 0$ and $df(x_0) = 0$, it holds:

$$0 = \tfrac{1}{2}(x_t - x_0)^\top \nabla^2 f(x_0)(x_t - x_0) + \epsilon(x_t - x_0)\|x_t - x_0\|^2.$$

Using the fact that ϕ is differentiable at 0 yields $x_t - x_0 = td\phi(0)[h] + o_{t\to 0}(t) = tk + o_{t\to 0}(t)$. Injecting this in the equation above yields

$$0 = t^2 \ \tfrac{1}{2}k^\top \nabla^2 f(x_0)k + o_{t\to 0}(t^2).$$

Hence, necessarily, $k^\top \nabla^2 f(x_0)k = 0$, which is a contradiction. This proves that $T_{x_0} \subset \ker(\nabla^2 f(x_0))$, and in particular, $d_0 \leq \dim(\ker(\nabla^2 f(x_0)))$. Since the rank of $\nabla^2 f(x_0)$ is actually $d - d_0$, the rank theorem shows that $\dim(\ker(\nabla^2 f(x_0))) = d_0$ and hence $T_{x_0} = \ker(\nabla^2 f(x_0))$.

To prove *b)*, we just need to prove that the restriction to any supplementary to the null-space of $\nabla^2 f(x_0)$ is positive definite. Using the small result at the beginning of the proof, any vector $k \in \mathbb{R}^d \setminus T_{x_0}$ satisfies $k^\top \nabla^2 f(x_0)k > 0$. In particular, this means that the restriction of $\nabla^2 f(x_0)$ to any supplementary subspace S of T_{x_0} is positive definite.

Step 2: applying the Morse lemma. Let $P = (P_1, P_2) \in O_d(\mathbb{R})$ be the matrix of an orthonormal basis adapted to the decomposition $\mathbb{R}^d = T_{x_0}^\perp \oplus T_{x_0}$. Note that $P_1 \in \mathbb{R}^{d\times(d-d_0)}$ and $P_2 \in \mathbb{R}^{d\times d_0}$ are also orthonormal matrices, and that since P_1 spans $T_{x_0}^\perp$, in particular $P_1^\top \nabla^2 f(x_0)P_1 \succ 0$.

Define $g : (x', y') \in \mathbb{R}^{d-d_0} \times \mathbb{R}^{d_0} \mapsto f(P_1x' + P_2y' + x_0) = f(\mathcal{A}(x', y'))$, where $\mathcal{A}(x', y') = P(x', y') + x_0$ is an isometry[3] ($\mathcal{A}^{-1}x = P^\top(x - x_0)$). We have $\nabla_{x'} g(0, 0) = P_1^\top \nabla f(x_0) = 0$ and

[3] An isometry is simply a map which preserves distances, and can be defined as an orthogonal transformation plus an affine shift.

$\nabla^2_{x'x'} g(0,0) = P_1^\top \nabla^2 f(x_0) P_1 \succ 0$. We can therefore apply the Morse lemma Lemma 9.10 to g: there exists two open neighborhoods of zero $V \subset \mathbb{R}^{d-d_0}, W \subset \mathbb{R}^{d_0}$ as well as $\varphi : W \to V$ of class C^{p-1} such that $\{(x', y') \in V \times W \ : \ \nabla_{x'} g(x', y') = 0\} = \{(x', y') \in V \times W \ : \ x' = \varphi(y')\}$ and $z : V \times W \to \mathbb{R}^{d-d_0}$ of class C^{p-2} such that

$$\forall (x', y') \in V \times W, \ g(x', y') = g(\varphi(y'), y') + \tfrac{1}{2} z(x', y')^\top H' z(x', y'), \tag{9.13}$$

where H' is the positive definite matrix $P_1^\top \nabla^2 f(x_0) P_1$.

Step 3: making the sum of squares appear. Since $H' \in \mathbb{S}_+(\mathbb{R}^{d-d_0})$ and $H' \succ 0$, we can decompose it using the spectral theorem: $H' = \sum_{i=1}^{d-d_0} \lambda_i u_i u_i^\top$ where the $\lambda_i > 0$. Defining $g_i(x', y') = \sqrt{\lambda_i / 2} u_i^\top z(x', y')$, Eq. (9.13) can be rewritten as

$$\forall (x', y') \in V \times W, \ g(x', y') = g(\varphi(y'), y') + \sum_{i=1}^{d-d_0} g_i^2(x', y'). \tag{9.14}$$

Note that the g_i are of class C^{p-2} since z is of class C^{p-2}. We see that if we can show that $g(\varphi(y'), y') = 0$ in a neighborhood of $(0, 0)$, since we can go back to the original coordinate system through $\mathcal{A}^{-1}$, we will have shown the theorem.

Step 4: characterizing $\mathcal{Z}$ in a neighborhood of x_0. Denote with $G_\varphi = \{(\varphi(y), y) \ : \ y \in W\}$ the graph of φ, and which is a sub-manifold of class C^{p-1} of $\mathbb{R}^{d-d_0} \times \mathbb{R}^d$ (see theorem 1.21, point (iv) by Lafontaine (2015)). Since $\mathcal{A}$ is an isometry, the set $\mathcal{A}(G_\varphi)$ is also a sub-manifold of class C^{p-1} of $\mathbb{R}^d$.

Let $\widetilde{W} = \phi^{-1}(\mathcal{A}(V \times W))$: it is an open neighborhood of 0. Note that $\phi(\widetilde{W}) \subset \mathcal{Z} \cap \mathcal{A}(V \times W)$ by assumption, and since for any $x \in \mathcal{Z}$, we have $\nabla f(x) = 0$, it holds in particular that for any $x \in \mathcal{Z} \cap \mathcal{A}(V \times W)$, we have $\nabla_{x'} g(\mathcal{A}^{-1}(x)) = P_1^\top \nabla f(x) = 0$. Hence, by the result of the Morse lemma, it holds $\mathcal{A}^{-1}(\phi(\widetilde{W})) \subset \mathcal{A}^{-1}(\mathcal{Z}) \cap (V \times W) \subset G_\varphi$.

Define $\psi : (x', y') \in V \times W \mapsto (x' - \varphi(y'), y')$ which is a C^{p-1} diffeomorphism onto its image with inverse $(t, u) \mapsto (t + \varphi(u), u)$. Note that ψ maps G_φ onto $\{0_{\mathbb{R}^{d-d_0}}\} \times W$. If π_2 denotes the canonical projection $\pi_2 : \mathbb{R}^{d-d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}^{d_0}$, we see that $\pi_2 \circ \psi$ maps G_φ injectively onto $W \subset \mathbb{R}^{d_0}$.

Take $\Phi = \pi_2 \circ \psi \circ \mathcal{A}^{-1} \circ \phi : \widetilde{W} \to \mathbb{R}^{d_0}$, which is well defined by definition of $\widetilde{W}$, and C^1 by composition. Note that it is an immersion at 0. Indeed *i)* ϕ maps 0 onto x_0 and is an immersion at 0 by assumption, hence $d\phi_0$ is injective; *ii)* $\psi \circ \mathcal{A}^{-1}$ is a C^{p-1} diffeomorphism from $\mathcal{A}(V \times W)$ (containing x_0) to its image, and hence its differential is invertible at x_0, and thus by composition, the differential $d(\psi \circ \mathcal{A}^{-1} \circ \phi)(0)$ is injective; *iii)* since $\mathcal{A}^{-1}(\phi(\widetilde{W})) \subset G_\varphi$ by a previous statement, and since $\psi(G_\varphi) \subset \{0\} \times W$ also by a previous statement, it, it holds that the differential $d(\psi \circ \mathcal{A}^{-1} \circ \phi)(0)\mathbb{R}^{d_0} \subset \{0\} \times \mathbb{R}^{d_0}$ and hence applying π_2 does not change the injectivity of the differential; hence Φ is an immersion at 0. But since $d\Phi_0$ is a linear map from $\mathbb{R}^{d_0}$ to $\mathbb{R}^{d_0}$, $d\Phi_0$ being injective is equivalent to $d\Phi_0$ being invertible. Hence, by the local inversion theorem Theorem 9.9, there exists an open neighborhood of 0 $\widetilde{W}' \subset \widetilde{W}$ and an open neighborhood of 0 $W' \subset W$ such that Φ is a C^1 diffeomorphism from $\widetilde{W}'$ to W'.

Define $U = (\pi_2 \circ \psi \circ \mathcal{A}^{-1})^{-1}(W') = \mathcal{A}(\psi^{-1}(\mathbb{R}^{d-d_0} \times W'))$,which is an open neighborhood of x_0. Note that since Φ is a diffeomorphism from $\widetilde{W}'$ to W', we have $\phi(\widetilde{W}') \subset U$. Moreover, since ψ is defined on $V \times W$, we have $U \subset \mathcal{A}(V \times W)$. Finally, let $u \in U \cap \mathcal{A}(G_\varphi)$. Since $u \in U$, there exists $\tilde{w}' \in \widetilde{W}'$ such that $\pi_2 \circ \psi \circ \mathcal{A}^{-1}(\phi(\tilde{w}')) = \pi_2 \circ \psi \circ \mathcal{A}^{-1}(u)$. Moreover, since $\pi_2 \circ \psi$ is injective on G_φ, and since both $\mathcal{A}^{-1}(\phi(\tilde{w}'))$ and $\mathcal{A}^{-1}(u)$ belong to G_φ (the first using the

previous point since $\widetilde{W'} \subset \widetilde{W}$ and the second by assumption), we have $\mathcal{A}^{-1}(\phi(\tilde{w}')) = \mathcal{A}^{-1}(u)$ and hence $u = \phi(\tilde{w}')$ since $\mathcal{A}$ is one to one. This shows that $U \cap \mathcal{A}(G_\varphi) \subset \phi(\widetilde{W'})$.

Moreover, a previous point shows that $\mathcal{A}^{-1}(\phi(\widetilde{W})) \subset \mathcal{A}^{-1}(\mathcal{Z}) \cap (V \times W) \subset G_\varphi$. Now since $\mathcal{A}$ is one to one and since $\widetilde{W'} \subset \widetilde{W}$ we have $\phi(\widetilde{W}) \subset \mathcal{Z} \cap (\mathcal{A}(V \times W)) \subset \mathcal{A}(G_\varphi)$. Since $\phi(\widetilde{W'}) \subset U$, we therefore have $\phi(\widetilde{W'}) \subset \mathcal{Z} \cap U \subset \mathcal{A}(G_\varphi) \cap U$. Combining this with the previous result, we finally have

$$\phi(\widetilde{W'}) \subset \mathcal{Z} \cap U \subset \mathcal{A}(G_\varphi) \cap U \subset \phi(\widetilde{W'}) \implies \phi(\widetilde{W'}) = \mathcal{Z} \cap U = \mathcal{A}(G_\varphi) \cap U. \tag{9.15}$$

Step 5: conclusion. Eq. (9.15) shows that $\phi(\widetilde{W'}) = \mathcal{Z} \cap U = \mathcal{A}(G_\varphi) \cap U$.

