Master's Thesis

Ulysse Marteau-Ferey, under the direction of Francis Bach and Alessandro Rudi

Abstract

Reproducing Kernel Hilbert Spaces (RKHS) provide a rigorous functional analysis framework to perform non-parametric learning. Kernel methods, optimization methods in these spaces, enjoy very nice statistical properties. However, up to recently, these methods scaled very badly in the number of data points, both in time and memory requirements, hence their limited applicability. However, the FALKON algorithm proposed by Rudi et al. has made a huge step in scaling down these requirements in the case of classical least-square regression, namely scaling the complexity to $O(n\sqrt{n})$ in time and $O(n)$ in memory while keeping optimal statistical properties. Our aim in this talk is to explore the possible extensions of the ideas behind Falkon to more complex loss functions, such as the logistic loss. These methods rely two main ideas: 1) reduction of the feature space using random projections and 2) the use of iterative solvers, combined with a good pre-conditioning

Contents

Chapter 1

Introduction

Large-scale supervised learning problems are ubiquitous in machine learning. Their goal is usually to learn from examples a function which generalizes well, i.e. which predicts well new data. Linear and parametric models are often very limited and do not allow to learn complex functions; it is therefore crucial to have tractable non-parametric methods. Among them, kernel methods are probably the ones with the best theoretical guarantees; however, their applicability to large-scale problems is still fairly limited as they scale very badly in the number of data points.

Overcoming these difficulties has led to a variety of practical approaches, such as stochastic methods and pre-conditioned extensions,as well as random projections to reduce the time complexity. Random projections have also helped reduce the memory costs; they include methods like Nystrom [4] or random features [6].

From a theoretical point of view, the key question has been keeping the balance between statistical accuracy and computational gain. The main trade-off element appears to be the intrinsic dimension of the problem, which is a way to formalize the way the complexity of the problem scales in terms of the number of data points.

Recently, many significant steps have been made to reduce both time and memory complexities in the least squares error case. Recent results have shown that using both ridge regression and random features or Nystrom, one can keep the statistical optimality of Kernel ridge regression while keeping the time complexity under $O(n^2)$. The last very significant step was made in [5], combining both Nystrom sampling and pre-conditioning strategies to achieve statistical optimality with complexities of $O(n\sqrt{n})$ in time and $O(n)$ in memory.

During this internship, our aim was to use the FALKON algorithm as a starting point to develop fast algorithms for other losses than the least square loss, for instance logistic or robust losses. Using the two main components of FALKON : Nystrom sub-sampling and using an iterative solver with a pre-conditioner, we tried developing general methods for quadratic problems, and apply them to perform second order methods on these more complex loss functions. We succeeded in doing the former but still have not solved the latter problem. However, we have numerical experiments which suggest that generalizing these methods should be possible from a theoretical point of view.

The rest of the report will be organized as follows: in section 2, we will present general information on kernel methods and motivations for extending FALKON to more general cases; in section 3, we present a modified version of the FALKON algorithm to deal with more general quadratic functions; in section 4, we present a statistical analysis of FALKON to quadratic losses and explain how we want to perform second order methods as well as the limitations of our approach for the moment. Finally, in section 5, we present a few experiments which a) show that we indeed achieve optimal statistical accuracy in the quadratic case and b) that we have hope of achieving good rates for certain losses such as the logistic regression.

Chapter 2

Backround

In this chapter, we introduce the backround of our work. In section 2.1, we introduce the basic framework and aim of non-parametric supervised learning. We then present kernel methods in section 2.2 more in detail, explaining which estimators they approximate as well as the basic computational difficulties they entail. In section 2.3, we present large-scale methods for the square loss and in particular FALKON, the method on which we will build throughout the rest of the report.

2.1 Supervised learning

The aim of supervised machine learning problems is, given $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, to learn a good predictor $\theta : \mathcal{X} \to \mathcal{Y}$ such that for any new $x \in \mathcal{X}$, $\theta(x)$ is a good prediction for the corresponding y. Throughout this report, we will make the following hypotheses to put this learning problem in a clear mathematical setting.

- the data space $\mathcal{X}$ is a polish space, the target space $\mathcal{Y} = \mathbb{R}$ and the observation space is $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Moreover, we represent the data $\mathcal{X}$ in a Hilbert space $\mathcal{H}$ with the function $\phi : \mathcal{X} \to \mathcal{H}$;
- the observations $z_i = (x_i, y_i) \in \mathcal{Z}$ are i.i.d. samples from the law of $Z = (X, Y)$ which is unknown; we assume that $||\phi(X)||_{\mathcal{H}} \leq \kappa$ almost surely for a certain constant κ. Moreover, we note $\theta_i = \phi(x_i) \in \mathcal{H}$ the representation of the i-th data point;
- the loss function $l : (z, y') \in \mathcal{Z} \times \mathbb{R} \to \mathbb{R}$ is made to depend on the full observation $z = (x, y)$ in the first coordinate and compares y' to the objective y. We assume that for all $z \in \mathcal{Z}$, the function $l(z, \cdot)$ is convex, three times differentiable.

Here, we see an element $\theta \in \mathcal{H}$ as a function on $\mathcal{X}$, by identifying $\forall x \in \mathcal{X},\ \theta(x) := \langle \theta, \phi(x) \rangle$: we can see $\mathcal{H}$ as a Hilbert space of predictors. Our aim is to find the best predictor, i.e. to minimize the *expected risk* $\mathcal{E}(\theta) = \mathbb{E}\left[l(Z, \theta(X))\right]$. Therefore, we aim at solving the following problem:

$$\inf_{\theta \in \mathcal{H}} \mathcal{E}(\theta),\ \mathcal{E}(\theta) = \mathbb{E}\left[l(Z, \langle \theta, \phi(X) \rangle)\right] \tag{$\mathcal{P}$}$$

Given $(x_1, y_1), ..., (x_n, y_n)$ i.i.d. samples from the law of (X, Y), our aim will be finding an estimator $\hat{\theta}$ with small *excess risk* $\mathcal{R}$:

$$\mathcal{R}(\hat{\theta}) := \mathcal{E}(\hat{\theta}) - \inf_{\theta \in \mathcal{H}} \mathcal{E}(\theta)$$

We give here some examples of loss functions which we may comment during the report. Recall that $l : \mathcal{Z} \times \mathbb{R} \to \mathbb{R}$ where an element $z \in \mathcal{Z}$ is of the form (x, y).

Examples of losses

- the square loss $l(z, y') = (y - y')^2$ which is a particular case of
- the weighed square loss $l(z, y') = w(z)(y - y')^2$ where w is a positive function on the support of Z;
- the robust regression : $l(z, y') = \sqrt{1 + (y - y')^2}$;

- loss functions of the type $l(z, y') = \varphi(yy')$ where φ is a three times differentiable non-negative convex function;
- the logistic regression which is an instance of the previous point with $\varphi(t) := \log(1 + \exp(-t))$.

Note on the representation $\phi : \mathcal{X} \to \mathcal{H}$: RKHS

A particular way of embedding the space $\mathcal{X}$ in a Hilbert space can be constructed as follows.

A continuous function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite kernel on $\mathcal{X}$ if for every $n \in \mathbb{N}$, for every $x_1, ..., x_n \in \mathcal{X}^n$, the matrix $(K(x_i, x_j))_{1 \leq i,j \leq n}$ is positive semi-definite.

Given a positive definite kernel K on a space $\mathcal{X}$, we can define $K_x := K(x, \cdot)$ for all $x \in \mathcal{X}$, and $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ the Hilbert space associated to the inner product defined by K that is

$$\mathcal{H} = \overline{\text{span}\,\{K_x \mid x \in \mathcal{X}\}},\ \forall x, x' \in \mathcal{H},\ \langle K_x, K_{x'} \rangle = K(x, x')$$

(This definition can be formalized, see [1])

The mapping $\phi : x \in \mathcal{X} \mapsto K_x \in \mathcal{H}$ embeds the space $\mathcal{X}$ in the Hilbert space $\mathcal{H}$ and this embedding reflects the similarity measure K. Note that $\mathcal{H}$ is a space of functions on $\mathcal{X}$. Moreover, we have that $\forall \theta \in \mathcal{H},\ \forall x \in \mathcal{X},\ \langle \theta, \phi(x) \rangle_{\mathcal{H}} = \langle \theta, K_x \rangle_{\mathcal{H}} = \theta(x)$. Moreover, for any $x, x' \in \mathcal{X},\ \phi(x) \cdot \phi(x') = K(x, x')$.

Thus, computing any coefficient of a matrix of the type $(\phi(x_i) \cdot \phi(\tilde{x}_j))_{\substack{1 \leq i \leq n \\ 1 \leq j \leq M}}$ is in the same complexity as a so-called *kernel evaluation*, which can be very fast and only uses the function K. In practice, all the matrices we will consider in our algorithms will be products of matrices of this form.

2.2 Regularized ERM and Kernel Ridge Regression

In this section, we present the main estimators we will consider for θ_*, the solution of our problem of interest $(\mathcal{P})$, obtained through regularized Expected Risk Minimization (sub-section 2.2.1). We will present these estimators as solutions to finite-dimensional convex problems (see sub-section 2.2.2) and make this problem explicit in the case of the Kernel Ridge Regression, the problem associated to the squared loss (see sub-section 2.2.3). This will illustrate the computational difficulties which motivated the introduction of the tools in the next section.

2.2.1 Finding a good estimator : Expected Risk Minimization

Recall that the aim is to find a good estimator of the solution to $(\mathcal{P})$:

$$\inf_{\theta \in \mathcal{H}} \mathcal{E}(\theta),\ \mathcal{E}(\theta) = \mathbb{E}\left[l(Z, \langle \theta, \phi(X) \rangle)\right] \tag{$\mathcal{P}$}$$

Natural estimators $\hat{\theta}$ come from the minimizing of the so-called empirical risk i.e. the solution to

$$\hat{\theta}_*^n \in \arg\min_{\theta} \hat{E}_n(\theta),\ \hat{E}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} l(z_i, \langle \theta_i, \theta \rangle) \tag{$\hat{\mathcal{P}}_n$}$$

However, since $\hat{E}_n$ is not strongly convex and because the problem is ill-conditioned, one usually considers the solution to the regularized ERM (Expected Risk Minimization), for a certain $\lambda > 0$:

$$\hat{\theta}_*^{\lambda,n} \in \arg\min_{\theta} \hat{E}_n^{\lambda}(\theta),\ \hat{E}_n^{\lambda}(\theta) := \frac{1}{n} \sum_{i=1}^{n} l(z_i, \langle \theta_i, \theta \rangle) + \frac{\lambda}{2} ||\theta||_{\mathcal{H}}^2 \tag{$\hat{\mathcal{P}}_n^{\lambda}$}$$

In the case where the loss we consider is the squared loss $l((x, y), y') = \frac{1}{2}|y - y'|^2$, this is called *Kernel Ridge Regression* problem (KRR).

The statistical properties of these estimators have been studied for certain specific losses and for different notions of risk. For example:

- if we consider the least-squares problem $\inf_{\theta \in \mathcal{H}} \mathbb{E}[|Y - \theta(X)|^2]$ which corresponds to the squared loss function $l((x, y), y') = |y - y'|^2$, the main result in [3] shows that the KRR estimator is optimal if $\lambda = n^{-1/2}$ and that
$$\mathcal{R}(\hat{\theta}_*^{\lambda,n}) = O(n^{-1/2}),\ \lambda = \frac{1}{\sqrt{n}} \tag{2.1}$$

- In the case where we are looking at the regularized risk $\mathcal{R}^\lambda$ (which is a bit different from the risk we are considering in this report), the main result in [8] shows that under certain conditions on the loss function

$$\mathcal{R}^\lambda(\hat{\theta}_*^{\lambda,n}) = O\left(\frac{1}{\lambda n}\right) \tag{2.2}$$

These statistical bounds are useful not only in the study of the general problem but in the way we should compute our estimator. Indeed, they show that we can "afford" committing a certain error if we compute our estimator $\hat{\theta}$ as a an approximation of the exact solution to a certain optimization problem (as in $(\hat{\mathcal{P}}_n^\lambda)$), namely an error of same order as the risk of this exact solution.

2.2.2 Optimization and finite dimensional formulation of ERM

We will now formulate the computing of this estimator as an optimization problem in order to simplify notations. We are actually solving a problem of the form

$$\theta_*^{\lambda,n} \in \arg\min_{\theta\in\mathcal{H}} E_n^\lambda(\theta),\ E_n^\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \varphi_i(\langle\theta,\theta_i\rangle_{\mathcal{H}}) + \frac{\lambda}{2}||\theta||_{\mathcal{H}}^2 \tag{$\mathcal{P}_n^\lambda$}$$

where the φ_i are three times differentiable convex functions.

Introducing the operator $S_n : \theta \in \mathcal{H} \mapsto n^{-1/2}(\theta\cdot\theta_i)_{1\leq i\leq n} \in \mathbb{R}^n$, we see that E_n^λ is of the form $\Phi\circ S_n + \frac{\lambda}{2}||\cdot||_{\mathcal{H}}^2$ where Φ is of the form $\Phi : v\in\mathbb{R}^n \mapsto \frac{1}{n}\sum_{i=1}^n \varphi_i(\sqrt{n}v_i)$, where the φ_i are convex and three times differentiable functions.

Denote with $\mathcal{H}_n = \text{span}\{\theta_i\ :\ 1\leq i\leq n\}$ which is also the range of S_n^* : it is easy to see that if there is a solution to $(\mathcal{P}_n^\lambda)$ in $\mathcal{H}$, then there exists a unique minimizer and it is in $\mathcal{H}_n$. This reduces problem $(\mathcal{P}_n^\lambda)$ to a n-dimensional problem by looking for a solution of the form $\theta = S_n^*\alpha = \frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha_i\theta_i$. If we denote with L_{nn} the matrix $S_nS_n^* = \frac{1}{n}\left(\langle\theta_i,\theta_j\rangle\right)_{1\leq i,j\leq n}$, then we can reformulate $(\mathcal{P}_n^\lambda)$ as the following finite dimensional problem:

$$\min_{\alpha\in\mathbb{R}^n} E_n^\lambda(\alpha),\ E_n^\lambda(\alpha) = \Phi\left(L_{nn}\alpha\right) + \frac{\lambda}{2}\alpha^T L_{nn}\alpha \tag{$\mathcal{P}_n^\lambda$}$$

2.2.3 The specific case of squared loss and KRR

In the case of the squared loss $l((x,y),y') = \frac{1}{2}|y-y'|^2$, problem $(\mathcal{P}_n^\lambda)$ becomes

$$\min_{\theta\in\mathcal{H}} \frac{1}{2n}\sum_{i=1}^n |y_i - \langle\theta,\theta_i\rangle_{\mathcal{H}}|^2 + \frac{\lambda}{2}||\theta||^2 \tag{2.3}$$

and its finite-dimensional form, the Kernel Ridge Regression, is the following:

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^n \left|\frac{y_i}{\sqrt{n}} - [L_{nn}\alpha]_i\right|^2 + \frac{\lambda}{2}\alpha^T L_{nn}\alpha \tag{2.4}$$

whose solution is simply $\alpha = (L_{nn}+\lambda I)^{-1}\frac{1}{\sqrt{n}}(y_i)_{1\leq i\leq n}$.

It is clear that the exact solving of this system cannot scale up to high dimensional problem as it is in $O(n^3)$. Taking into account (2.1), we can solve afford to solve this problem with an error of $O(n^{-1/2})$, using iterative solvers for example, and obtain a time complexity of $O(n^2)$. This is still too large for problems with millions of data points. This motivated the introduction of new large-scale methods for the squared loss, described in the next section.

2.3 Large scale methods for squared loss

In this section, we present the basis of large scaled methods for the squared loss, based on dimension reduction, and then describe informally the FALKON algorithm, a large-scale method for the squared loss introduced in [5], based on the addition of iterative solvers with pre-conditioning.

In order to reduce time and memory complexities, we reduce the dimension of the space of solutions by considering a subspace $\mathcal{H}_M$ of $\mathcal{H}$ of dimension at most $M << n$. We define this subspace as the image of a mapping $S_M^* : \mathbb{R}^M \to \mathcal{H}$, seeing it as a dual map to be coherent with our notations S_n. The following example will be crucial in the developments of this report.

Example 1 (Important form for S_M^*)**.** *The following form of $S_M^* : \mathbb{R}^M \to \mathcal{H}$ will be of particular interest to us in many respects. We will formulate them as assumptions, to refer to them later on.*

- *The first interesting forms are S_M^* which satisfy*

$$\forall \alpha \in \mathbb{R}^M,\ S_M^* \alpha = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} \alpha_i \phi(\tilde{x}_j) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} \alpha_i \tilde{\theta}_j \text{ where } (\tilde{x}_j)_{1 \leq j \leq M} \in \mathcal{X}^M,\ \tilde{\theta}_j = \phi(\tilde{x}_j) \tag{M-1}$$

- *We can also consider the following stronger assumption:*

$$S_M^* \text{ satisfies (M-1) and } (\tilde{x}_j)_{1 \leq j \leq M} \text{ is a sub-family of } (x_i)_{1 \leq i \leq n} \tag{M-2}$$

When restricting to $\mathcal{H}_M$, problem (2.3) becomes

$$\min_{\theta \in \mathcal{H}_M} \frac{1}{2n} \sum_{i=1}^{n} |y_i - \langle \theta, \theta_i \rangle_{\mathcal{H}}|^2 + \frac{\lambda}{2} ||\theta||^2 \tag{2.5}$$

Define $L_{MM} = S_M S_M^*$ and $L_{nM} = S_n S_M^*$ and assume any coefficient of these matrices is computable in near-constant time (in the case of (M-1), they are simple kernel evaluations). The solution of this problem $\theta_*^{\lambda,M}$ is of the form $S_M \alpha_*^{\lambda,M}$ where $\alpha_*^{\lambda,M}$ is solution to

$$\min_{\alpha \in \mathbb{R}^M} \frac{1}{2n} \sum_{i=1}^{n} \left| \frac{y_i}{\sqrt{n}} - [L_{nM}\alpha]_i \right|^2 + \frac{\lambda}{2} \alpha^T L_{MM} \alpha \tag{2.6}$$

that is $\alpha_*^{\lambda,M} = \left(L_{nM}^T L_{nM} + \lambda L_{MM}\right)^{-1} L_{nM}^T \frac{1}{\sqrt{n}} (y_i)_{1 \leq i \leq n}$.

Solving this problem exactly is still too costly as computing $L_{nM}^T L_{nM}$ will take $O(nM^2)$ (for statistical reasons, $M \approx n^{1/2}$). In [5], Rudi et al. introduce a new algorithm which avoids this computation time by combining iterative methods and pre-conditioning.

2.3.1 Falkon for squared loss

The idea of the original FALKON algorithm is the use of iterative solvers and pre-conditioning. The principle of iterative solvers is the following : given a positive definite symmetric matrix $A \in \mathcal{S}_n^{++}$ and a vector $b \in \mathbb{R}^n$, solve $Ax = b$ not directly but by considering it as a convex optimization problem ($\min_x \frac{1}{2} x^T A x - b^T x$). One can then use an iterative method such as gradient descent with a certain rate $\tau : x_{t+1} \leftarrow x_t + \tau(b - Ax_t)$.

The interesting thing about these methods is that they only rely on *matrix-vector products*, and that therefore, if we were to apply it to $\left(L_{nM}^T L_{nM} + \lambda L_{MM}\right)$, an iteration would only cost $O(nM)$ since the matrix would never be computed explicitly.

The number of iterations needed in such algorithms, however, is controlled by the *condition number* of A, where the condition is defined as the ration between the largest and lowest eigenvalue of A:

$$\text{Cond}(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

The convergence is given usually by a bound of the type $||x_t - A^{-1}b||_A \leq \epsilon ||A^{-1}b||_A$ if $t = \Omega(\text{Cond}(A) \log \frac{1}{\epsilon})$. Thus, if we control the condition number of the matrix, we can solve the problem with certain precision ϵ in quasi-constant time.

Coming back to (2.6), we see that it is equivalent to solving the linear system $H\alpha = \Gamma$ where $H = \left(L_{nM}^T L_{nM} + \lambda L_{MM}\right)$ and $\Gamma = L_{nM}^T \frac{1}{\sqrt{n}} (y_i)_{1 \leq i \leq n}$. However, we have no control on the condition number of H which is necessary to apply an iterative solver in an effective way. That is why we *pre-condition* the matrix H.

The idea of pre-conditioning is to find a matrix B called a preconditioner so that $B^T H B$ has a small condition number. A good pre-conditioner typically satisfies $BB^T \approx \left(L_{nM}^T L_{nM} + \lambda L_{MM}\right)^{-1}$. We then compute quickly an approximation to the solution of the system $B^T H B \beta = B^T \Gamma$ (since the problem is well conditioned) and set $\alpha = B\beta$.