One the one hand, this shows that $U \cap \mathcal{Z}$ is the intersection between an open set U and a sub-manifold $\mathcal{A}(G_\varphi)$ of $\mathbb{R}^d$ of class C^{p-1} (since it is the composition of the graph of φ which is C^{p-1}, which is a C^{p-1} manifold by Lafontaine (2015), by an isometry which is in particular a diffeomorphism). Moreover, since ϕ is a C^k immersion which is a homeomorphism on its image, $\phi(\widetilde{W'})$ is a sub-manifold of class C^k. Thus, $U \cap \mathcal{Z}$ is a sub-manifold of $\mathbb{R}^d$ of class $C^{\max(k,p-1)}$.

On the other, since $\mathcal{A}^{-1}(U) \subset V \times W$, Eq. (9.14) becomes

$$\forall u \in U,\ g(\mathcal{A}^{-1}(u)) = g(\varphi(y'), y') + \sum_{i=1}^{d-d_0} g_i^2(\mathcal{A}^{-1}(u)),\ (x', y') = \mathcal{A}^{-1}(u). \tag{9.16}$$

Let $u \in U$ and write $(x', y') = \mathcal{A}^{-1}(u)$. First, note that $\mathcal{A}(\varphi(y'), y') \in \mathcal{A}(G_\varphi)$. Moreover, since $\mathcal{A}^{-1}u \in \psi^{-1}(\mathbb{R}^{d-d_0} \times W')$ by definition of U, this shows that $y' \in W'$ and hence $(\varphi(y'), y') = \psi^{-1}(0, y') \in \psi^{-1}(\mathbb{R}^{d-d_0} \times W')$. This in turn shows that $\mathcal{A}(\varphi(y'), y') \in U$. Hence, $\mathcal{A}(\varphi(y'), y') \in \mathcal{A}(G_\varphi) \cap U = \mathcal{Z} \cap U$ and thus $g((\varphi(y'), y')) = f(\mathcal{A}(\varphi(y'), y')) = 0$. Finally, using this in Eq. (9.16), recalling that $g = f \circ \mathcal{A}$, and defining $f_i : u \in U \mapsto g_i(\mathcal{A}^{-1}u)$, we have

$$\forall u \in U,\ f(u) = \sum_{i=1}^{d-d_0} f_i^2(u). \tag{9.17}$$

We see that f_i is of class C^{p-2} since g_i was of class C^{p-2} and $\mathcal{A}^{-1}$ is an isometry; this concludes the proof of the theorem. □

9.5 Discussion and possible extensions

In this work, we have provided second order sufficient conditions in order for a non-negative C^p function to be written as a sum of squares of C^{p-2} functions. We hope this will help provide a theoretical basis to algorithms which use functional sum of squares methods (Rudi, Marteau-Ferey, and Bach, 2020; Vacher, Muzellec, Rudi, Bach, and Vialard, 2021; Rudi and Ciliberto, 2021), which rely on the smoothness of such decompositions. As these conditions are sufficient and not necessary, one main problems is understanding this gap. This seems a highly difficult, and while very interesting, we present three other more reachable subjects for future work.

The first is to have an explicit bound for the number of squares needed in the sum of squares decomposition in the compact case. We believe that using finer tools from differentiable geometry, we should be able to obtain a bound depending on meaningful quantities, and upper bounded by a constant n_d depending only on the dimension d of the manifold on which the function is defined.

The second is, as in the polynomial case, to handle functions f which are non-negative on a constrained set defined by inequalities $f_i \geq 0$. More precisely, we would like to show second order sufficient conditions to write $f = g + \sum_i g_i f_i$ where the g, g_i are sum of squares of regular functions when f and the f_i are regular. This would open up the field of constrained optimization for methods developed for functions, such as kernel sum of squares (Rudi, Marteau-Ferey, and Bach, 2020). In the polynomial case, such conditions are given by so-called *Positivstellensätzen* (Putinar, 1993; Stengle, 1974; Schmügden, 1991), but usually assume the polynomial is positive. Second order conditions have been developed more specifically by Marshall (2006) to deal with non-negative polynomials with zeros.

The third is to handle functions with conic outputs which are more general than the non-negativity one. For example, in order to represent functions which have values in a cone defined by linear inequalities (Marteau-Ferey, Bach, and Rudi, 2020) or in the PSD cone (Muzellec, Bach, and Rudi, 2022).

9.A Around partitions of unity and gluing functions

In this section, we detail a few topological properties of manifolds, in order to *a)* decompose a manifold or a sub-manifold in connected components and *b)* use partitions of unity as a tool to glue functions together. These specific properties are needed for Sec. 9.3 .2. For basics on topological spaces (what is a topology, the notion of continuity, of homeomorphism), we refer to Chapter 1 by Jänich (1980). Main references for manifold are the works by Lafontaine (2015); Spivak (1999); Paulin (2006). Recall from Sec. 9.3 .3 the definition of a manifold M equipped with its atlas $\mathcal{A}$ of class C^k, and of a chart on M. A chart ϕ is said to be of class $C^{k'}$ for $k' \leq k$ if it compatible with the atlas up to k' smoothness, i.e. if the transitions maps $\phi \circ \phi_i^{-1}$ and $\phi_i \circ \phi^{-1}$ are all $C^{k'}$. A priori, the atlas of a manifold of class C^k is not unique in the sense that more than one atlas generate the same structure. To make it so, and to be able to say **the atlas** of M of class C^k, we consider the **maximal atlas** on M, i.e. the collection of all charts of class C^k on M.

9.A .1 Paracompactness and partitions of unity

The main point of asking a (differential) manifold to be second countable and Hausdorff, (and not just to be locally homeomorphic to $\mathbb{R}^d$), is for the manifold to be paracompact, and and hence to be equipped with partitions of unity. In this section, we introduce the main definitions and results on this topic.

Recall that a family of subsets (U_α) of a space X is said to be a covering of X if $\bigcup_\alpha U_\alpha = X$. It is said to be locally finite if for any $x \in X$, there exists an open neighborhood U of x which intersects only a finite number of the U_α. A family (V_β) is said to be a refinement of (U_α) if for all β, there exists an α such that $V_\beta \subset U_\alpha$.

A topological space X is said to be **paracompact** if for any open covering (U_α) of X, there exists an open refinement (V_β) of (U_α) such that (V_β) is locally finite, and is an open covering of X. The following lemma is proved in the first part of proposition 2.3 by Paulin (2006) or can be found in Theorem 2.13 by Spivak (1999).

Lemma 9.7. *A manifold is paracompact.*

Note that for Spivak (1999), a manifold is defined to be a metric space locally like $\mathbb{R}^d$. In proposition 2.2 by Paulin (2006), it is shown that being metric and second countable is equivalent to the countable Hausdorff condition (under the condition of being locally homeomorphic to $\mathbb{R}^d$). The condition by Spivak (1999) is a bit more general; in fact, it allows a manifold M to be a union of a possible non-countable connected component (as theorem 2 by Spivak (1999) shows that any connected component of a metric space locally homeomorphic to $\mathbb{R}^d$ is actually second countable).

Paracompactness is an important property as it yields the existence of partitions of unity. The following lemma is standard (a proof can be found by Paulin (2006), proposition 2.3). The result is of course also true for $k = 0$, but is more technical to prove.

Lemma 9.8 (Paulin (2006, Standard gluing lemma))**.** *Let $(U_i)_{i\in I}$ be an open covering of a manifold M of class C^k (i.e. $\bigcup_{i\in I} U_i = M$). There exists a family of functions $\chi_i : M \to [0, 1]$ of class C^k such that $\mathrm{supp}(\chi_i) \subset U_i$ for all $i \in I$ and with locally finite support satisfying:*

$$\sum_{i\in I} \chi_i = 1.$$

We now prove the two technical results need in Sec. 9.3 .2.

Proof of Lemma 9.3. The proof of this lemma is immediate. Indeed, by multiplication, we already know that χg is well defined and C^q on U. Moreover, for any point x in $V = M \setminus \text{supp}(\chi)$, which is an open set, $(\chi g)(x) = 0$ (by definition if $x \in M \setminus U$ and since $\chi(x) = 0$ if $x \in U$) and hence is C^q on V. Since $V \cup U = M$ as $\text{supp}(\chi) \subset U$, the property holds. Moreover, since $\chi g = 0$ on V, $\text{supp}(\chi g) \subset \text{supp}(\chi) \subset U$. □

Proof of Lemma 9.4. The proof of this result is a consequence of Lemma 9.8. Indeed, this result shows that there exists a family of function $\widetilde{\chi}_i : M \to [0, 1]$ of class C^k such that *i)* for all $i \in I$, $\text{supp}(\widetilde{\chi}_i) \subset U_i$, *ii)* the support of $(\widetilde{\chi}_i)$ is locally finite and *iii)* $\sum_i \widetilde{\chi}_i = 1$.

Define $\phi = \sum_i \widetilde{\chi}_i^2$. Since $\sum_i \widetilde{\chi}_i = 1$, and $\widetilde{\chi} \geq 0$, necessarily, $\phi > 0$. Hence $\sqrt{\phi}$ is of class C^k, and hence $\chi_i := \widetilde{\chi}_i / \sqrt{\phi}$ is of class C^k, and satisfies all the desired properties. □

9.A .2 Connected components

Connectedness is a key topological notion for manifolds, and allows to decompose manifold into separate blocks. Recall that two points x, x' of a topological set X are connected if there exists no two open sets U, V such that $X = U \cup V$, $x \in U$ and $x' \in V$. Since being connected is an equivalence relation, we can partition X in classes with respect to that relation, which are called "connected components". Connected components are both open and closed[4]. On a connected component of a manifold, the dimension d of the charts $\phi : U \to \mathbb{R}^d$ is the same, and is called the dimension of that connected component (for more details, see any of the references on manifolds). Note that as a manifold M is assumed to be second-countable, it has at most a countable number of connected components. Recall that a sub-manifold is defined in the main text as follows (such a definition can be found in section 2.4.2 by Paulin (2006)).

Definition 9.4. *Let M be a manifold of class $C^{k'}$, $k \leq k'$. N is a sub-manifold of M of class C^k if for any $x \in N$, and any chart $\phi : U \to \mathbb{R}^d$ defined around x and of class C^k, $\phi(U \cap N)$ is a sub-manifold (in the sense of $\mathbb{R}^d$, see Sec. 9.2 .1) of $\mathbb{R}^d$ around $\phi(x)$. It is equivalent to ask the existence of one such chart per point x.*

Let N be a sub-manifold of class C^k of a manifold M of class $C^{k'}$. Then it is naturally a manifold of class C^k in its own right. Indeed, consider that *i)* N is equipped with the topology of M, i.e. V is open in N iif $V = U \cap N$ for some open set of M, and *ii)* the atlas of N is (the completion of) the set of restrictions of charts $\phi|_{U \cap N}$ where $\phi : U \to \mathbb{R}^d$ is a C^k chart on M such that $\phi(U \cap N) \subset \mathbb{R}^{d'} \times \{0\}$, where d' is the dimension of N at $x \in U$ (we identify $\mathbb{R}^{d'} \times \{0\} \approx \mathbb{R}^{d'}$). Note that the second-countable Hausdorff condition directly follows from that of M. Moreover, the C^k compatibility of the charts is evident. From now on, when considering a sub-manifold $N \subset M$ as a manifold, it will be with this structure. The reason for the introduction of all these concepts is to obtain the following lemma, which while it seems natural, we have not found as such in the literature.

Lemma 9.9. *Let N be a sub-manifold of a manifold M. Let $(N_i)_{i \in I}$ be the connected components of N. There exists a collection of disjoint open sets $(U_i)_{i \in I}$ of M such that each $N_i \subset U_i$.*

Proof. This proof relies mainly on paracompactness.

Step 1. For all $x \in N$, there exists U_x an open set in M such that $\overline{U_x} \cap N$ is included in the unique connected component of x in N. Indeed, by definition 9.4, there exists a chart $\phi : U \to N$ where U is an open neighborhood of x. But since $\phi(U \cap N)$ is a sub-manifold of $\mathbb{R}^d$ of class C^k around $\phi(x)$, by Theorem 2.5 by Paulin (2006), there exists a C^k diffeomorphism

[4] For more details on connected components, see the work by Jänich (1980)

$\psi : (\phi(x), V) \to (0, W)$ where V such that $\psi(\phi(U \cap N) \cap V) = W \cap (\mathbb{R}^{d'} \times \{0\})$ for some d'. Taking $\widetilde{\phi} = \psi \circ \phi$ on $\widetilde{U} = \phi^{-1}(V) \cap U$, we have a chart of class C^k around x such that $\widetilde{\phi} : \widetilde{U} \to W \subset \mathbb{R}^d$ such that $\widetilde{\phi}(\widetilde{U} \cap N) = W \cap (\mathbb{R}^{d'} \times \{0\})$. Now let r be a radius such that the closed ball $\overline{B}(0, r) \subset W$. Set $U_x = \widetilde{\phi}^{-1}(B(0, r))$, which is an open neighborhood of x included in $\widetilde{U}$. Note that $\overline{U_x} \subset \widetilde{\phi}^{-1}(\overline{B}(0, r)) \subset \widetilde{U}$ since $\widetilde{\phi}$ is continuous. Since $\overline{U}_x \cap N = \widetilde{\phi}^{-1}(\overline{B}(0, r) \cap (\mathbb{R}^{d'} \times \{0\}))$ which is connected, we have that $\overline{U}_x \cap N$ is connected and hence is included in the unique connected component of x.

Step 2. Consider the collection of open sets U_x. By paracompactness of $U := \bigcup_{x \in N} U_x$ (it is a manifold), we can find an open cover (U_α) of U which is locally finite, and which still satisfies the condition that for all α, $\overline{U}_\alpha \cap N$ is included in at most one connected component of N. Let $(N_i)_{i \in I}$ denote the connected components of N. For $i \in I$, let $(V_{i,\alpha})_{\alpha \in A_i}$ denote the collection of open sets U_α such that $\overline{U}_\alpha \cap N \subset N_i$ and $\overline{U}_\alpha \cap N \neq \emptyset$. These collections satisfy *a)* the $(V_{i,\alpha})_{\alpha \in A_i}$ cover N_i; *b)* the collection $(V_{i,\alpha})_{i \in I,\ \alpha \in A_i}$ has locally finite support; and *c)* $\overline{V_{i,\alpha}} \cap N_i \subset N_i$ for all $i \in I,\ \alpha \in A_i$.

Step 3. For all $i \in I$, define $F_i = \bigcup_{j \in I \setminus \{i\}, \beta \in A_j} \overline{V}_{j,\beta}$ and for all $\alpha \in A_i$, consider the set $W_{i,\alpha} = V_{i,\alpha} \setminus F_i$. $W_{i,\alpha}$ is open, and $W_{i,\alpha} \cap N = V_{i,\alpha} \cap N$. Indeed, let $x \in W_{i,\alpha}$. Since the $(V_{j,\beta})$ are locally finite, there exists $V_x \subset V_{i,\alpha}$ such that V_x intersects a finite number of the $V_{j,\beta}$ and hence of the $\overline{V}_{j,\beta}$. Hence, $V_x \setminus F_i$ is still open. Hence $W_{i,\alpha}$ is open. The second condition comes from the fact that $\overline{V}_{i,\alpha} \cap N \subset N_i$, and that the connected components are disjoint Finally, taking $W_i = \bigcup_{\alpha \in A_i} W_{i,\alpha}$, the W_i satisfy all the desired properties (they are disjoint thanks to the previous point and cover N_i since the $V_{i,\alpha}$ covered N_i). □

9.B Morse lemma

In order for this article to be self contained, we restate the following classical lemmas from differential geometry and topology. Recall that a C^k-diffeomorphism is a map $\phi : U \subset \mathbb{R}^d \to V \subset \mathbb{R}^{d'}$ which is of class C^k and whose inverse is of class C^k (in that case, necessarily, $d = d'$). The following results are classical.

Theorem 9.9 (Theorem 1.13 by Lafontaine (2015))**.** *Let $f : (x_0, \mathbb{R}^d) \to \mathbb{R}^d$ be a function of class C^k ($k \geq 1$) defined around x_0 and such that $df(x_0)$ is invertible. Then there exists a neighborhood U of x_0 such that $f(U)$ is open and $f : U \to f(U)$ is a C^k diffeomorphism.*

Theorem 9.10 (Theorem 1.18 by Lafontaine (2015))**.** *Let $f : (x_0, \mathbb{R}^{d_1}) \to (y_0, \mathbb{R}^{d_2})$ be a function of class C^k ($k \geq 1$) defined around x_0 s.t. $df(x_0)$ is surjective and $f(x_0) = y_0$. Then there exists an open neighborhood U of x_0 in $\mathbb{R}^{d_1}$, V of y_0 in $\mathbb{R}^{d_2}$ as well as a function $g : V \to U$ of class C^k such that $g(y_0) = x_0$ and $f \circ g = Id_{\mathbb{R}^{d_2}}$*

We restate and reprove Lemma C.6.1 by Hörmander (2007), which is a generalization of the so-called Morse Lemma (see lemma 2.2 by Milnor (1963)), and which is the basis of Morse Theory. We will consider a function of two variables $f(x, y)$ defined on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. We will denote with $\nabla_x f(x, y)$ its gradient with respect to the first variable taken at point (x, y); it is an element of $\mathbb{R}^{d_1}$. Similarly, we will use the notation $\nabla^2_{xx} f(x, y) \in \mathbb{R}^{d_1 \times d_1}$ to denote the Hessian matrix taken with respect to the first coordinate at point (x, y). It is symmetric.

Lemma 9.10 (Hörmander (2007, Lemma C.6.1))**.** *Let $d_1, d_2 \in \mathbb{N}, p \in \mathbb{N}$ with $p \geq 2$. Let $f : (x, y) \in U_0 \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \mapsto f(x, y) \in \mathbb{R}$ be C^p function defined on a neighborhood U_0 of $(0, 0)$. Assume that $\nabla_x f(0, 0) = 0$ and that $H := \nabla^2_{xx} f(0, 0)$ is non-singular.*

There exists an open convex neighborhood V of 0 in $\mathbb{R}^{d_1}$ and an open convex neighborhood W of

0 in $\mathbb{R}^{d_2}$ such that $V \times W \subset U_0$, a map $\varphi \in C^{p-1}(W, V)$ and a map $z \in C^{p-2}(V \times W, \mathbb{R}^{d_1})$ such that for any $(x, y) \in V \times W$ $\nabla_x f(x, y) = 0$ if, and only if $x = \varphi(y)$, and

$$\forall (x, y) \in V \times W, \;\; f(x, y) = f(\varphi(y), y) + \frac{1}{2} z(x, y)^\top H z(x, y). \tag{9.18}$$

To simplify the proof, we first show an intermediate result which gives φ.

Lemma 9.11. *Under the assumptions of Lemma 9.10, there exists two open convex neighborhoods of zero $V_0 \subset \mathbb{R}^{d_1}$, $W_0 \subset \mathbb{R}^{d_2}$ and $\varphi : W_0 \to V_0$ of class C^{p-1} such that a) $V_0 \times W_0 \subset U_0$ and b) $\forall (x, y) \in V_0 \times W_0$, $\nabla_x f(x, y) = 0 \Leftrightarrow x = \varphi(y)$.*

Proof. Consider the map $\psi : (x, y) \in U_0 \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \mapsto (\nabla_x f(x, y), y)$. Its jacobian at $(0, 0)$ is of the form $\begin{pmatrix} H & \star \\ 0 & \mathbf{I}_{d_2} \end{pmatrix}$. Since H is non-singular, this matrix is non-singular. Applying the local inversion lemma Theorem 9.9, there exists an open neighborhood $U_1 \subset U_0$ such that ψ is a C^{p-1} diffeomorphism from U_1 to $\psi(U_1)$.

Let $\widetilde{V}_0 \subset \mathbb{R}^{d_1}$, $\widetilde{W}_0 \subset \mathbb{R}^{d_2}$ be open convex neighborhoods of 0 such that $\widetilde{V}_0 \times \widetilde{W}_0 \subset U_1 \cap \psi(U_1)$. Define $\varphi : w \in \widetilde{W}_0 \mapsto \pi_1(\psi^{-1}(0, w)) \in \mathbb{R}^{d_1}$. Defining $V_0 = \tilde{V}_0$ and $W_0 \subset \mathbb{R}^{d_2}$ to be an open convex neighborhood of 0 included in $\varphi^{-1}(V_0) \cap \widetilde{W}_0$, we have $\varphi(W_0) \subset V_0$ and $V_0 \times W_0 \subset U_1 \subset U_0$.

Moreover, for any $(x, y) \in V_0 \times W_0 \subset U_1 \cap \psi(U_1)$, $\nabla_1 f(x, y) = 0$ iif $\psi(x, y) = (0, y) \in \psi(U_1)$, iif $(x, y) = \psi^{-1}(0, y) = (\varphi(y), y)$, iif $x = \varphi(y)$. □

We can now prove our main result.