In [5], S_M^* is obtained under the form (M-2) by randomly sub-sampling M points from (x_i). H is preconditioned by setting $BB^T = \left(L_{MM}^T L_{MM} + \lambda L_{MM}\right)^{-1}$, which is a problem which costs $O(M^3)$ and running a conjugate gradient algorithm with a constant number of iterations t. The total complexity of the algorithm is then in order $O(nMt + M^3)$, and by taking $M = \Omega(\sqrt{n})$, optimal statistical accuracy is reached for $t = \Omega(\log n)$. Thus, FALKON is kernel method which reaches statistical optimality with complexities of $O(n\sqrt{n})$ in time and $O(n)$ in memory for the squared loss problem, making it the first large scale kernel method.

Chapter 3

FALKON for smooth losses

In this internship, our goal was to start from FALKON, which solves $(\mathcal{P})$ for the squared loss ($\inf_\theta \mathbf{E}[|Y - \theta(X)|^2]$), to create a large-scale kernel method for solving problem $(\mathcal{P})$ for more general smooth losses ($\inf_\theta \mathbf{E}[l(Z, \theta(X))]$). We would like to keep the complexities of $O(n\sqrt{n})$ in time and $O(n)$ in memory to scale well in the number of points for methods such as such as for logistic regression or robust regression.

3.1 Generalizing FALKON to smooth losses : reduction to a quadratic optimization problem

In the general case, recall that minimizing the regularized ERM and finding $\hat{\theta}_*^{\lambda,n}$, the solution of $(\hat{\mathcal{P}}_n^\lambda)$ can be seen as an instance of an optimization problem of the type

$$\theta_*^{\lambda,n} \in \arg\min_{\theta \in \mathcal{H}} E_n^\lambda(\theta),\ E_n^\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \varphi_i(\langle \theta, \theta_i \rangle_{\mathcal{H}}) + \frac{\lambda}{2}||\theta||_{\mathcal{H}}^2 \tag{$\mathcal{P}_n^\lambda$}$$

where $\varphi_i(\theta \cdot \theta_i) = l(z_i, \theta \cdot \theta_i) = l(z_i, \theta(x_i))$.

We can apply the same dimension reduction principle as in section 2.3 and perform the minimization not on $\mathcal{H}$ or equivalently $\mathcal{H}_n$ but on a subspace $\mathcal{H}_M \hookrightarrow \mathcal{H}$ given as the image of a certain linear map $S_M^* : \mathbb{R}^M \to \mathcal{H}$ (we give important examples of form of maps in section 2.3)

$$\theta_*^{\lambda,M} \in \arg\min_{\theta \in \mathcal{H}_M} E_n^\lambda(\theta) \text{ or } \arg\min_{\theta \in \mathcal{H}} \tilde{E}_M^\lambda(\theta),\ \tilde{E}_M^\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \varphi_i(\langle \theta, P_M\theta_i \rangle_{\mathcal{H}}) + \frac{\lambda}{2}||\theta||_{\mathcal{H}}^2 \tag{$\tilde{\mathcal{P}}_M^\lambda$}$$

Where P_M is the orthogonal projection on $\mathcal{H}_M$. As in the squared loss case, defining $L_{MM} = S_M S_M^*$ and $L_{nM} = S_n S_M^*$, solving problem $(\tilde{\mathcal{P}}_M^\lambda)$ is equivalent to solving

$$\min_{\alpha \in \mathbb{R}^M} \tilde{E}_M^\lambda(\alpha) = \frac{1}{n}\sum_{i=1}^n \varphi_i\left(\sqrt{n}[L_{nM}\alpha]_i\right) + \frac{\lambda}{2}\alpha^T L_{MM}\alpha \tag{$\tilde{\mathcal{P}}_M^\lambda$}$$

and taking $\theta = S_M^*\alpha$.

For now, *assuming that $\theta_*^{\lambda,M}$ will have good statistical properties* if considered as a proxy for $\theta_*^{\lambda,n}$ in a statistical setting, suppose that our aim is to compute a good approximation of $\theta_*^{\lambda,M}$. Since FALKON heavily relies on pre-conditioning, transposing it directly to the solving of $(\tilde{\mathcal{P}}_M^\lambda)$ is impossible.

On the other hand, second order methods like the newton method rely on iteratively solving quadratic approximations of the function we want to minimize. Our goal in this section is therefore to generalize FALKON to solving quadratic problems of the form

$$\arg\min_{\theta \in \mathcal{H}_M} \frac{1}{n}\sum_{i=1}^n w_i(\theta \cdot \theta_i)^2 - \frac{1}{\sqrt{n}}\sum_{i=1}^n b_i\theta \cdot \theta_i + \frac{\lambda}{2}||\theta||^2 \tag{$\mathcal{Q}_M^\lambda$}$$

because taylor expensions of $\tilde{E}_M^\lambda$ around $\theta_0 \in \mathcal{H}_M$ can be put into this form for $w_i = \varphi''(\theta_0 \cdot \theta_i)$ and $b_i = w_i\theta_0 \cdot \theta_i - \varphi'(\theta_0 \cdot \theta_i)$. The solution to this quadratic approximation is called Newton step at θ_0, and can in turn be used to compute $\theta_*^{\lambda,M}$ using a second-order method.

3.2 Extending FALKON to more general quadratic optimization problems

As explained in the previous section, the main work of this report is to generalize the FALKON algorithm to solve problems of type $(\mathcal{Q}_M^\lambda)$ ("with weights"). Hence we suppose we are given the following:

- a positive diagonal matrix $W_n = \operatorname{diag}(w_i)_{1\leq i\leq n} \in \mathbb{R}^{n\times n}$;
- a vector $b = (b_i) \in \mathbb{R}^n$

One way to see FALKON is to consider problem $(\tilde{\mathcal{P}}_M^\lambda)$ and approximately solve this problem in the "quadratic case", that is where $\varphi_i : t \in \mathbb{R} \mapsto \frac{1}{2}w_i t^2 - b_i t$. We can write this problem down in the following condensed way (we denote with $\theta_*^{\lambda,M}$ the solution to the FALKON problem in the rest of the chapter)

$$\theta_*^{\lambda,M} \in \arg\min_{\theta\in\mathcal{H}} \frac{1}{2}\langle\theta, P_M \underbrace{S_n^* W_n S_n}_{C_n} P_M\theta\rangle_{\mathcal{H}} - \langle P_M S_n^* b, \theta\rangle_{\mathcal{H}} + \frac{\lambda}{2}||\theta||_{\mathcal{H}}^2 \tag{$\mathcal{Q}_M^\lambda$}$$

Where P_M is the orthogonal projector on $\mathcal{H}_M$. Note that $C_n = \frac{1}{n}\sum_{i=1}^n w_i\theta_i \otimes \theta_i = \sum_{i=1}^n (\sqrt{w_i}\theta_i) \otimes (\sqrt{w_i}\theta_i)$. The solution of $(\mathcal{Q}_M^\lambda)$ is

$$\theta_*^{\lambda,M} = (P_M S_n^* W_n S_n P_M + \lambda I)^{-1} P_M S_n^* b \tag{3.1}$$

and can be obtained by solving the equivalent finite dimensional problem

$$\alpha_*^{\lambda,M} \in \arg\min_{\alpha\in\mathbb{R}^M} \frac{1}{2}\langle\alpha, \left(L_{nM}^T W_n L_{nM} + \lambda L_{MM}\right)\alpha\rangle_{\mathbb{R}^M} - \langle L_{Mn} b, \alpha\rangle_{\mathbb{R}^M} \tag{$\mathcal{Q}_M^\lambda$}$$

setting $\theta_*^{\lambda,M} = S_M^* \alpha_*^{\lambda,M}$.

The original FALKON solves this problem for $W_n = I_n$. In this slightly more general case, we keep the core idea of combining iterative solvers and pre-conditioning to solve $(\mathcal{Q}_M^\lambda)$.

In this context, we are solving the linear system $H\alpha = \Gamma$ where $H = \left(L_{nM}^T W_n L_{nM} + \lambda L_{MM}\right)$ and $\Gamma = L_{nM}^T b$. As in the original FALKON algorithm explained in section 2.3.1, to apply the conjugate gradient algorithm in an effective way, we need to pre-condition H.

3.2.1 The algorithm

In this section, we describe the FALKON algorithm with weights and particularly the construction of the pre-conditioner. In this section, we suppose S_M is given; the way we select those points will be discussed later. FALKON therefore simply solves the optimization system $(\mathcal{Q}_M^\lambda)$ approximately.

Recall that and ideal pre-conditioner would satisfy $BB^T = \left(L_{nM}^T W_n L_{nM} + \lambda L_{MM}\right)^{-1}$. However, computing this pre-conditioner would be, in terms of complexity, equivalent to directly computing a solution to $(\mathcal{Q}_M^\lambda)$. Instead, we compute a pre-conditioner by sub-sampling Q elements from $\{1, ..., n\}$ according to a certain probability distribution.

Throughout the report, we will adopt and discuss three different sub-sampling strategies which we can use to sample either the M points or the Q points. We describe two of them below; the third one will be developed specifically for the statistical case and will be described in C.

- **Uniform sampling** : we select indices in $\{1, ..., n\}$ without replacement;
- **Sampling according to the weights** : we select indices in $\{1, ..., n\}$ according to the probability vector $p_i = \frac{w_i}{\sum_{j=1}^n w_j} = \frac{w_i}{n\overline{w}}$ where $\overline{w} = \frac{\sum_{j=1}^n w_j}{n}$

Throughout the rest of the report, given a sub-family $J = (j_k)$ of size p in $\{1, ..., n\}$, denote with W_J the diagonal matrix of size p with coefficients $(W_J)_{kk} = w_{j_k}$. When the context is clear (i.e. the size of a family refers to this particular family), we will simply write $W_p = W_J$;

Definition 1 (Sub-sampling)**.** *Suppose we are given a probability vector on $\{1, ..., n\}$ which we denote with $(p_i)_{1\leq i\leq n}$. Let $\bar{J} = (j_k)_{1\leq k\leq Q}$ be Q iid samples from p.*

We define

- $D_Q \in \mathbb{R}^{Q\times Q}$ *the diagonal matrix with coefficients* $(D_Q)_{kk} = \sqrt{\frac{1}{np_{j_k}}},\ 1 \leq k \leq Q$*;*
- $S_Q : x \in \mathcal{H} \longmapsto \frac{1}{\sqrt{Q}} \left(x \cdot \theta_{j_k}\right)_{1\leq k \leq Q} \in \mathbb{R}^Q$*;*
- $L_{MQ} := S_M S_Q^* \in \mathbb{R}^{M\times Q}$ *and* $L_{QM} = S_Q S_M^* \in \mathbb{R}^{Q\times M}$

Note that since the Q points are sub-sampled respecting the condition (M-2), the computation of L_{QM} is easy, especially if the Kernel is easily computable.