Proof of Lemma 9.10. Fix V_0, W_0 satisfying the properties of Lemma 9.11. Let $(x, y) \in V_0 \times W_0$. For $t \in [0, 1]$, define $x_t = \varphi(y) + t(x - \varphi(y))$. By convexity of V_0, $(x_t, y) \in V_0 \times W_0 \subset U_1 \subset U_0$ for all $t \in [0, 1]$. Thus, the map $g : t \in [0, 1] \mapsto f(x_t, y)$ is well defined, and we can apply the Taylor formula $g(1) = g(0) + g'(0) + \int_0^1 (1 - t) g''(t) dt$ and the fact that $g'(0) = \nabla_x f(\varphi(y), y) \cdot (x - \varphi(y)) = 0$ to obtain

$$f(x, y) = f(\varphi(y), y) + (x - \varphi(y))^\top \left(\int_0^1 (1 - t) \nabla^2_{xx} f(x_t, y) dt \right) (x - \varphi(y))$$

Defining $B : V_0 \times W_0 \to S(\mathbb{R}^{d_1})$, such that $B(x, y) := 2 \int_0^1 (1 - t) \nabla^2_{xx} f(x_t, y) dt$, the previous equation can simply be written $f(x, y) = f(\varphi(y), y) + \frac{1}{2}(x - \varphi(y))^\top B(x, y)(x - \varphi(y))$. Note that $B \in C^{p-2}(V_1 \times W_1, S(\mathbb{R}^{d_1}))$ and $B(0, 0) = H$. Now define $G : R \in \mathbb{R}^{d_1 \times d_1} \mapsto R^\top H R \in S(\mathbb{R}^{d_1})$ which is C^∞ and whose differential in $\mathbf{I}_{\mathbb{R}^{d_1}}$ is surjective (Hörmander, 2007). Theorem 9.10 shows there exists an neighborhood $\mathcal{O}$ of H in $S(\mathbb{R}^{d_1})$ and a C^∞ function $F : \mathcal{O} \to \mathbb{R}^{d_1 \times d_1}$ such that $(G \circ F)(B) = B$ for all $B \in \mathcal{O}$. Let $V \subset \mathbb{R}^{d_1}, \widetilde{W} \subset \mathbb{R}^{d_2}$ be two open convex neighborhoods of 0 such that $V \times \widetilde{W} \subset B^{-1}(\mathcal{O})$. Let W be an open convex neighborhood of 0 such that $W \subset \widetilde{W} \cap \varphi^{-1}(V)$ and define $z(x, y) = (F \circ B)(x, y)(x - \varphi(y))$. z satisfies Eq. (9.18). □

Chapter 10

Conclusion and perspectives

Contents

10.1 Summary of the thesis 445
10.2 Research perspectives 446

10.1 Summary of the thesis

In this thesis, we have explored three main topics, building on the great versatility of kernel methods. First, we focused on a statistical learning issue, extending fast rates and algorithms which existed in the least-square setting to other losses, and in particular to the logistic regression setting. Second, we developed a kernel based model to handle non-negativity constraints (PSD models), opening the way to the modelling of functions with constrained outputs, while keeping the good properties of kernel methods in terms of algorithms and approximation results. We then used it to model probability densities and to sample from the associated distribution. Third, using the analogy between PSD models and the sum of squares model for polynomials, we considered the problem of global optimization of regular functions using a so-called "kernel sum of squares" approach. We designed an algorithm which approximately minimizes a regular function using n evaluations points, and provided guarantees which shows that this algorithm adapts to the regularity of the function to be minimized, needing less evaluation points to reach a certain precision if the function is more regular.

In all these different settings, we have tried to put modelling at the center of the analysis. We have tried to show how the statistical and algorithmic properties can be characterized using properties of the models we consider, and argue that choosing the right model for the right task remains a (if not the) important choice of a practitionner, but also of the theoretician. While this is only one point of view, we believe it to be a nice way of looking at the different machine learning and applied mathematics problems from a central perspective.

We now summarize the contributions of this thesis in (slightly) greater detail.

In part I, we extended the classical "slow rates" known for logistic regression to match the "fast rates" known for least squares (Caponnetto and De Vito, 2007; Blanchard and Mücke, 2018). We introduced two key quantities : the bias and the effective dimension (see definition 2.1), which charaterize the complexity of the problem in this setting, and which generalize the corresponding quantities defined in the least-square case. These quantities as well as the derived trade-offs and

rates, helped us understand what makes a logistic regression problem either complicated or simple, and the interaction between the complexity of the problem and the necessary regularization. These quantities have allowed us to show that under certain assumptions, regularized empirical risk minimization can leverage the regularity of the solution and of the RKHS, and perform better than the slow rates $O(1/\sqrt{n})$, going as fast as $O(1/n)$ in the most favorable settings. In certain of these favorable settings, we have shown that the regularization needs to be chosen very small, *i.e.*, of order $O(1/n)$, in order to obtain optimal rates. From a more applied perspective, we have also derived results based on the bias-variance decomposition to design a meothod to solve the empirical risk minimization problem which is both efficient and statistically optimal. In particular, we have adapted the dimension reduction techniques of least-squares (Rudi, Camoriano, and Rosasco, 2015) to handle the fact that *a priori*, the empirical risk minimization problem is represented in a large dimensional space. We then proposed a novel second-order scheme, resembling an interior point method, to optimize the modified problem in a way which is less depednent on the conditioning of the problem than first order methods : this was crucial because as was shown in the statistical part and was verified empirically on real world data, the regularization can be very small. Finally, we derived optimal rates for the algorithm, as in the work by Rudi, Carratino, and Rosasco (2017) in the least squares setting, showing that we have an optimal algorithm with limited dependence to the condition number, and low complexity.

In part II, we introduced and analyzed a model for non-negative functions based on kernels (PSD models), which is PSD representable, *i.e.*, is parametrized by a (subspace of a) PSD matrix cone. We ansalysed the basic properties of this model in terms of approximation and optimization guarantees. We showed that it comes with the same advantages as (non-parametric) linear models, inheriting the nice properties of kernel methods to model real or vector valued functions. We have applied this model to sampling from an un-normalized density function, using it to model the density in a way where we can easily sample from it.

Finally in part III, we have used PSD models in a completely different context, to derive an algorithm for global (non-convex) optimization via function evaluations by modelling a non-negativity constrain using a PSD model. We derived guarantees for that algorithm when minimizing a function defined on a bounded domain of $\mathbb{R}^d$, assuming that it is regular, of class C^r. In particular, we showed that our algorithm almost matches the worst case lower bounds for global optimization of a C^r function based on function evaluations given by Novak (2006), showing that our global optimization algorithm reaches error ε using roughly $n = \varepsilon^{-d/r}$ evaluation points. The method to obtain these guarantees was crucially based on a decomposition of the non-negative function $f - \min(f)$ as a sum of squares of regular functions. In a second work in this part, we therefore explored second-order sufficient conditions in order to decompose C^r functions as sums of squares of C^{r-2} functions, in order to understand what can be achieved by our algorithm in the context of C^r regularity, and extend the cornerstone results to handle functions with continuous sets of minimizers and which are defined on a manifold, thus paving the way to extend the guarantees of the method to a wider variety of spaces when working with the C^r-type regularity.

10.2 Research perspectives

We conclude this manuscript with some perspectives and open questions related to the different areas and problems we have worked with. We will keep the same structure as that of the thesis, developing questions and perspectives raised by each part. Note that this is an informal presentation of these perspectives, and that the conjectures we will formulate are not made to be exact; we will omit certain elements for simplicity, keeping to the main ideas.

10.2 .1 Beyond least squares

In part I, we extended results from the least-squares setting to the class of generalized self-concordant functions, in terms of statistical rates for the regularized empirical risk minimizer (defining the bias, effective dimension, and showing a bias-variance decomposition), as well as algorithms.

The least-squares setting has served as a major model not only in itself, but also as a toy model that can be analyzed in order to understand wider phenomena, which are observed on other loss functions. Many results, statistical and algorithmic, exist in the least squares setting, and their proof critically relies on the simple form of the loss function, and the closed form solutions for the least-squares problem. Extending these results to other losses, starting with the widespread logistic loss for classification, is an obvious perspective and is directly in the line of what we have done in part I. As we will see, there is still a lot to be done, especially in the stochastic setting.

Related works

Before formulating open questions and resarch perspectives, we wish to come back on two works which are related to ours, and which already provide additional steps in generalizing results from least squares to the generalized self concordant setting.

Library including our method. Meanti, Carratino, Rosasco, and Rudi (2020) have developed a library, based on fast kernel computation routines (Charlier, Feydy, Glaunes, Collin, and Durif, 2021), and computations on GPUs, in order to make kernel methods scalable to billions of points. In their library, they include least square regression as well as our method for logistic regression.

Extending the statistical rates to very regular problems. Beugnot, Mairal, and Rudi (2021) have developed a statistical analysis of the performance of regularized empirical risk minimization for GSC functions in the case where the target function is very regular. Indeed, note that in main corollary 1, we have assumed that the source condition $f_* = \mathbf{H}^{r-1/2}h$ holds for some $h \in \mathcal{H}$ with $r \leq 1$.

In the least squares case, when $r \geq 1$, the Tikhonov regularization is not optimal anymore : it behaves as if the functions satisfies the source condition with $r = 1/2$ and does not leverage the additionnal regularity. This is referred to in the literature on least squares as the saturation effect of Tikhonov regularization (Gerfo, Rosasco, Odone, Vito, and Verri, 2008), and is linked to the fact that the bias term introduced in definition 2.1 cannot go faster than $b_\lambda \leq R^2\lambda^2$ due to the specific form of Tikhonov regularization. In the work by Blanchard and Mücke (2018), it is shown that the optimal rates when $r \geq 1$ match the expression in Eq. (2.12), that is there is no saturation effect in the optimal rates, and that these rates can be achieved usign a different spectral regularization than Tikhonov, satisfying a certain qualification property.

The same phenomenon happens in the context of GSC function : our bound on the bias term does not allow a faster convergence of the bias than the one obtained for $r = 1$. Beugnot, Mairal, and Rudi (2021) circumvent this problem by introducing an estimator based on iterating Tikhonov regularization steps. They define :

$$\widehat{f}^t_{n,\lambda} = \arg\min_{f\in\mathcal{H}} \widehat{R}_n(f) + \tfrac{\lambda}{2}\|f - \widehat{f}^{t-1}_{n,\lambda}\|^2_{\mathcal{H}}, \qquad \widehat{f}^0_{n,\lambda} = 0. \tag{10.1}$$

Analyzing this method as a form of proximal algorithm, they show that as soon as t is large enough, the qualification of the regularization is large enough to achieve optimal rates, without

saturation effect.

Perspectives from a statistics point of view

From a statistics point of view, we see two main perspectives. The first one seems relatively easily reachable to us, while the second is more vague as it goes further away from our setting.

Minimax rates. The first is to formally derive minimax upper and lower rates in the setting of GSC losses. For the upper rates, we believe that using Main theorem 1 as a starting point, and applying the same method as in the work by Blanchard and Mücke (2018), we would obtain minimax upper bounds (of course, in the $r \geq 1/2$ setting, we should not use Tikhonov regularization but the one proposed by Beugnot, Mairal, and Rudi (2021)).

Conjecture 10.1: Minimax upper bounds for GSC losses

Fix $\mathsf{b}_1, \mathsf{b}_2 > 0$, $r > 1/2$, $Q > 0$,$R > 0$, and $b \geq 1$ and let $\mathcal{M} = \mathcal{M}(R, r, \mathsf{Q}, b, \mathsf{b}_1, \mathsf{b}_2)$, as defined in main corollary 1. Let θ denote the variables $\theta = (\mathsf{Q}, R)$. Setting

$$\lambda_{n,\theta} = \min\left(\left(\frac{\mathsf{Q}^2}{R^2 n}\right)^{\frac{1}{2br+1}}, 1\right), a_{n,\theta} = R^2\left(\frac{\mathsf{Q}^2}{R^2 n}\right)^{\frac{2br}{2br+1}}, \tag{10.2}$$

for all $\psi(\cdot) = |\cdot|^p$ for $p > 0$, the following minimax rates of convergence hold with rate $a_{n,\theta}$:

$$\sup_{\theta\in\Theta} \limsup_{n\to+\infty} \sup_{\rho\in\mathcal{M}_\theta} \mathbb{E}_{\rho^{\otimes n}}\left[\psi\left(\frac{\mathcal{R}(\widehat{f}_{n,\lambda_{n,\theta}}) - \mathcal{R}(f_\rho)}{r_{n,\theta}}\right)\right] < \infty \tag{10.3}$$

We also believe that analoguous lower rates can be obtained, using the fact that GSC losses are essentially equal to a quadratic function in a small region around their minimum. However, one would have to be more careful than in the least squared case, as that quadratic depends on the distribution. The proof by Blanchard and Mücke (2018) using the scheme by Tsybakov (2008) will have to be studied in depth to see if it can be generalized as such, as we would have to find densities $\rho_1, ..., \rho_N$ which are close enough to be compared using a quadratic approximation (*i.e.*, they must induce functions f_{ρ_i} which are in the same Dikin ellipsoids), while far enough from each other to obtain the desired lower bound.

The misspecified setting. Our second perspective is to derive rates and provable algorithms in the misspecified setting, that is when we do not assume that f_ρ belongs to the space $\mathcal{H}$. In the least squares case, the degree of misspecification can be quantified by the kernel operator T_k which compares the space $L^2(\mathcal{X}, \rho_\mathcal{X})$ with the space $\mathcal{H}$, and rates can still be obtained for certain algorithms (although they are usually upper rates and not optimal rates). We would have to define similar objects for more general losses, where the natural metric is not the L^2 metric but (in the case of GSC losses), the metric induced by the Hessian at the optimum, which is a form of weighed L^2 metric.

Extending fast rates of stochastic gradient optimization

In the least squares setting, a vast literature exists on the problem of minimizing the square loss when having access only to a stochastic oracle of the gradient (Dieuleveut, Flammarion, and Bach, 2017). Recall that in the kernel least squares setting, the goal is to solve

$$\min_{f\in\mathcal{H}} \mathcal{R}(f) = \tfrac{1}{2}\mathbb{E}\left[\|f(X) - Y\|^2\right] = \tfrac{1}{2}\langle f, \Sigma f\rangle_\mathcal{H} - \mathbb{E}\left[\langle f, Y k_X\rangle_\mathcal{H}\right] + \text{cste}, \tag{10.4}$$

where Σ is the covariance operator of the kernel k with respect to the measure $\rho_{\mathcal{X}}$ of X. Consider the following regularized stochastic gradient descent scheme :

$$\forall n \geq 1, \ f_n = f_{n-1} - \gamma G_n(f_{n-1}) - \gamma\lambda(f_{n-1} - f_0), \tag{10.5}$$

where $G_n(f_{n-1})$ is an unbiased estimator of $\nabla\mathcal{R}(f_{n-1})$ (we assume $\mathbb{E}\left[G_n(f)|\mathcal{F}_{n-1}\right] = \nabla\mathcal{R}(f)$, where $\mathcal{F}_{n-1}$ is the σ-algebra generated by $f_0, ..., f_{n-1}$. Write $G_n(f) = \nabla\mathcal{R}(f) - \xi_n$, where ξ_i is the noise of the gradient. Under the *additive noise assumption*, that is $\xi_n = y_n k_{x_n} - \mathbb{E}\left[Yk_X\right]$, Bach and Moulines (2013) show the following bounds on the performance of $\overline{f}_n = \frac{1}{n}\sum_{i=1}^n f_i$:

$$\mathbb{E}\left[\mathcal{R}(\overline{f}_n)\right] - \mathcal{R}(f_\rho) \leq \left(\lambda + \frac{1}{\gamma n}\right)^2 \|\Sigma^{1/2}(\Sigma + \lambda\mathbf{I})^{-1}(f_0 - f_\rho)\|^2 + \frac{\tau^2 \operatorname{Tr}(\Sigma^2(\Sigma + \lambda\mathbf{I})^{-2})}{n} \tag{10.6}$$

This bound shows a real bias variance trade off.

- The first term is a bias term controlled by the regularity of $f_0 - f_\rho$.
- The second term is a variance term; in particular, it is controlled by the quantity $\operatorname{Tr}(\Sigma^2(\Sigma + \lambda\mathbf{I})^{-2}) \leq d_\lambda$ where d_λ is the effective dimension for least squares introduced in chapter 2.

Note that Dieuleveut, Flammarion, and Bach (2017) provide bounds for accelerated methods as well under the additive noise assumption. Moreover, they prove similar bounds in the case where there is multiplicative and additive noise (see Dieuleveut, Flammarion, and Bach (2017), Sec. 3. for details), although not for accelerated methods.

From our point of view, it would be very interesting to extend these rates to the logistic regression and GSC functions setting. Indeed, the fact that both a complex bias and variance term appear, which match those we encounter in the setting described in chapter 2, shows that there is hope to derive similar rates with meaningful quantities in the logistic case. However, the extension is not trivial, and exploring this area of research would help us understand if properties like GSC actually help us to analyze first order methods or not.

10.2 .2 A broader understanding of PSD models

In part II, we introduced PSD models, which are linear models of the form $g_A(x) = \langle\phi(x), A\phi(x)\rangle_H$ for a given feature map ϕ representing $\mathcal{X}$ in a Hilbert space H, and where A is constrained to belong to the PSD cone of compact positive semidefinite operators. Denote with SOS^H this set of functions. While we have presented some key properties of these models, and their applications to sampling, a lot of questions and directions are left to explore to better understand their advantages and limits.

We will divide this section into three parts. First, we will present our main lines of inquiry in order to better understand and use PSD models from a generic point of view (statistics, algorithms). Then, we will focus on sampling, and present research directions motivated by chapter 6. Finally, we will take a step back and raise some questions about modeling constraints in general.

Towards a better understanding of approximation and algorithmic properties of PSD models

Approximation properties. As is usually the case for non-parametric models such as RKHSs, it is not enough to know that our model for non-negative functions can approximate any non-negative functions pointwise or on a compact set to get precise rates non asymptotic rates of convergence.

Indeed, in order to get more precise (fast) rates, as in the least squares and GSC settings described in part I, we usually need the target function to lie in the space SOS^H, or in an interpolation space between SOS^H and another "natural" space for the problem at hand (such as $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ in the least squares setting). In any case, we need to better understand the conditions for a non negative function to be in SOS^H, that is for the model to be well-specified (see chapters 7 and 9 as well as Sec. 10.2 .3). Finding such conditions is therefore a natural perspective. Note that this has been treated in part for functions defined on $\mathbb{R}^d$ an on differentiable manifolds in the case where the kernel quantifies the C^r regularity of a function, such as Sobolev RKHSs (see chapters 8 and 9). Obtaining such results for different spaces $\mathcal{X}$ and kernels on $\mathcal{X}$ would be an interesting direction to investigate. Moreover, the results we have derived so far for representability as a sum of squares of functions show that A is finite rank. This leads to the following question : does allowing the number of squares to go to infinity help have better approximation properties ? Finally, a question can be raised in relation to the importance of the spectral norm chosen on SOS^H, as it characterizes the regularization. Given a situation or problem, can we have elements to characterize the ideal spectral norm to use (for now, we have heuristically chosen the trace norm as the optimal solutions are assumed to be finite rank) ?

Dimension reduction. A second key element which we must investigate further, and that has both statistical and algorithmic repercussions, is the dimension reduction aspect, for analoguous reasons to the ones described in chapter 2. Indeed, even though there is a reprenter theorem for PSD models Theorem 5.1, it is not always applicable (for example, when we must enforce an affine constraint such as $\int_{\mathcal{X}} g_A(x)dx = 1$ in the case of densities), and leads to a SDP of size n which can quickly become untractable. It is therefore crucial to reduce the dimension of the resulting SDP. The way presented by Rudi and Ciliberto (2021); Marteau-Ferey, Bach, and Rudi (2022a) is to sample m points $\widetilde{x}_1, ..., \widetilde{x}_m$ uniformly from the domain $\mathcal{X}$ and to use the following finite dimensional model, parametrized by $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^m)$ and defined as $\sum_{ij} B_{ij} k(x, \tilde{x}_i)k(x, \tilde{x}_j)$. This model is included in the set $\mathrm{SOS}^{\mathcal{H}}$ where $\mathcal{H}$ is the RKHS associated to the kernel k. This corresponds to uniform Nyström sampling, described in chapter 2. Note that sampling the m points uniformly does not seem to be the best idea *a priori*, and understanding good ways of selecting the centroids is also a very interesting perspective (we will detail this perspective a bit more in the next section).

Library. Finally, as is done by Lasserre (2010), it is crucial that we implement a fast, well documented library implementing PSD models in order to explore the properties of this model in different settings, relying on fast SDP solvers. This is important both to test the quality and flexibility of the method on real data sets and loss, as well as to get some insight on what the right setting is for this model.

Perspectives around the use of PSD models to do sampling

In chapter 6, we derived an algorithm to approximately sample from an un-normalized distribution with density p by modelling that density with a PSD model from which we can sample. This algorithm models the density by fitting it to a PSD model $\widetilde{p}$ using m centers $\widetilde{x}_1, ..., \widetilde{x}_m$, which are sampled uniformly at random from $\mathcal{X}$ (we look for $\widetilde{p}$ in the form $\sum_{ij} B_{ij} k(x, \tilde{x}_i)k(x, \tilde{x}_j)$).

Doing better than approximation on a uniform subset of points. Sampling the m basis points uniformly at random from $\mathcal{X}$ seems to be a bit of a waste. If, for example, the distributions p which we have to sample from has support essentially contained in a small dimensional suspace of $\mathbb{R}^d$, the m centers will probably never capture any information on the distribution p. An important open problem is therefore to design sampling methods which can identify a region in which the support of p is included.

An idea to do this is to consider a damped method. Assume that p is of the form $p_\lambda(x) = e^{-\lambda V(x)}$ where V is a regular potential and λ is an inverse-temperature parameter. A natural idea in order to sample from the normalized version of p_λ would be to create successive approximations $\widetilde{p}_t$ of $e^{-\lambda_t V}$ for a decreasing sequence of $\lambda_t \downarrow \lambda$, starting at $\lambda_0 = 0$ (in that case, the measure is uniform). To approximate $p_t := p_{\lambda_t}$ with $\widetilde{p}_t$, we would use m centers computed by sampling from the approximation $\widetilde{p}_{t-1}$ of p_{t-1}. In practice, at least in low dimensional settings ($d \leq 15$), this method seems to be quite efficient. However, we still lack proper theoretical results, and would like to explore it in future work. Note that applying such a scheme could help sample points to perform dimension reduction.

Applications. Following the previous line of thought, one interesting application of being able to sample from a density proportional to $e^{-\lambda V(x)}$ would be to sample from Gibbs measures, which are exactly of this form. This is a very important in statistical physics and chemistry, as they appear as stationnary measures (*i.e.*, limit states, see the book by Lelièvre, Rousset, and Stoltz (2010)). This is a field where Monte-Carlo Markov chain (MCMC) methods are widely used. These methods suffer from the fact that they do not automatically generate i.i.d. samples, and hence get stuck in certain regions for a long time (this is related to called the multimodal and entropy barrier problems). We can therefore try to see if our method can help in that setting, by either directly sampling from the Gibbs measure, or by providing good candidates in order to reduce the amount of samples rejected in a MCMC methods which uses rejection sampling.

One other interesting application would be to sample from a uniform measure defined on a subset of $\mathcal{X}$. This would require to abandon the C^r regularity condition, and to look instead at the regularity of the subset or a regularization of the indicator function of the subset.

A more general question : how to model constrained outputs

In part II, we have explored the modelling of a non-negativity constraint using PSD models. There are of course a lot of interesting problems where the constraint on the output is not a non-negativity constraint, but a constraint of another type. We could look for a predictor with outputs in :

- an affine cone;
- a PSD cone;
- a compact set.