Using these sampled vectors, we will find a pre-conditioner for H.
To do so, intuitively, we do the following approximation (in the uniform case) :

$$H \approx (S_M S_Q^* W_Q^2 S_Q S_M^* + \lambda S_M S_M^*)$$

and look for a preconditioner of the form $BB^T = (S_M S_Q^* W_Q^2 S_Q S_M^* + \lambda S_M S_M^*)^{-1}$
In the non uniform case, we add a sampling term as we can see in the following formal definition.

Operation	**Time Complexity**	**Memory Complexity**
Computing and saving T, U	M^3	M^2
Computing and saving A	$M^3 + MQ^2 + QM^2$	$M^2 + MQ$
Computing and saving $B\beta$ or $B^T\beta$ for $\beta \in \mathbb{R}^m$	M^2	M
Computing and saving $H\alpha$ for $\alpha \in \mathbb{R}^M$	nM	n
Computing $\tilde{H}\alpha$ for $\alpha \in \mathbb{R}^M$	nM	n
Total complexity with T iterations	$M^3 + MQ^2 + TnM$	$\max(n, M^2, MQ)$

Figure 3.1: Time and memory complexities in computing the FALKON estimator

Definition 2 (pre-conditioner)**.** *Let* $U = \in \mathbb{R}^{M\times m}$ *be a partial isometry such that* $U^T U = I$ *and* $T \in \mathbb{R}^{m\times m}$ *a triangular matrix such that*

$$UT^T T U^T = S_M S_M^*$$

$A \in R^{m\times m}$ *a triangular matrix such that*

$$A^T A = T^{-T} U^T \left(L_{MQ} D_Q W_Q D_Q L_{QM}\right) U T^{-1} + \lambda I_m$$

Then we define the preconditioner as follows :

$$B = UT^{-1}A^{-1} \in \mathbb{R}^{M\times m}$$

Numerically computing the pre-conditioner can be done effectively using the following QR and Cholesky decompositions

$$(U, \_) = qr(S_M S_M^*),\ T = \mathrm{chol}(U^T L_{MM} U),\ A = \mathrm{chol}(T^{-T} U^T L_{MQ} D_Q W_Q D_Q L_{QM} U T^{-1} + \lambda I_m)$$

Definition 3 (preconditioned problem)**.** *Using the preconditioner* B *defined in definition 2, define the preconditioned matrix*

$$\tilde{H} := B^T \left(L_{Mn} W_n L_{nM} + \lambda L_{MM}\right) B \in R^{m\times m}$$

Given Γ *and defining* $\tilde{\Gamma} = B^T \Gamma$*, the preconditioned problem is*

$$\beta_*^{\lambda,M,Q} \in \arg\min_{\beta} \frac{1}{2}\beta^T \tilde{H}\beta - (B^T\Gamma)^T\beta = \beta^T \tilde{H}\beta - \tilde{\Gamma}^T\beta \tag{3.2}$$

Definition 4 (FALKON computations)**.** *To compute the solution $\beta_*^{\lambda,M,Q}$ to the preconditioned problem* (3.2)*, define $\beta_*^{\lambda,M,Q,t}$ to be the t-th iteration of the conjugate gradient algorithm with matrices $\tilde{H}$ and $\tilde{\Gamma}$*

The resulting approximation of $\alpha_*^{\lambda,M}$ in problem $(\mathcal{Q}_M^\lambda)$ is $\alpha_*^{\lambda,M,Q,t} := B\beta_*^{\lambda,M,Q,t}$. We define the FALKON estimator to be the element $\mathcal{H}_M$ associated to $\alpha_*^{\lambda,M,Q,t}$ which we write $\theta_*^{\lambda,M,Q,t}$.

In order to measure the complexity of the algorithm, we measure the time complexity in number of matrix coefficient evaluations (which we suppose to be fast, which is the case when they are kernel evaluations for instance). As for the memory complexity, we measure it in the number of coefficients we need to have saved at a given time. Note that given a matrix of size (n, M) and a vector in $\mathbb{R}^M$, we can compute the matrix vector multiplication by computing stocking one block of the matrix at a time and therefore achieve a memory complexity of order n if $M < n$. We present a table of the step-by-step complexity of the FALKON algorithm in figure 3.2.1.

The algorithm itself can be found in algorithm 1.

Algorithm 1 FALKON method

```
Require: $b$, $W_n$ $x_1, ..., x_n$, $\tilde{J}$ for the $M$ points, $\lambda$, number of iterations in the conjugate gradient algorithm $t$,
  {When we write @ in front of a matrix, we mean the function associated to this matrix}
  Sample $Q$ indexes and compute $L_{MM}$ and $L_{MQ}$
  {Computing the pre-conditioner}
  $(U, \_) \leftarrow qr(L_{MM})$
  $T \leftarrow \mathrm{chol}(U^T L_{MM} U)$
  $A \leftarrow \mathrm{chol}\left(T^{-T} U^T L_{MQ} W_Q^{1/2} D_Q^2 W_Q^{1/2} L_{QM} U T^{-1} + \lambda I_m\right)$
  $@B \leftarrow \left(\beta \mapsto U T^{-1} A^{-1} \beta\right)$
  $@H, \Gamma \leftarrow \left(\alpha \mapsto @L_{Mn}\left(W_n @L_{nM}(\alpha)\right) + \lambda L_{MM}\alpha\right), @L_{Mn} b$ {Original system}
  $@\tilde{H}, \tilde{\Gamma} \leftarrow \left(\beta \mapsto @B^T(@H(@B(\beta)))\right), @B^T(\Gamma)$ {Preconditioned system}
  Perform $t$ steps of the conjugate gradient algorithm to retrieve $\beta_*^{\lambda,M,Q,t}$
  $\alpha_*^{\lambda,M,Q,t} \leftarrow @B(\beta_*^{\lambda,M,Q,t})$
  return $\alpha_*^{\lambda,M,Q,t}$ the FALKON estimator
```

Chapter 4

Theoretical results

In this chapter, we present theoretical results on the FALKON algorithm we have just derived. In section 4.1, we show that under certain conditions on Q, our algorithm is effective in solving the optimization problem $(\mathcal{Q}_M^\lambda)$. In section 4.2, we show that in our general statistical setting, in the case of a quadratic loss, the FALKON estimator can achieve statistical optimality with complexities of $O(n\sqrt{n})$ in time and $O(n)$ in memory. Finally, in section 4.3, we present the directions of future work in solving more general problems using second-order methods.

4.1 Results for FALKON with weights in the optimization setting

In this section, we consider FALKON as a solver of problem $(\mathcal{Q}_M^\lambda)$ and evaluate its performance as such (and not as that of the underlying statistical problem like in next sections).

The speed at which we solve $(\mathcal{Q}_M^\lambda)$ is determined by the condition number of $\tilde{H}$, the result of the pre-conditioning. In lemma 10, we obtain the following result :

Lemma 1 (bound on the condition number)**.** *Let Q be an integer, and suppose we sample $(\bar{x}_1, \bar{y}_1), ..., (\bar{x}_Q, \bar{y}_Q)$ from $(x_1, y_1), ..., (x_n, y_n)$ according to the probability vector $p_i := \frac{w_i}{\sum_{\tilde{i}=1}^n w_{\tilde{i}}}$, $1 \leq i \leq n$. Let $0 < \lambda \leq ||C_n||$ $\eta > 0$ and $\delta > 0$.*

Then with probability at least $1 - \delta$, if

$$Q \geq 8d\left[\overline{w}N_\infty(\lambda)\left(\left(\frac{1+\eta}{\eta}\right)^2 + \frac{1+\eta}{3\eta}\right) + \frac{1+\eta}{3\eta}\right]$$

with $d = \log \frac{8\overline{w}\kappa^2}{\lambda\delta}$ and $N_\infty(\lambda) = \sup_{1\leq i\leq n} ||C_{\lambda n}^{-1/2}\theta_i||^2 \leq \frac{\kappa^2}{\lambda}$, then

$$\mathit{Cond}(\tilde{H}) \leq 1 + \eta$$

This result shows that as soon as $Q \geq \Omega(\frac{1}{\lambda})$, the condition number of the pre-conditioned matrix $\tilde{H}$ is bounded by a constant, and therefore that our pre-conditioning is effective. This yields the following convergence result of the falkon method towards $\theta_*^{\lambda,M}$ (it is a restatement of lemma 24).

Proposition 1 (performance of the FALKON wrt $\theta_*^{\lambda,M}$)**.** *Let Q be an integer, and suppose we sample the Q points according to the probability vector $p_i := \frac{w_i}{\sum_{\tilde{i}=1}^n w_{\tilde{i}}}$, $1 \leq i \leq n$. Let $0 < \lambda \leq ||C_n||$ and $\delta > 0$.*

If $Q \geq d\,(16.4\overline{w}N_\infty(\lambda) + 3.4)$, then with probability at least $1 - \delta$,

$$\forall t \geq 0, \left[\tilde{E}_M^\lambda(\theta_*^{\lambda,M,Q,t}) - \tilde{E}_M^\lambda(\theta_*^{\lambda,M})\right]^{1/2} \leq 2e^{-t}\tilde{E}_M^\lambda\left(\theta_*^{\lambda,M}\right)^{1/2}$$

where $d = \log \frac{8\overline{w}\kappa^2}{\lambda\delta}$ and $N_\infty(\lambda) = \sup_{1\leq i\leq n} ||C_{\lambda n}^{-1/2}\theta_i||^2 \leq \frac{\kappa^2}{\lambda}$.

Again, in this proposition, we see that $\theta_*^{\lambda,M,Q,t}$ converges exponentially fast to $\theta_*^{\lambda,M}$ as soon as the condition number is bounded and therefore as soon as $Q \geq \Omega(1/\lambda)$.

Note: We can get similar results for *random sub-sampling* instead of sampling along the weights; the major change is that instead of a $\overline{w}$, a $\sup_{1\leq i\leq n} w_i$ appears in the bounds, which makes them worse.

4.2 Theoretical Guarantees of FALKON with weights in the statistical setting

In this section, we show how FALKON with weights can directly provide good estimators for quadratic machine learning problems of the form $(\mathcal{P})$, generalizing the statistical results in [5] with the same complexites of $O(n\sqrt{n})$ in time and $O(n)$ in memory.

Consider problem $(\mathcal{P})$ with losses of the type

- $l(z, y') = \frac{1}{2}w(z)|y' - r(z)|^2$
- $l(z, y') = \frac{1}{2}w(z)(y')^2 - b(z)y'$

These formulations are equivalent under the following assumptions which we will keep throughout this entire statistical part.