Proposing a good model in these different situations proves to be quite a challenge, for which PSD models are not always adapted. Note that in the work by Marteau-Ferey, Bach, and Rudi (2020), we show that it is easy to generalize the non negative setting to the affine cone setting. Handling PSD cones has partially been done by Muzellec, Bach, and Rudi (2022). Note that in that setting, the SOS guarantees for approximation are weaker than those existing for the non-negative case.

The real difficulty, of course, is to provide a good model for outputs in compact sets. Even in the case where the compact set is of the form $[a, b]$, this is not easy. Indeed, if we use the PSD model naively, we would ask $f - a = g_1 \in \mathrm{SOS}^H$ and $b - f = g_2 \in \mathrm{SOS}^H$, and hence, in particular, $g_1 + g_2 = a + b$. This affine constraint means that we cannot apply the representer theorem.

10.2 .3 Going forward with kernel sum of squares

In part III and in particular in chapter 7, we have made a parallel between the method to perform global optimization proposed in chapter 8 and the one used in polynomial optimization. In this section, we present future directions of research which are, for the greater part, inspired by this parallel.

Towards a more complete understanding of the properties of kernel sums of squares in the manifold setting

Constrained case. As we have seen in chapter 7, constraints are necessary in the polynomial case to guarantee the convergence of the moment-SOS hierarchy. A first interesting question is whether or not it is possible to adapt Main theorem 8 to the constrained setting for kernel sum of squares. More formally, we wish to find sufficient conditions in order for a function f which is non-negative on a set $K = \{x \in \mathcal{X} \ : \ f_j(x) \geq 0, \ 1 \leq j \leq m\}$ to be written in the form :

$$f = \sigma_0 + \sum_{j=1}^{m} \sigma_j f_j, \qquad (\sigma_j)_{0 \leq j \leq m} \in \mathrm{SOS}^{\mathcal{H}}. \tag{10.7}$$

We claim this is possible under certain second order conditions which resemble those made by Marshall (2006) in the polynomial case, and which correspond to the "easy" second order condition for optimality given by Gilbert (2020) called the linear independence qualification of constraints (see p.182, QC-IL). For simplicity, we will not write down the full condition here. However, we believe that the result is true, and that we can assume that the σ_j are simply squares of functions for $1 \leq j \leq m$. The proof will rely on the introduction of a well chosen Lagrangian which incorporates only the activated constraints, and which are assumed to always be linearly independent.

Bounding the number of functions in the SOS decomposition. Going into the proof of Main theorem 8, we can see that in order to decompose a function f satisfying the assumptions of the theorem into a finite sum of squares, we glue local decompositions as the sum of d squares. While in Main theorem 8 does not provide any guarantees on the necessary number of functions, we make the following conjecture, with the same assumptions.

Conjecture 10.2: Number of functions in the decomposition

Under the same assumptions as Main theorem 8, there exists a finite number of functions $f_1, ..., f_N \in C^{p-2}(M)$ with $N = O(d^3)$ such that

$$\forall x \in M, \ f(x) = \sum_{i=1}^{N} f_i(x)^2. \tag{10.8}$$

A very rough sketch of proof of this result is the following. Take one connected component of the set of zeros, N. We use the tubular neighborhood theorem to show that there exists an open set U containing N in M such that $\nu_M N$ is diffeomorpic to U, where $\nu_M N$ is called the normal bundle to N in M. We therefore assume that f is defined on $\nu_M N$. We perform a handle decomposition (Wall, 2016) on N and show that we can build a Riemannian structure on the fibers of $\nu_M N$, such that $f(p, v) = \langle v, v \rangle_p$ by successively gluing such sructures using the handle decomposition ($\langle \cdot, \cdot \rangle_p$ denotes the riemannian metric at point p on the fiber above p). We then use the Nash embedding theorem, to show that there exists a neighborhood around each connected component of the minimizers such that f is decomposable as a sum of $O(d^3)$ squares.

We conclude by gluing intelligently all these decompositions, in particular assuming the open sets containing each connected component N are disjoint.

Moments. In Sec. 7.3 .3, we showed that we can define the cone of moments in the kernel setting, in a slightly more elaborate way than for polynomials as we do not have the nice product structure in the kernel case anymore. Moreover, when constraints are added, the problem will become even more complicated, as we will have to consider the products of three functions, and hence the space $\mathcal{H}^{\otimes 3}$, with kernel k^3. We really wish to explore this moments point of view of kernels, as we hope it will help us better understand where the extraction procedure does not work in the kernel setting. Moreover, we expect it to help us perform relaxations which keep the lower bound guarantees of the polynomial setting.

General objective

After having explained what we believe to be open problems within reach, let us take a step back, and formulate two long term goals.

- We would like to see if it is possible to find interesting classes of functions which are adapted to our problems and which provide better guarantees (*i.e.*, that are really compatible with PSD representations).
- We would like to move out of the setting of $\mathbb{R}^d$ and C^b regularity, and instead focus on different types of regularity, such as the low dimensionality of the support for example.

Can we provide good a posteriori guarantees ?

As we have seen in chapter 7, the fact that we sample the equality constraint in the kernel SOS setting kills the guarantees we used to have at the step before, and for polynomials, which is that we have a lower bound of the original problem.

An interesting research direction in this context is to see if we can adapt the moment formulation or the sum of squares formulation to keep a guarantee, and even have stronger ones (upper bounds and lower bounds). This is done in a work by Woodworth, Bach, and Rudi (2022) in the case where we have access to the discrete Fourier transform of the function we want to minimize. Indeed, in this work, all errors (optimization, statistical) are explicitely controlled and allow to derive a certificate whose bound is fairly tight.

Applying the method to typical problems tackled by moment-SOS hierarchies in the polynomial case

In the book by Lasserre (2010), there is an entire section where polynomial optimization is used to tackle certain problems, ranging from optimal control to probability theory. In particular, these problems can usually be formulated as optimization problems on non-negative measures, or optimization problems with non-negativity constraints on a function (*i.e.*, we ask for a measure to be non-negative or a function to be non-negative).

Berthier, Carpentier, Rudi, and Bach (2021) have started applying the kernel sum of squares method to optimal control. This is quite difficult, as the optimal control problems are usually highly irregular. On the bright side, it is almost impossible to solve them in an efficient way, and so it is relevant to try to apply our method even if the needed guarantees do not hold *a priori*.

Without making a list of the different field of applications which can be found in the book by Lasserre (2010), we can cite the particular example of finding Lyapunov functions for certain dynamics using kernel sum of squares, which we would like to study in the future.

Making a toolbox

Finally, as we can see with the development of gloptipoly, building a good toolbox with our tools for global optimization will be necessary both for us, in order to see if we are really able to tackle real-life problems with an off the shelf method, and for users if they are to try it in a user friendly way. However, we believe we must wait for the method to be slightly more mature to start implementing and publishing this. In particular, we have to see if there is no way of having a posteriori guarantees, and implement automatic ways to tune the hyperparameter. Finally, we must incorporate the constrained setting from an algorithmic point of view, in order to deal with these problems as well.

References

Murat A. Erdogdu and Andrea Montanari. Convergence rates of sub-sampled Newton methods. Technical Report 1508.02810, ArXiv, 2015.

Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(Mar):803–826, 2009.

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Elsevier, 2003.

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *J. Mach. Learn. Res.*, 18(1):4148–4187, jan 2017.

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

M. Montaz Ali, Charoenchai Khompatraporn, and Zelda B. Zabinsky. A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. *Journal of Global Optimization*, 31(4):635–672, 2005. doi: 10.1007/s10898-004-9972-2. URL https://doi.org/10.1007/s10898-004-9972-2.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the Symposium on Theory of Computing*, pages 1200–1205, 2017.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Emil Artin. Über die Zerlegung definiter Funktionen in Quadrate. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 5:100–115, 1927.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017a.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017b.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with

convergence rate $O(1/n)$. *CoRR*, abs/1306.2119, 2013. URL http://arxiv.org/abs/1306.2119.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with Sparsity-Inducing Penalties. *Foundations and Trends in Machine Learning*, pages –, 2011. doi: 10.1561/2200000015. URL https://hal.archives-ouvertes.fr/hal-00613125.

Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6, 2004.

J. Andrew Bagnell and Amir-massoud Farahmand. Learning positive functions in a Hilbert space. In *NIPS Workshop on Optimization (OPT2015)*, pages 3240–3255, 2015.

P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(1):4308, 2014. doi: 10.1038/ncomms5308. URL https://doi.org/10.1038/ncomms5308.

Richard E. Barlow and Hugh D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

M. S. Bartlett. Approximate confidence intervals. *Biometrika*, 40(1/2):12–19, 1953.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4), aug 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214{%}2F009053605000000282.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1903070116.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

Eloïse Berthier, Justin Carpentier, Alessandro Rudi, and Francis Bach. Infinite-Dimensional Sums-of-Squares for Optimal Control. working paper or preprint, October 2021. URL https://hal.archives-ouvertes.fr/hal-03377120.

Gaspard Beugnot, Julien Mairal, and Alessandro Rudi. Beyond tikhonov: faster learning with self-concordant losses, via iterative regularization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=7_M2f2DEIEK.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

Mathieu Blondel, Akinori Fujino, and Naonori Ueda. Convex factorization machines. In *Joint*

European Conference on Machine Learning and Knowledge Discovery in Databases, pages 19–35. Springer, 2015.

Raghu Bollapragada, Richard H. Byrd, and Jorge Nocedal. Exact and inexact subsampled newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2018.

Howard D. Bondell, Brian J. Reich, and Huixia Wang. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

Jean-Michel Bony. Sur l'inégalité de Fefferman-Phong. *Séminaire Équations aux dérivées partielles (Polytechnique) dit aussi "Séminaire Goulaouic-Schwartz"*, 1998-1999. URL http://www.numdam.org/item/SEDP_1998-1999____A3_0/. talk:3.

Jean-Michel Bony. Sommes de carrés de fonctions dérivables. *Bulletin de la Société Mathématique de France*, 133(4):619–639, 2005. URL http://bonyjm.perso.math.cnrs.fr/S2KR_bullSMF.pdf.

Jean-Michel Bony, Fabrizio Broglia, Ferruccio Colombini, and Ludovico Pernazza. Nonnegative functions as squares or sums of squares. *Journal of Functional Analysis*, 232(1):137–147, 2006.

Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-Lopez, Leslie O'Bray, and Bastian Rieck. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712, 2020. doi: 10.1561/2200000076. URL https://doi.org/10.1561{%}2F2200000076.

Jonathan M. Borwein and Adrian S. Lewis. *Convex analysis and nonlinear optimization: theory and examples; 2nd ed.* CMS Books in Mathematics. Springer, Dordrecht, 2010. URL http://cds.cern.ch/record/1616007.

Léon Bottou and Olivier Bousquet. The trade-offs of large scale learning. In *Advances in Neural Information Processing systems*, pages 161–168, 2008.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Stéphane Boucheron and Pascal Massart. A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150(3-4):405–433, 2011.

Stephane Boucheron, Olivier Bousquet, and Gabor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Stéphane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: a non Asymptotic Theory of Independence*. Oxford University Press, 2013. URL https://hal.inria.fr/hal-00942704.

Christos Boutsidis and Alex Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1301–1340, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Susanne Brenner and Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15. Springer Science & Business Media, 2007.