The main statistical problem we will consider is therefore

$$\theta_* \in \arg\min_{\theta \in \mathcal{H}} \mathbb{E}[l(Z, \langle \phi(X), \theta \rangle_{\mathcal{H}})] = \frac{1}{2}\langle \theta, \mathbb{E}[w(Z)\phi(X) \otimes \phi(X)]\theta \rangle_{\mathcal{H}} - \langle \mathbb{E}[b(Z)], b \rangle_{\mathcal{H}} \tag{$\mathcal{Q}$}$$

Hypotheses

In order for this problem to be defined/solvable, we will make the following hypothesis

- we assume that the data is bounded by a constant κ :

$$||\phi(X)||_{\mathcal{H}} \leq \kappa \text{ almost surely} \tag{S-1}$$

- Denote with w the so-called *weight function* : we assume it has the following properties :

$$w \text{ positive measurable on } \mathcal{Z} \text{ and } ||w||_\infty := \inf_{t>0} \{|w(Z)| \leq t \ a.s.\} \leq \infty \tag{S-2}$$

  Note that the first assumption makes the two losses equivalent. Moreover, this conditions implies that $||w||_1 = \mathbb{E}[w(Z)] \leq ||w||_\infty \leq \infty$.

- r is called the *target function* (indeed we wish to lessen the gap between y' and r); we will usually use one of the following hypotheses:

$$||r||^2_{L^2(\mathcal{Z}, wd\rho)} = \mathbb{E}[r(Z)^2 w(Z)] \leq \infty \Leftrightarrow ||b||^2_{L^2(\mathcal{Z}, w^{-1}d\rho)} = \mathbb{E}[b(Z)^2 w^{-1}(Z)] \leq \infty \tag{S-3}$$

$$\mathbb{E}[|r(Z)|w(Z)] \leq \infty \Leftrightarrow \mathbb{E}[|b(Z)|] \leq \infty \tag{S-3 bis}$$

  Note that (S-3) implies (S-3 bis)

- We will assume that problem $(\mathcal{Q})$ has a solution, that is

$$\exists \theta_* \in \arg\min_{\theta \in \mathcal{H}} \mathbb{E}[l(Z, \langle \phi(X), \theta \rangle_{\mathcal{H}})] \tag{S-4}$$

4.2.1 Bounds for the risk of the algorithm

Using these assumptions and the FALKON algorithm, we have the following bounds for the FALKON estimator (this is a consequence of theorem D.2).

Theorem 1. *Assume* (S-1), (S-2) *and* (S-4), *and that b is bounded or r is bounded (one can make either hypothesis). There exists constants n_0, t_0 and c_0 such that for any $n \geq n_0$ and any $\delta > 0$, if*

$$\lambda = \frac{1}{\sqrt{n}}, \ M \geq 5\left(||w||_\infty \kappa^2 + 1\right)\sqrt{n} \log \frac{44\kappa^2 ||w||_1 n}{\delta}, \ Q \geq 17\left((||w||_1 + 0.08||w||_\infty)\kappa^2 + 1\right)\sqrt{n}\log\frac{264||w||_1\kappa^2 n}{\delta}$$

Then if $t \geq \frac{1}{2}\log n + t_0$,

$$\mathcal{R}(\hat{\theta}_*^{\lambda,M,t}) \leq \frac{c_0 \log^2 \frac{6}{\delta}}{\sqrt{n}}$$

Where c_0, t_0 and n_0 do not depend on λ, M, Q, t, n and c_0 does not depend on δ, and using random sampling for M and weights sampling for Q.

4.2.2 Fast rates and Nystrom leverage scores

In this section, we present a slightly more general result and conditions with which we can obtain faster rates. Define :

- $C = \mathbb{E}[w(Z)\phi(X) \otimes \phi(X)]$ the co-variance operator, such that for any $\theta, \theta' \in \mathcal{H}$, $\mathbb{E}\left[w(Z)\theta(X)\theta'(X)\right] =: \langle \theta, \theta' \rangle_{L^2(Z,w)} = \langle \theta, C\theta' \rangle_{\mathcal{H}}$;
- $\mathcal{N}(\lambda) := \mathrm{Tr}((C + \lambda I)^{-1}C)$ the effective dimension of the problem.

In order to get faster rates, one must make two types of hypothesis :

- A source condition : there exists $s \in [0, 1/2]$ such that $C^{-s}\theta_*$ exists. s controls the regularity of the solution θ_* in the space of solutions; $s = 0$ is simply the previous case.
- A size condition of the type $\mathcal{N}(\lambda) = O(\lambda^{-\gamma})$ where $\gamma \in (0, 1]$; γ is a parameter which measures the size of the RKHS $\mathcal{H}$ with respect to the problem. $\gamma = 1$ is the generic case.

We also mention a different way of sub-sampling Q and M points which is defined in appendix C : Nystrom leverage scores, which allow us to reduce the number of necessary M points and Q points even more.

Theorem 2 (Fast rates)**.** *Assume* (S-1), (S-2) *and* (S-4)*, and that r is bounded (one can make either hypothesis). There exists constants* n_0*,* t_0 *and* c_0 *such that for any* $n \geq n_0$ *and any* $\delta > 0$*, if*

$$\lambda = n^{-\frac{1}{1+2s+\gamma}}, \ t \geq \frac{1}{2}\log n + t_0$$

Then

$$\mathcal{R}(\hat{\theta}_*^{\lambda,M,t}) \leq c_0 \left(\log^2 \frac{6}{\delta}\right) n^{-\frac{1+2s}{1+2s+\gamma}}$$

The bounds needed for M *are the following depending on the different sampling schemes, setting* $d = \log \frac{132\, Tr(C)}{\lambda\delta} \leq \log \frac{132||w||_1\kappa^2}{\lambda\delta}$*:*

- *In the random case,* $M \geq (d - \log \frac{22}{4})(4.5\mathcal{N}_w(\lambda) + 2)$
- *In the Nystrom case* $M \geq \frac{8d}{3} + 21.2dq^2\mathcal{N}(\lambda)$
- *in the weighted case,* $M \geq \frac{8d}{3} + 21.2d\overline{||w||_1}\mathcal{N}_\infty(\lambda)$

And on Q*, setting* $\tilde{d} = \log \frac{264\, Tr(C)}{\lambda\delta} \leq \log \frac{234||w||_1\kappa^2}{\lambda\delta}$ *:*

- *In the weighted case* $Q \geq \tilde{d}\left(16.4\overline{||w||_1}\mathcal{N}_\infty(\lambda) + 3.4\right)$
- *In the Nystrom case* $Q \geq \tilde{d}\left(44q^2\mathcal{N}(\lambda) + 3.4\right)$

4.3 Second-order methods

Main direction of research

In this section, we do not assume the loss is quadratic anymore, but come back to a convex three times differentiable loss function l. In order to find a good estimator $\hat{\theta}$ of the solution to problem $(\mathcal{P}) : \min_{\theta \in \mathcal{H}} \mathbb{E}[l(Z, \theta(X))]$, we would like to proceed in the following way:

- Find a statistical bound showing that for a certain $M << n$ (and a certain sub-sampling strategy), the estimator
$$\hat{\theta}_*^{\lambda,M} \in \arg\min_{\theta \in \mathcal{H}} \hat{E}_M^\lambda(\theta) := \frac{1}{n}\sum_{i=1}^n l(z_i, \langle \theta, P_M\theta_i \rangle_{\mathcal{H}}) + \frac{\lambda}{2}||\theta||^2 \qquad (\hat{\mathcal{P}}_M^\lambda)$$
has a small excess risk, i.e. find a reasonable bound on $\mathcal{R}(\hat{\theta}_*^{\lambda,M})$;
- Find an effective way of computing an approximation to this $\hat{\theta}_*^{\lambda,M}$ using second order methods and the FALKON algorithm.

For the moment, we have not yet succeeded in proving good statistical bounds for $\mathcal{R}(\hat{\theta}_*^{\lambda,M})$.

Applying FALKON to compute a newton step

In order to minimize $\hat{E}_M^\lambda$, we would like to apply second order methods to use FALKON. The typical second-order method is the Newton scheme, which works as follows:

- Start at $\theta_0 \in \mathcal{H}_M$ (usually 0)
- Given the t-step θ_t, compute the second-order approximation of $\hat{E}_M^\lambda$ around θ_t:

$$\hat{E}_M^\lambda(\theta) \approx \hat{E}_M^\lambda(\theta_t) + \nabla \hat{E}_M^\lambda(\theta_t) \cdot (\theta - \theta_t) + \frac{1}{2}\langle \theta - \theta_0, \nabla^2 \hat{E}_M^\lambda(\theta_t)(\theta - \theta_t)\rangle_{\mathcal{H}}$$

- Then set θ_{t+1} to be the minimizer of the quadratic approximation.$\Delta_t = \theta_{t+1} - \theta_t$ is the **newton step** and in closed form

$$\nabla^2 \hat{E}_M^\lambda(\theta_t)\Delta_t = -\nabla \hat{E}_M^\lambda(\theta_t) \tag{4.1}$$

Computing the Hessian and gradient of $\hat{E}_M^\lambda$, it is easy to see that equation (4.1) is of the form which FALKON can deal with as explained in 3.1. Indeed,

$$\forall \theta \in \mathcal{H},\ \nabla^2 \hat{E}_M^\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \underbrace{l''(z_i, \theta(x_i))}_{:=w_i(z)} \theta_i \otimes \theta_i + \lambda I$$

In particular, applying directly lemma 24, we have that

Proposition 2 (performance of the FALKON to compute a newton step)**.** *Suppose we are at θ_0 and let Δ be the full newton step at point θ_0, i.e. $\Delta = -\nabla^2 \hat{E}_M^\lambda(\theta_0)^{-1}\nabla \hat{E}_M^\lambda(\theta_0)$ Let Q be an integer, and suppose we sample the Q points according to the probability vector $p_i := \frac{l''(z_i,\theta(x_i))}{\sum_{i=1}^n l''(z_j,\theta(x_j))}$, $1 \leq i \leq n$. Let $0 < \lambda \leq ||\nabla^2 \hat{E}_M^\lambda(\theta)||$ and $\delta > 0$. Let $\tilde{\Delta}$ be the approximation of Δ obtained by solving the pre-conditioned system after t iterations.*

If $Q \geq d\left(16.4\frac{\sum_{i=1}^n l''(z_i,\theta(x_i))}{n}N_\infty(\lambda) + 3.4\right)$, then with probability at least $1 - \delta$,

$$||\Delta - \tilde{\Delta}||_{\nabla^2 \hat{E}_M^\lambda(\theta_0)} \leq 2e^{-t}\nu(\theta_0)$$

where $d = \log\frac{8\overline{w}\kappa^2}{\lambda\delta}$ and $N_\infty(\lambda) = \sup_{1\leq i\leq n}||C_{\lambda n}^{-1/2}\theta_i||^2 \leq \frac{\kappa^2}{\lambda}$. $\nu(\theta_0) := \langle \nabla \hat{E}_M^\lambda(\theta_0),\ \nabla^2 \hat{E}_M^\lambda(\theta_0)^{-1}\nabla \hat{E}_M^\lambda(\theta_0)\rangle_{\mathcal{H}}$ is called the newton decrement at point θ_0.

Making second order methods work: problems and possible directions

The previous proposition shows that we are able to effectively compute an approximation of a newton step of the function $\hat{E}_M^\lambda$ at any point using FALKON. However, to the best of our knowledge, it is very hard to prove the fast convergence of a Newton method. Indeed, there exists a fast regime for Newton methods, but only once the starting point θ_0 is sufficiently close to the optimal value (this region is called the Dynkin ellipsoid).

For example, if we assume that the function l satisfies the generalized self-concordance hypothesis (see [2] or [9]), that is $|l'''(z,\cdot)| \leq Cl''(z,\cdot)$, then $\hat{E}_M^\lambda$ is $C\kappa$ generalized self-concordant, that is

$$\forall \theta \in \mathcal{H},\ h, \delta \in \mathcal{H},\ \left|D^3 \hat{E}_M^\lambda(\theta)[h,\delta,\delta]\right| \leq C\kappa||h||_{\mathcal{H}}||\delta||^2_{\nabla^2 \hat{E}_M^\lambda(\theta)}$$

Typically, this is the case for the logistic loss with $C = 1$. This combined with the fact that $\hat{E}_M^\lambda$ is λ strongly convex shows that the Newton method converges quadratically in the region $\nu(\theta) \leq \mathrm{Cste}\sqrt{\lambda}$. However, outside this region, the full newton step algorithm has no guarantee of converging quickly, and even no guarantee of converging at all ! (we can modify it a bit using a learning rate, i.e. set $\theta_{t+1} = \theta_t + \tau\Delta_t$ to make it converge but no good time guaranties).

For the moment, we are therefore constrained by the size of this ellipsoid which is very small a priori. However, we aim to find conditions under which this ellipsoid can be made bigger, so that we could somehow get in it with with another less effective algorithm and then use FALKON to compute newton steps.

Chapter 5

Experiments

Here, we present FALKON's performace on two large-scale data-sets. As we are only in the beginning of our experimentation, we have not proceeded to all of the classical pre-processing for these different data-sets but rather on making a small comparison between the FALKON method and SAGA, both on the least squares problem and the logistic regression. We indicate the performance of FALKON in terms of classification error for both losses; however, our tuning of the parameters being very rough, they are not to be considered the baseline results of the method. For more complete results on least squares, see [5] which performs an in-depth experimentation of the performance of FALKON

In these experiments, we used a single TESLA P100 GPU with 16GB of RAM. We consider the two following datasets:

SUSY ($n = 5 \times 10^6$, $d = 18$, binary classification) We used a Gaussian Kernel with $\sigma = 10$, $\lambda = 10^{-6}$ for both least squares and logistic regression. When comparing to SAGA, we extracted $n = 10^5$ elements and used $M = 3000$ Nystrom points in order for SAGA to still be tractable.

HIGGS ($n = 1.1 \times 10^7$, $d = 28$, binary classification) We used a Gaussian Kernel with $\sigma = 10$, $\lambda = 10^{-4}$ for both least squares and logistic regression. We did not perform the usual pre-processing of the Higgs data-set (substracting the mean of each feature and dividing by its variance). When comparing to SAGA, we extracted $n = 10^5$ elements and used $M = 3000$ Nystrom points in order for SAGA to still be tractable.

We obtained the values of λ and σ by doing a quick cross-validation on the parameters.These parameters should be tuned more precisely.

Note on SAGA in the kernel setting

In order to perform SAGA in an efficient way, we proceed in the following way. Using the statistical properties of Nystrom points, we consider a Nystromized SAGA to reduce computations, i.e. we solve a problem of the type :

$$\min_{\alpha \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^{n} l(z_i, [L_{nM}\alpha]_i) + \frac{\lambda}{2} \alpha^T L_{MM} \alpha$$

In order to easily fall in the SAGA framework, we compute a square root $T^T T = L_{MM}$ and set β to be the new variable $\beta = T\alpha$ such that the problem is of the form

$$\min_{\beta \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^{n} l(z_i, [\tilde{L}_{nM}\beta]_i) + \frac{\lambda}{2} ||\beta||^2$$

where $\tilde{L}_{nM} = L_{nM} T^{-1}$. We do this to find a problem which is regularized in the natural norm and thus that can be solved using SAGA.

5.1 FALKON for the least squares regression

Here, we consider the least squares loss.

In figure 5.2, we compare the performance of FALKON and SAGA in terms of time. In figure 5.2, we compare the performance of FALKON and SAGA in terms of epochs, where an epoch is defined (in the FALKON case) as one evaluation in the conjugate gradient algorithm. We see that in terms of time, FALKON clearly outstrips

SAGA whereas they are comparable in terms of epochs even if FALKON seems to generalize better (an epoch in SAGA is much more expensive in terms of time).

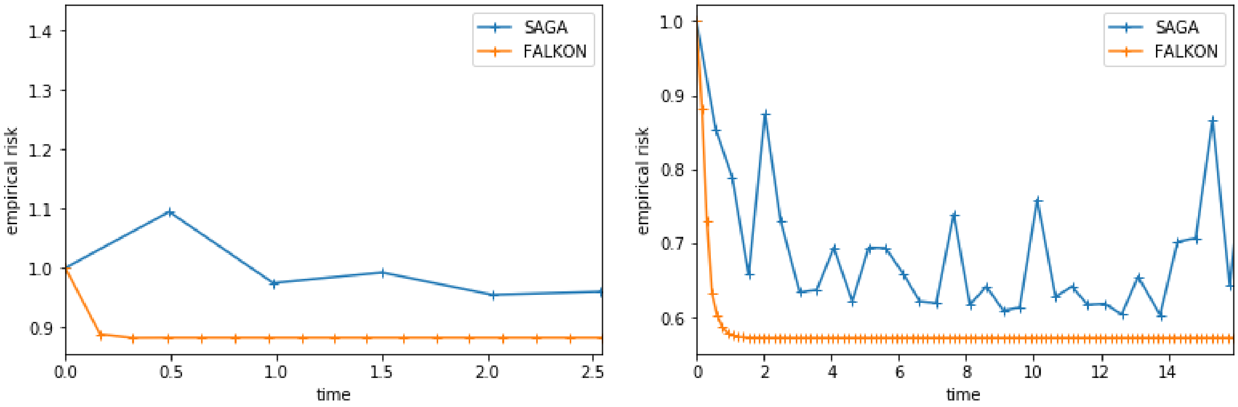

Figure 5.1: Comparison between SAGA and FALKON on the HIGGS (left) and SUSY (right) data sets for least squares regression as a function of time

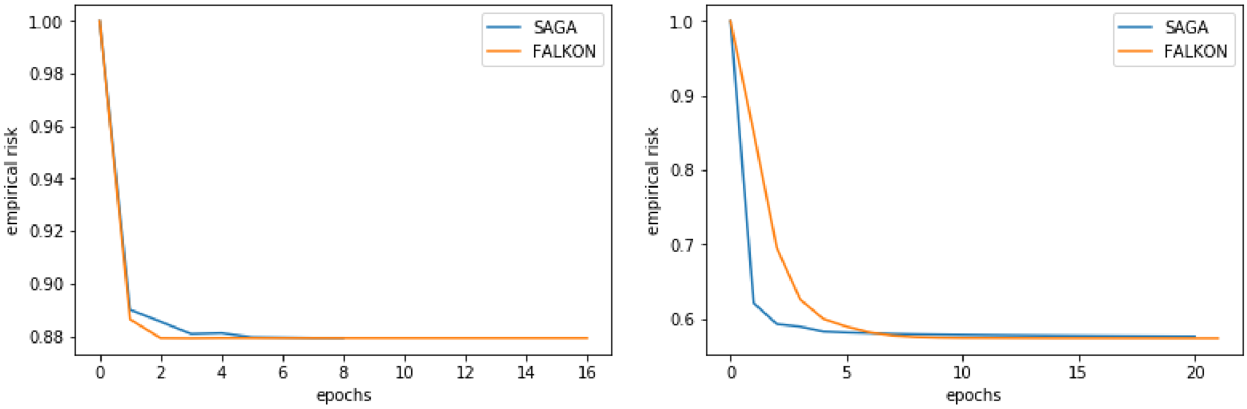

Figure 5.2: Comparison between SAGA and FALKON on the HIGGS (left) and SUSY (right) data sets for least squares regression as a function of epochs

If we use the complete power of FALKON to solve the least squares problem, we obtain the results classified in figure 5.1.

5.2 Experiments on logistic regression

To perform the FALKON method for logistic regression, we used a full newton method starting at 0. While we cannot prove that this converges to the optimum, this is a standard way of optimizing the logistic regression.