Haïm Brezis and Petru Mironescu. Gagliardo-nirenberg, composition and products in fractional sobolev spaces. 2001.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8. URL https://doi.org/10.1007/s10208-006-0196-8.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Reproducing kernel Hilbert spaces and Mercer theorem. *arXiv Mathematics e-prints*, art. math/0504071, April 2005.

Nicolo Cesa-Bianchi, Yishay Mansour, and Ohad Shamir. On the complexity of learning with kernels. In *Conference on Learning Theory*, pages 297–325, 2015.

Benjamin Charlier, Jean Feydy, Joan Alexis Glaunes, FranÃ§ois-David Collin, and Ghislain Durif. Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021. URL http://jmlr.org/papers/v22/20-275.html.

Caroline Chaux, Patrick L. Combettes, Jean-Christophe Pesquet, and Valérie R. Wajs. A variational formulation for frame-based inverse problems. *Inverse Problems*, 23(4):1495–1518, 2007.

Elliott Ward Cheney and William Allan Light. *A Course in Approximation Theory*, volume 101. American Mathematical Soc., 2009.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

Aaron Defazio. A simple practical accelerated method for finite sums. In *Advances in Neural Information Processing Systems*, pages 676–684, 2016.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing systems*, pages 1646–1654, 2014.

P. Del Moral and A. Niclas. A Taylor expansion of the square root matrix function. *Journal of Mathematical Analysis and Applications*, 465(1):259 – 266, 2018.

Peter Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*. Springer, 2011.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101): 1–51, 2017. URL http://jmlr.org/papers/v18/16-335.html.

Petros Drineas, Michael W Mahoney, Shan Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media, 1996.

C. Fefferman and D. H. Phong. On positivity of pseudo-differential operators. *Proceedings of the National Academy of Sciences*, 75(10):4673–4674, 1978. ISSN 0027-8424. doi: 10.1073/pnas.75.10.4673. URL https://www.pnas.org/content/75/10/4673.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. *Proceedings of COLT*, 2018.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

L. Lo Gerfo, Lorenzo Rosasco, Francesca Odone, E. De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

Jean-Charles Gilbert. Fragments d'optimisation différentiable – théorie et algorithmes, 2020. URL https://hal.inria.fr/hal-03347060/document. Syllabus de cours à l'ENSTA, Paris.

Israel Gohberg, Seymour Goldberg, and Marinus A. Kaashoek. *Basic Classes of Linear Operators*. Birkhäuser, Basel, 2004.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

Robert Gower, Filip Hanzely, Peter Richtárik, and Sebastian U. Stich. Accelerated stochastic matrix inversion: general theory and speeding up BFGS rules for faster second-order optimization. In *Advances in Neural Information Processing Systems*, pages 1619–1629, 2018.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

Zaid Harchaoui and Francis Bach. Image classification with segmentation graph kernels. pages 1–8, 07 2007. ISBN 1-4244-1180-7. doi: 10.1109/CVPR.2007.383049.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: tight bounds for stochastic and online optimization. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 197–209, 2014.

J. William Helton and Jiawang Nie. Semidefinite representation of convex sets. *Mathematical Programming*, 122(1):21–64, 2010. doi: 10.1007/s10107-008-0240-y. URL https://doi.org/10.1007/s10107-008-0240-y.

Didier Henrion, Jean-Bernard Lasserre, and Johan Löfberg. Gloptipoly 3: Moments, optimization and semidefinite programming. *Optimization Methods and Software*, 24, 10 2007. doi: 10.1080/10556780802699201.

Didier Henrion, Milan Korda, and Jean Bernard Lasserre. *The Moment-SOS Hierarchy*. WORLD SCIENTIFIC (EUROPE), 2020. doi: 10.1142/q0252. URL https://www.worldscientific.com/doi/abs/10.1142/q0252.

David Hilbert. Ueber die Darstellung definiter Formen als Summe von Formenquadraten. *Mathematische Annalen*, 32(3):342–350, 1888. doi: 10.1007/BF01443605. URL https://doi.org/10.1007/BF01443605.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods*, 5(1):77–88, 1976.

Lars Hörmander. Hypoelliptic second order differential equations. *Acta Mathematica*, 119: 147–171, 1967.

Lars Hörmander. *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Springer, 2015.

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, 2012.

L. Hörmander. *The Analysis of Linear Partial Differential Operators III*. Classics in Mathematics. Springer, Berlin, 2007. ISBN 978-3-540-49937-4. Pseudo-differential operators, Reprint of the 1994 edition.

Viktor V. Ivanov. On optimum minimization algorithms in classes of differentiable functions. In *Doklady Akademii Nauk*, volume 201, pages 527–530. Russian Academy of Sciences, 1971.

T. Jaakkola, Mark E. Diekhans, and David Haussler. Using the fisher kernel method to detect remote protein homologies. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, pages 149–58, 1999.

Momin Jamil and Xin She Yang. A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2):150, 2013. doi: 10.1504/ijmmno.2013.055204.

Klaus Jänich. *Topologie*. Springer-Verlag, Berlin-New York, 1980. ISBN 3-540-10183-7.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.

Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. Online and stochastic gradient methods for non-decomposable loss functions. *Advances in Neural Information Processing Systems*, 1, 10 2014.

Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Global linear convergence of newton's method without strong-convexity or lipschitz gradients. *CoRR*, abs/1806.00413, 2018. URL http://arxiv.org/abs/1806.00413.

S. Sathiya Keerthi, K. B. Duan, Shirish Krishnaj Shevade, and Aun Neow Poo. A fast dual algorithm for kernel logistic regression. *Machine learning*, 61(1-3):151–165, 2005.

J. Thomas King and David Chillingworth. Approximation of generalized inverses by iterated regularization. *Numerical Functional Analysis and Optimization*, 1(5):499–513, 1979. doi: 10.1080/01630567908816031. URL https://doi.org/10.1080/01630567908816031.

Etienne Klerk and Monique Laurent. On the lasserre hierarchy of semidefinite programming relaxations of convex polynomial optimization problems. *SIAM Journal on Optimization*, 21: 824–832, 07 2011. doi: 10.1137/100814147.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 1477–1485, 2015.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

Jacques Lafontaine. *An introduction to differential manifolds*. Springer, 2015.

Jean B. Lasserre. Convexity in semi algebraic geometry and polynomial optimization. *SIAM Journal on Optimization*, 19(4):1995–2014, 2009. doi: 10.1137/080728214. URL https://doi.org/10.1137/080728214.

Jean-Bernard Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

Jean-Bernard Lasserre. A sum of squares approximation of nonnegative polynomials. *SIAM Review*, 49(4):651–669, 2007.

Jean-Bernard Lasserre. *Moments, Positive Polynomials and their Applications*, volume 1. World Scientific, 2010.

Jean-Bernard Lasserre. A new look at nonnegativity on closed sets and polynomial optimization. *SIAM Journal on Optimization*, 21(3):864–885, 2011.

Jean-Bernard Lasserre, Kim-Chuan Toh, and Shouguang Yang. A bounded degree SOS hierarchy for polynomial optimization. *EURO Journal on Computational Optimization*, 5(1-2):87–117, 2017.

Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

Monique Laurent and Lucas Slot. An effective version of schmüdgen's positivstellensatz for the hypercube, 2021. URL https://arxiv.org/abs/2109.09528.

Quoc V. Le, Alex J. Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496, 2005.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. *Free Energy Computations*. IMPERIAL COLLEGE PRESS, 2010. doi: 10.1142/p579. URL https://www.worldscientific.com/doi/abs/10.1142/p579.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Publishing Company, Incorporated, 2008. ISBN 0387763694.

Grace Lo Yang (auth.) Lucien Le Cam. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer US, 1990. ISBN 9781468403794; 1468403796; 9781468403770; 146840377X.

Molga M. and Smutnicki C. Test functions for optimization needs, 2005.

Murray Marshall. Representations of non-negative polynomials having finitely many zeros. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 15(3):599–609, 2006. doi: 10.5802/afst.1131. URL afst.centre-mersenne.org/item/AFST_2006_6_15_3_599_0/.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Newton methods for ill-conditioned generalized self-concordant losses. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/60495b4e033e9f60b32a6607b587aadd-Paper.pdf.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2294–2340. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/marteau-ferey19a.html.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12816–12826. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/968b15768f3d19770471e9436d97913c-Paper.pdf.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Sampling from arbitrary functions via psd models. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2823–2861. PMLR, 28–30 Mar 2022a. URL https://proceedings.mlr.press/v151/marteau-ferey22a.html.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Second order conditions to decompose smooth functions as sums of squares, 2022b. URL https://arxiv.org/abs/2202.13729.

Swann Marx, Edouard Pauwels, Tillmann Weisser, Didier Henrion, and Jean Lasserre. Semi-algebraic approximation using Christoffel-Darboux kernel. Technical Report 1904.01833, ArXiv, 2019.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: Handling billions of points efficiently. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14410–14422. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a59afb1b7d82ec353921a55c579ee26d-Paper.pdf.

Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. Technical Report 1605.01288, ArXiv, 2016.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. 2019. doi: 10.48550/ARXIV.1908.05355. URL https://arxiv.org/abs/1908.05355.

J. Mercer. Functions of positive and negative type, and their connection with the theory of

integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446, 1909.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

John W. Milnor. *Morse theory*. Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51. Princeton University Press, Princeton, N.J., 1963. URL http://www.maths.ed.ac.uk/~aar/papers/milnmors.pdf.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN 978-0-262-03940-6.

E. H. Moore. Abstract for "On properly positive Hermitian matrices". *Bull. Am. Math. Soc.*, 23(2):66–67, 1916. ISSN 0002-9904. URL http://www.ams.org/journals/bull/1916-23-02/S0002-9904-1916-02866-7/S0002-9904-1916-02866-7.pdf. Unpublished address. JFM:46.0165.03.

Eliakim Hastings Moore. *General analysis*, volume 1 of *Memoirs of the American Philosophical Society*. American Philosophical Society, Philadelphia, 1935. Edited by R. W. Barnard. Zbl:0013.11605. JFM:61.0433.06.

Jaouad Mourtada and Stéphane Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research*, 23(31):1–49, 2022. URL http://jmlr.org/papers/v23/20-782.html.

Boris Muzellec, Francis Bach, and Alessandro Rudi. Learning psd-valued functions using kernel sums-of-squares, 2022.

Francis J. Narcowich, Joseph D. Ward, and Holger Wendland. Refined error estimates for radial basis function interpolation. *Constructive Approximation*, 19(4):541–564, 2003.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 2004.

Yuri E Nesterov. A method for solving the convex programming problem with convergence rate o(1/k$\hat{2}$). *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983. URL https://cir.nii.ac.jp/crid/1571135650017614976.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

Jiawang Nie. Optimality conditions and finite convergence of Lasserre's hierarchy. *Mathematical Programming*, 146(1-2):97–121, 2014.

Jiawang Nie and Markus Schweighofer. On the complexity of Putinar's Positivstellensatz. *Journal of Complexity*, 23(1):135–150, 2007. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2006.07.002. URL https://www.sciencedirect.com/science/article/pii/S0885064X0600080X.

Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, 1992.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006.

Erich Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349. Springer, 2006.

Erich Novak and Henryk Woźniakowski. *Tractability of Multivariate Problems: Standard Information for Functionals*, volume 12. European Mathematical Society, 2008.

Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

Roland Opfer. Multiscale kernels. *Advances in Computational Mathematics*, 25(4):357–380, 2006. doi: 10.1007/s10444-004-7622-3. URL https://doi.org/10.1007/s10444-004-7622-3.

Michael A. Osborne, Roman Garnett, and Stephen J. Roberts. Gaussian processes for global optimization. In *International Conference on Learning and Intelligent Optimization (LION3)*, pages 1–15, 2009.

Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. Technical Report 1810.06838, arXiv, 2018.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, jan 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL https://doi.org/10.1561/2400000003.

Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003. doi: 10.1007/s10107-003-0387-5. URL https://doi.org/10.1007/s10107-003-0387-5.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zach DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Frédéric Paulin. Géométrie différentielle élémentaire, 2006. URL https://www.imo.universite-paris-saclay.fr/~paulin/notescours/cours_geodiff.pdf.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge University Press, 2016.

Mathew Penrose. *Random geometric graphs*, volume 5. Oxford University Press, 2003.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 8125–8135, 2018.

Mihai Putinar. Positive Polynomials on Compact Semi-algebraic Sets. *Indiana Univ. Math. J.*, 42:969–984, 1993. ISSN 0022-2518.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*, 2015.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Michael Reed. *Methods of Modern Mathematical Physics: Functional Analysis*. Elsevier, 1980.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Farbod Roosta-Khorasani and Michael W. Mahoney. Sub-sampled Newton methods. *Math. Program.*, 174(1-2):293–326, 2019.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf.

Alessandro Rudi and Carlo Ciliberto. Psd representations for effective probability models. *Advances in Neural Information Processing Systems*, 34, 2021.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in Neural Information Processing Systems*, 30:3215–3225, 2017.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898, 2017.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pages 5672–5682, 2018.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations, 2020. URL https://arxiv.org/abs/2012.11978.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2003. ISBN 0898715342.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer International Publishing, 2015. doi: 10.1007/978-3-319-20828-2. URL https://doi.org/10.1007/978-3-319-20828-2.

Konrad Schmügden. The k-moment problem for compact semi-algebraic sets. *Mathematische Annalen*, pages 203–206, 1991.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Winfried Sickel. Superposition of functions in sobolev spaces of fractional order. a survey. *Banach Center Publications*, 27:481–497, 1992.

Lucas Slot. Sum-of-squares hierarchies for polynomial optimization and the christoffel-darboux kernel, 2021. URL https://arxiv.org/abs/2111.04610.

Lucas Slot and Monique Laurent. Near-optimal analysis of Lasserre's univariate measure-based bounds for multivariate polynomial optimization. *Mathematical Programming*, pages 1–18, 2020a.

Lucas Slot and Monique Laurent. Improved convergence analysis of lasserre's measure-based upper bounds for polynomial minimization on compact sets. *Mathematical Programming*, 01 2020b. doi: 10.1007/s10107-020-01468-3.

Lucas Slot and Monique Laurent. Near-optimal analysis of lasserre's univariate measure-based bounds for multivariate polynomial optimization. *Mathematical Programming*, 188 (2):443–460, 2021. doi: 10.1007/s10107-020-01586-y. URL https://doi.org/10.1007/s10107-020-01586-y.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume One. Publish or Perish, Inc., Houston, Texas, third edition, 1999.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552. 2009.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL http://jmlr.org/papers/v11/sriperumbudur10a.html.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proc. COLT*, 2009.

Gilbert Stengle. A Nullstellensatz and a Positivstellensatz in Semialgebraic Geometry. *Mathematische Annalen*, 207:87–98, 1974. URL http://eudml.org/doc/162533.

T. Sun and Q. Tran-Dinh. Generalized Self-Concordant Functions: A Recipe for Newton-Type Methods. *ArXiv e-prints*, March 2017.

Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

Ichiro Takeuchi, Quoc Le, Timothy Sears, and Alexander Smola. Nonparametric quantile regression. *Journal of Machine Learning Research*, 01 2005.

M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28 – 76, 1994. doi: 10.1214/aop/1176988847. URL https://doi.org/10.1214/aop/1176988847.

Daniel Tataru. On the fefferman–phong inequality and related problems. *Communications in Partial Differential Equations*, 27(11-12):2101–2138, 2002. doi: 10.1081/PDE-120016155. URL https://doi.org/10.1081/PDE-120016155.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society series b-methodological*, 58:267–288, 1996.

Ilya O Tolstikhin, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/5055cbf43fac3f7e2336b27310f0b9ef-Paper.pdf.

Hans Triebel. *Theory of Function Spaces III*, volume 100. Birkhäuser Basel, 2006.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Stephen Tu, Rebecca Roelofs, Shivaram Venkataraman, and Benjamin Recht. Large scale kernel learning using block coordinate descent. *arXiv preprint arXiv:1602.05310*, 2016.

Levent Tuncel. *Potential Reduction and Primal-Dual Methods*, pages 235–265. Springer US, Boston, MA, 2000. ISBN 978-1-4615-4381-7. doi: 10.1007/978-1-4615-4381-7_9. URL https://doi.org/10.1007/978-1-4615-4381-7_9.

Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A Dimension-free Computational Upper-bound for Smooth Optimal Transport Estimation. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4143–4173. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/vacher21a.html.

Sara A. Van de Geer. High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Tim Van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Grace Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

C. T. C. Wall. *Theory of handle decompositions*, page 129–166. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016. doi: 10.1017/CBO9781316597835.006.

Joachim Weidmann. *Linear Operators in Hilbert Spaces*, volume 68. Springer Science & Business Media, 1980.

Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*, volume 2. MIT Press, 2006.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

Blake Woodworth, Francis Bach, and Alessandro Rudi. Non-convex optimization with certificates and fast rates through kernel sums of squares, 2022. URL https://arxiv.org/abs/2204.04970.

Yingxiang Yang, Haoxiang Wang, Negar Kiyavash, and Niao He. Learning positive functions with pseudo mirror descent. In *Advances in Neural Information Processing Systems*, pages 14144–14154, 2019.

Ming Yuan and Grace Wahba. Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & Probability Letters*, 69(1):11–20, 2004.

Vadim Yurinsky. *Sums and Gaussian vectors, volume 1617 of Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

S Zaremba. L'équation biharmonique et une class remarquable de functions fondamentales harmoniques. *Bull. Int. Acad. Sci. Cracovie*, 3:147–196, 1907.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, 2008.

Kaiwen Zhou, Fanhua Shang, and James Cheng. A simple stochastic variance reduced algorithm with fast convergence rates. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5980–5989. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/zhou18c.html.

Kaiwen Zhou, Qinghua Ding, Fanhua Shang, James Cheng, Danli Li, and Zhi-Quan Luo. Direct acceleration of saga using sampled negative momentum. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1602–1610. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/zhou19c.html.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B (Statistical Methodology)*, 67(2):301–320, 2005.

RÉSUMÉ

Cette thèse se situe à l'interaction entre les domaines de l'apprentissage statistique, de l'optimisation, et de la modélisation. Nous étudions différents modèles de fonctions, fondés sur les espaces de Hilbert à noyau reproduisant.
Nous commençons par étudier les propriétés statistiques et algorithmiques de l'estimateur de minimisation du risque empirique régularisé, dans le cadre de la minimisation du risque (en espérance) sur des espaces à noyau reproduisant. Nous généralisons la décomposition biais-variance très précise bien connue dans le cadre de la régression linéaire à une plus grande classe de fonctions en controlant précisément l'évolution locale de leurs Hessiennes. Cette classe de fonctions appelées fonctions autoconcordante généralisées contient la perte logistique, qui est très utilisée en classification. Nous étendons également les taux de convergence rapides ainsi que les algorithmes rapides établis dans le cadre de la régression linéaire à cette classe de fonctions, caractérisant au passage la difficulté du problème statistique sous-jacent à l'aide de deux quantités caractérisant la régularité de la solution, ainsi que celle de l'espace à noyau reproduisant utilisé.
Dans un deuxième temps, nous introduisons un modèle pour les fonctions positives, qui est linéairement paramétré par un opérateur symmetrique sur un espace à noyau reproduisant, et où la contrainte de positivité est garantie au moyen d'une contrainte de positivité semidéfinie sur cet opérateur. Nous étudions les propriétés de ce modèle, et montrons qu'il a de bonnes propriétés d'approximation, qu'il préserve la convexité des fonctions de perte standard en apprentissage statistique, et qu'il est intégrable et différentiable en forme close lorsqu'il est défini avec des noyaux particuliers. En outre, cette dernière propriété nous permet d'utiliser ce modèle pour estimer des densités de probabilité, et nous développons un algorithme d'échantillonage d'une densité de probabilité quelconque, accessible à partir de ses valeurs en certains points.
Enfin, nous appliquons ce modèle de fonctions positives afin d'approcher le minimum global d'une fonction régulière. Dans le même esprit que les hiérarchies de Lasserre pour l'optimisation polynomiale, nous cherchons le minimum global d'une fonction comme sa plus grande borne inférieure, et affinons les garanties d'approximation du paragraphe précédent afin d'obtenir un algorithme quasi-optimal pour trouver le minimum global d'une fonction à partir de ses valeurs en un nombre donné de points. Nous établissons également des résultats théoriques permettant de généraliser cette approche à des fonctions définies sur des variétés.

MOTS CLÉS

Apprentissage statistique, estimation non paramétrique, espaces de Hilbert à noyau reproduisant, modèles semidéfinis positifs, sommes de carrés, optimisation globale, méthode du deuxième ordre, méthode de Newton, conditions du deuxième ordre, échantillonnage.

ABSTRACT

In this thesis at the boundary between statistical learning, optimization and modelling, we study different non-parametric models for functions based on reproducing kernel Hilbert spaces.
We first study the expected risk minimization problem through the solving of the regularized empirical risk minimization problem on a reproducing kernel Hilbert space. We extend the precise bias-variance trade offs known for least squares regression to a broader class of functions by controlling the evolution of their Hessians. This class of functions, called generalized self-concordant functions, includes the logistic loss, which is widely used for classification. We also extend the fast rates and fast algorithms known for least squares to this class of functions, characterizing the difficulty of a problem through interpretable quantities.
We then introduce a model for non-negative functions, which is linearly parametrized by a symmetric operator and which enforces non-negativity through a positive semidefinite constraint on this operator. We study the properties of this model and show that it has good approximation properties, preserves the convexity of machine learning loss functions, and is integrable in closed form in certain cases. In particular, this last property allows us to use this model to approximate probability densities, and we develop a sampling algorithm based on these models to sample from an arbitrary probability density known via function evaluations.
Finally, we apply this model for non-negative functions in order to approximate the global minimum of a function with certain regularity properties. In the same spirit as moment-SOS hierarchies for polynomial optimization, we look for the global minimum as the maximum lower bound of the function, and refine our approximation guarantees from the previous paragraph in order to obtain a near optimal algorithm to solve global optimization given a fixed number of function evaluations. We also pave the way to generalize this approach to functions defined on manifolds, by introducing a second order sufficient condition for a non-negative regular function to be decomposable as sums of squares of regular functions.

KEYWORDS

Statistical learning, non-parametric estimation, reproducing Kernel Hilbert spaces, PSD models, sum of squares, global optimization, second order methods, Newton methods, second order conditions, sampling.