In figure 5.1, we compare the performance of FALKON and SAGA in terms of time. In figure 5.1, we compare the performance of FALKON and SAGA in terms of epochs, where an epoch is defined (in the FALKON case) as one iteration in the newton scheme. We make the same observation as previously, that is that in terms of time, FALKON clearly outstrips SAGA whereas they are comparable in terms of epochs even if FALKON seems to generalize better (an epoch in SAGA is much more expensive).

data set	number of training points	M,Q	number of test points	time (s)	classification error (%)
SUSY	4×10^6	10^4	2×10^5	300	19.905
HIGGS	10^7	10^4	2×10^5	300	34.7065

Figure 5.3: Time and classification error on a test set after training with FALKON for the least squares regression

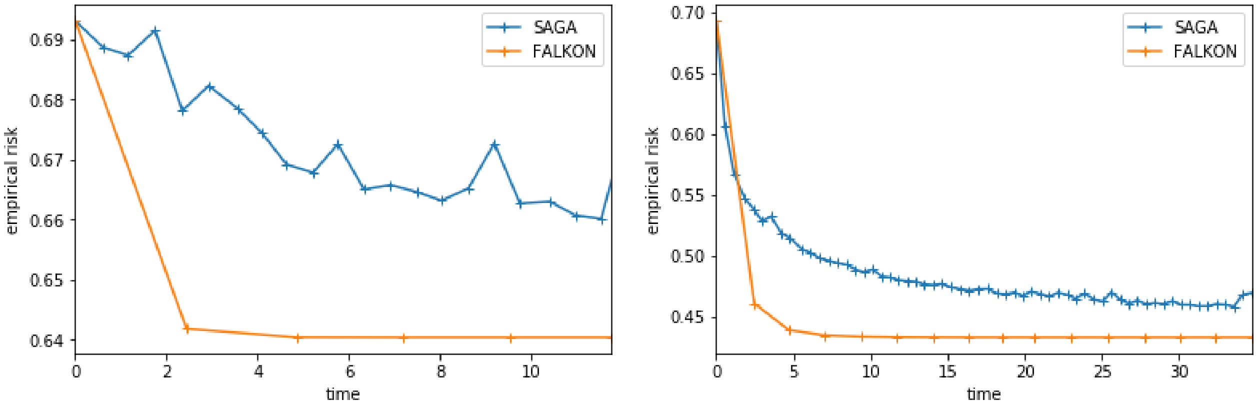

Figure 5.4: Comparison between SAGA and FALKON on the HIGGS (left) and SUSY (right) data sets for the logistic regression as a function of time

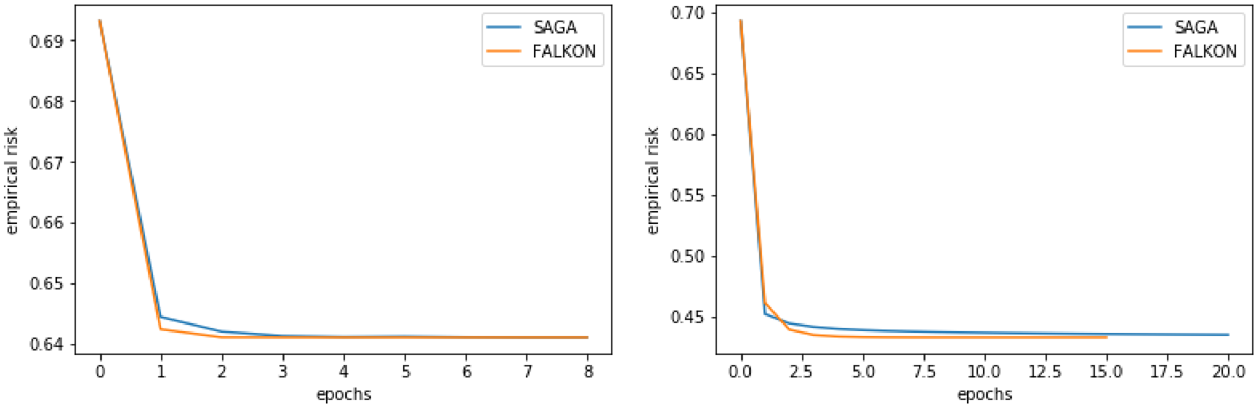

Figure 5.5: Comparison between SAGA and FALKON on the HIGGS (left) and SUSY (right) data sets for logistic regression as a function of epochs

If we use the complete power of FALKON to solve logistic regression on a significant part of the data set, we obtain results which are classified in figure 5.2.

data set	number of training points	M,Q	number of test points	time (s)	classification error (%)
SUSY	4×10^6	10^4	2×10^5	2000	19.835
HIGGS	10^7	10^4	2×10^5	3000	36.0305

Figure 5.6: Time and classification error on a test set after training for logistic regression

Conclusion

In this report, we have tried extending FALKON in order to find methods capable of dealing with large scale Kernel methods. Combining random projections and pre-conditioning, we have succeed in generalizing FALKON to solve a wider range of quadratic learning problems with complexity guarantees of $O(n\sqrt{n})$ in time and $O(n)$ in memory (up to log factors), making weighed kernel ridge regression tractable for large-scale problems.

We have also tried using FALKON in order to solve non-quadratic learning problems using second order methods. Experimentally, they seem to be promising for certain losses like the logistic loss, if we perform a Newton method.

The aim for future work is to try to get satisfactory theoretical results and rates for second order methods. This entails both finding good statistical bounds on the estimator we compute with FALKON as well as guaranteeing that this computation can be done in a limited number of iterations. In particular, a key element in this analysis seems to be the better control of the so-called Dynkin ellipsoid, the region in which second-order methods perform very well.

Bibliography

[1] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society **68** (1950), 337–404.

[2] Francis R. Bach and Eric Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n)*, CoRR **abs/1306.2119** (2013).

[3] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Found. Comput. Math. **7** (2007), no. 3, 331–368.

[4] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco, *Less is more: Nyström computational regularization*, Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (Cambridge, MA, USA), NIPS'15, MIT Press, 2015, pp. 1657–1665.

[5] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco, *FALKON: an optimal large scale kernel method*, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 3891–3901.

[6] Alessandro Rudi and Lorenzo Rosasco, *Generalization properties of learning with random features*, Advances in Neural Information Processing Systems 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), Curran Associates, Inc., 2017, pp. 3215–3225.

[7] Y. Saad, *Iterative methods for sparse linear systems*, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003.

[8] Karthik Sridharan, Shai Shalev-shwartz, and Nathan Srebro, *Fast rates for regularized objectives*, Advances in Neural Information Processing Systems 21 (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), Curran Associates, Inc., 2009, pp. 1545–1552.

[9] T. Sun and Q. Tran-Dinh, *Generalized Self-Concordant Functions: A Recipe for Newton-Type Methods*, ArXiv e-prints (2017).

Appendix A

Definitions and notations

A.1 Optimization setting

In the entire analysis, we will use the letter n to denote functions or operators attached to the family $(x_i)_{1\leq i\leq n}$, Q to denote functions or operators attached to the sub-family $(\bar{x}_j)_{1\leq j\leq Q}$ which is a sub-family of (x_i) used to compute the pre-conditioner and the letter M to denote functions or operators linked to the reduction of dimension. Recall the following definitions:

Definition 5 (Global problem). *Recall that we are given a positive diagonal matrix $W_n \in \mathbb{R}^{n\times n}$. Moreover, we define :*

- $S_n : \mathcal{H} \to \mathbb{R}^n$ *such that* $S_n(x) = \frac{1}{\sqrt{n}}\,(x\cdot\theta_i)_{1\leq i\leq n}$, *with adjoint*
- $S_n^* : \mathbb{R}^n \to \mathcal{H}$ *such that* $S_n^*\,(a_i)_{1\leq i\leq n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n a_i v_i$
- $C_n : \mathcal{H} \to \mathcal{H}$ *such that* $C_n = \frac{1}{n}\sum_{i=1}^n w_i\theta_i\otimes\theta_i = S_n^* W_n S_n$
- $C_{\lambda n} = C_n + \lambda I$

Definition 6 (reduction to $\mathcal{H}_M$). *In the whole of the report, we assume that we are given a linear map*

$$S_M : \mathcal{H} \to \mathbb{R}^M$$

and its dual map:

$$S_M^* : \mathbb{R}^M \to \mathcal{H}$$

We denote with $\mathcal{H}_M$ the range of S_M^* and P_M the orthogonal projector on $\mathcal{H}_M$.

Definition 7 (sub-sampling the Q-Points). *Recall that we sub-sample the Q points from the (x_i) according to a certain probability vector p. Denote with $(i_1, ..., i_Q)$ the sampled indices. We define*

- $W_Q = Diag\,(w_{i_k})_{1\leq k\leq Q} \in \mathbb{R}^{Q\times Q}$ *and* $D_Q = Diag\left(\sqrt{\frac{1}{np_{i_k}}}\right)_{1\leq k\leq Q}$;
- $S_Q : \theta \in \mathcal{H} \longmapsto \frac{1}{\sqrt{Q}}\,(x\cdot\theta_{j_k})_{1\leq k\leq Q} \ \in \mathbb{R}^Q$
- $G_Q = S_Q^* D_Q W_Q D_Q S_Q = \frac{1}{Q}\sum_{k=1}^Q \left(\sqrt{\frac{w_{i_k}}{np_{i_k}}}\theta_{i_k}\right)\otimes\left(\sqrt{\frac{w_{i_k}}{np_{i_k}}}\theta_{i_k}\right)$ *and* $G_{\lambda Q} = \hat{G}_Q + \lambda I$

Definition 8 (matrices). *We define*

- $L_{nn} := S_n S_n^* \in \mathbb{R}^{n\times n}$ *and* $L_{MM} := S_n S_n^* \in \mathbb{R}^{M\times M}$;
- $L_{Mn} := S_M S_n^* \in \mathbb{R}^{M\times n}$ *and* $L_{nM} = S_n S_M^* \in \mathbb{R}^{n\times M}$;
- $L_{MQ} := S_M S_Q^* \in \mathbb{R}^{M\times Q}$ *and* $L_{QM} = S_Q S_M^* \in \mathbb{R}^{Q\times M}$

Note that we always consider the computation of one of the coefficient of these matrices to be possible/in constant time.

Definition 9 (Intrinsic dimensions). *Define*

$$N_\infty(\lambda) := \sup_{1\leq i\leq n} ||C_{\lambda n}^{-1/2}\theta_i||^2$$

We have $N_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$.

A.2 Statistical setting

Recall that $L^2(Z)$ is the space of measurable functions $\psi : \mathcal{Z} \to \mathbb{R}$ such that $\mathbb{E}[|\psi(Z)|^2] < +\infty$.

Definition 10. *Under the assumptions above, for any $\theta \in \mathcal{H},\ \psi \in L^2(Z)$*

- *$S : \mathcal{H} \to L^2(Z)$ such that $S\theta : (x, y) \mapsto \langle \theta, \phi(x) \rangle$ with adjoint*
- *$S^* : L^2(Z) \to \mathcal{H}$ such that $S^*\psi = \mathbb{E}[\psi(Z)\phi(X)]$*
- *$W : L^2(Z) \to L^2(Z)$ such that $W\psi = w\psi$*
- *$C : \mathcal{H} \to \mathcal{H}$ such that $C = S^*WS$. Note that $C = \mathbb{E}[w(Z)\ \phi(X) \otimes \phi(X)]$*

We also use all the previous notations for discrete operators, but with a hat on top to signify their dependance on the data.

Let us introduce the two following quantities :

Definition 11 (effective dimension)**.** *Define*

$$\mathcal{N}(\lambda) := Tr\left(C(C + \lambda I)^{-1}\right),\ \mathcal{N}_\infty(\lambda) := \sup_{x\in\mathcal{X}} ||\phi(x)(C + \lambda I)^{-1/2}||^2,\ \mathcal{N}_w(\lambda) := \sup_{x,y} ||\sqrt{w(x, y)}\phi(x)(C + \lambda I)^{-1/2}||^2$$

We have $\mathcal{N}(\lambda) \le ||w||_1\mathcal{N}_\infty(\lambda)$, $\mathcal{N}_w(\lambda) \le ||w||_\infty\mathcal{N}_\infty(\lambda)$ and $\mathcal{N}_\infty(\lambda) \le \frac{\kappa^2}{\lambda}$. $\mathcal{N}(\lambda)$ is called the effective dimension and is somehow a measure of the complexity of the space $\mathcal{H}$ with respect to the measure $wd\rho$.

Appendix B

Analytic results

In this section, we provide purely analytic results, seeing FALKON as the solving of the optimization problem $(\mathcal{Q}_M^\lambda)$. These results will then be exploited in all settings, changing the notations depending on whether we are in the optimization setting or statistical setting.

B.1 Controlling the condition number of $\tilde{H}$

The objective of this section is to bound the condition number of the matrix $\tilde{H}$ (see definition 3) associated to the pre-conditioned linear system. We wish to do so by bounding this condition number by quantities we can more easily control in probability (see lemma 5).

Lemma 2. *Let $\lambda > 0$, $\tilde{H}$ as in definition 3, B, A as in definition 2. The matrix $\tilde{H}$ is characterized by :*

$$\tilde{H} = A^{-T} V^* (C_n + \lambda I) V A^{-1}, \text{ where } V = S_M B A$$

Moreover, V is a partial isometry such that $V^* V = I_m$ and V has the range of S_M^* which is $\mathcal{H}_M$.

Proof. First, decompose the matrix we precondition in the following way :

$$\begin{aligned} H &= (L_{Mn} W_n L_{nM} + \lambda L_{MM}) \\ &= (S_M S_n^* W_n S_n S_M^* + \lambda S_M S_M^*) \\ &= S_M (C_n + \lambda I) S_M^* \end{aligned}$$

Since by definition of V, $VA^{-1} = S_M^* B$,

$$A^{-T} V^* (C_n + \lambda I) V A^{-1} = B^T S_M (C_n + \lambda I) S_M^* B = B^T H B$$

where the last equality comes from the preliminary calculation. By definition of $\tilde{H}$, we have shown that $A^{-T} V^* (\hat{C}_n + \lambda I) V A^{-1} = \tilde{H}$.

To obtain the partial isometry, note that using the definitions of A, U, T in definition 2, we find $V = S_M^* B A = S_M^* U T^{-1}$. Then, using equation 1. in proposition **??**, we find that

$$\begin{aligned} V^* V &= T^{-T} U^T S_M S_M^* U T^{-1} \\ &= T^{-T} U^T U T^T T U^T U T^{-1} = I_m \end{aligned}$$

Finally, since $S_M S_M^* = (UT^T)(UT^T)^T$, and T is invertible, the range of U is the same as the range of S_M. Thus, the range of $S_M^* U$ is equal to the range of S_M and thus the range of $V = S_M^* U T^{-1}$ is also the range of S_M^*.

□

Lemma 3. *$\tilde{H} = I + E$ where $E = A^{-T} V^* (C_n - \hat{G}_Q) V A^{-1}$. In particular, if $||E|| < 1$, then*

$$\mathit{cond}(\tilde{H}) \leq \frac{1 + ||E||}{1 - ||E||}$$

Before proving this lemma, we will prove the following technical lemma, which shows that our preconditioner satisfies the good matrix properties (see discussion before definition 2)

Lemma 4 (technical properties of the preconditioner)**.** *The preconditioner we defined in Definition 2 satisfies the following property:*

$$B^T\left(S_M S_Q^* D_Q W_Q D_Q S_Q S_M^* + \lambda n S_M S_M^*\right) B = B^T S_M \left(\hat{G}_Q + \lambda I\right) S_M^* B = I_m \tag{B.1}$$

Proof. Using the matrices U, T, A introduced in definition 2, we have by definition of A:

$$\begin{aligned} UT^T A^T A T U^T &= UT^T\left(T^{-T} U^T S_M S_Q^* D_Q W_Q D_Q S_Q S_M^* U T^{-1} + \lambda I_m\right) T U^T \\ &= UU^T\left(S_M S_Q^* D_Q W_Q D_Q S_Q S_M^*\right) UU^T + \lambda U T^T T U^T \end{aligned}$$

Using the fact that $UU^T S_M = S_M$ (the span of S_M is the span of $S_M S_M^*$) and $UT^T TU^T = S_M S_M^*$ (see definition 2),

$$UT^T A^T A T U^T = S_M S_Q^* D_Q W_Q D_Q S_Q S_M^* + \lambda S_M S_M^*$$

Using this equality, since $B = UT^{-1}A^{-1}$ and $U^T U = I_m$ (see definition 2), equation (B.1) falls directly. □

We can now prove lemma 3, simply using the previous technical lemma:

Proof. We need to show that $A^{-T}V^*(\hat{G}_Q + \lambda I)VA^{-1} = I$. To do so, develop this using operators :

$$A^{-T}V^*(G_Q + \lambda I)VA^{-1} = B^T S_M\left(G_Q + \lambda I\right) S_M^* B$$

We then conclude using the (B.1) □

Lemma 5. *Let E be as defined in 3. Then*

$$||E|| \leq ||G_{\lambda Q}^{-1/2}\left(C_n - G_Q\right) G_{\lambda Q}^{-1/2}|| \tag{B.2}$$

Proof. First note that in the proof of lemma 3, we prove that $A^{-T}V^*(G_Q + \lambda I)VA^{-1} = I$ and thus

$$||A^{-T}V^* G_{\lambda Q}^{1/2}||^2 = ||A^{-T}V^*(G_Q + \lambda I)VA^{-1}|| = ||I|| = 1$$

Then, we multiply and divide by $G_{\lambda Q}^{-1/2}$ to obtain

$$\begin{aligned} ||E|| &= ||A^{-T}V^*(C_n - G_Q)VA^{-1}|| \\ &= ||A^{-T}V^* G_{\lambda Q}^{1/2} G_{\lambda Q}^{-1/2}(C_n - G_Q) G_{\lambda Q}^{-1/2} G_{\lambda Q}^{1/2} V A^{-1}|| \\ &\leq ||A^{-T}V^* G_{\lambda Q}^{1/2}||\ ||G_{\lambda Q}^{-1/2}(C_n - G_Q) G_{\lambda Q}^{-1/2}||\ ||G_{\lambda Q}^{1/2} V A^{-1}|| \\ &= ||A^{-T}V^* G_{\lambda Q}^{1/2}||^2\ ||G_{\lambda Q}^{-1/2}(C_n - G_Q) G_{\lambda Q}^{-1/2}|| = ||G_{\lambda Q}^{-1/2}(C_n - G_Q) G_{\lambda Q}^{-1/2}|| \end{aligned}$$

□

B.2 Representation of the FALKON estimator

Lemma 6 (Representation of the Falkon estimator in $\mathcal{H}$)**.**

$$\theta_*^{\lambda,M,Q,t} = S_M^* \alpha_*^{\lambda,M,Q,t} = S_M^* B \beta_*^{\lambda,M,Q,t}$$

Moreover, we have

$$\theta_*^{\lambda,M} = S_M^* B \beta_*^{\lambda,M,Q}$$

Proof. The fact that $\theta_*^{\lambda,M,Q,t} = S_M^* \alpha_*^{\lambda,M,Q,t} = S_M^* B \beta_*^{\lambda,M,Q,t}$ is just an immediate consequence of the definition of the FALKON estimator.

Then note that by definition, $\beta_*^{\lambda,M,Q} = \tilde{H}^{-1} B^T S_M S_n^* b$, so that

$$S_M^* B \beta_*^{\lambda,M,Q} = \left(S_M^* B \tilde{H}^{-1} B^T S_M \right) S_n^* b$$

Using the notations in Lemma 2, we see that $\tilde{H}^{-1} = A \left(V^* C_n V + \lambda I\right)^{-1} A^T$ and thus, using the fact that $V = S_M B A$,

$$S_M^* B \tilde{H}^{-1} B^T S_M = V \left(V^* C_n V + \lambda I\right)^{-1} V^*$$

and therefore

$$S_M^* B \beta_*^{\lambda,M,Q} = V \left(V^* C_n V + \lambda I\right)^{-1} V^* S_n^* b$$

On the other hand, using the fact that V is a partial isometry with range $\mathcal{H}_M$ (see lemma 2), we see that $VV^* = P_M$ where P_M is the orthogonal projection on $\mathcal{H}_M$. Starting from the characterization of $\theta_*^{\lambda,M}$ in (3.1),

$$\begin{aligned}
\theta_*^{\lambda,M} = \left(P_M S_n^* W_n S_n P_M + \lambda I\right)^{-1} P_M S_n^* b &= P_M S_n^* W_n^{1/2} \left(W_n^{1/2} S_n P_M S_n^* W_n^{1/2} + \lambda I \right)^{-1} W_n^{-1/2} b \\
&= V V^* S_n^* W_n^{1/2} \left(W_n^{1/2} S_n V V^* S_n^* W_n^{1/2} + \lambda I_n \right)^{-1} W_n^{-1/2} b \\
&= V \left(V^* S_n^* W_n S_n V + \lambda I_m \right)^{-1} V^* S_n^* b
\end{aligned}$$

which gives us the result.

□

Lemma 7. *Recall that $\beta_*^{\lambda,M,Q}$ is the solution to $\tilde{H}\beta_*^{\lambda,M,Q} = B^T S_M S_n^* b$. Then*

$$||\tilde{H}^{1/2} \beta_*^{\lambda,M,Q}||_{\mathbb{R}^m} = || \left(P_M C_n P_M + \lambda I\right)^{-1/2} P_M S_n^* b|| = ||C_{\lambda n}^{1/2} \theta_*^{\lambda,M}||$$

Proof. Using the definition of $\tilde{H}$, we can develop

$$\begin{aligned}
||\tilde{H}^{1/2} \beta_\infty||^2 &= ||\tilde{H}^{-1/2} B^T S_M S_n^* b||^2 & &= b^T S_n S_M^* B \tilde{H}^{-1} B^T S_M S_n^* b \\
&= b^T S_n V \left(V^* C_n V + \lambda I\right)^{-1} V^* S_n^* b & &= b^T S_n P_M \left(P_M C_n P_M + \lambda I\right)^{-1} P_M S_n^* b \\
&= || \left(P_M C_n P_M + \lambda I\right)^{-1/2} P_M S_n^* b||^2 & &= ||C_{\lambda n}^{1/2} \theta_*^{\lambda,M}||^2
\end{aligned}$$

□

This yields the following lemma, using theorem 6.6 of [7]

Lemma 8 (performance of the conjugate gradient method). *We can bound the performance of the conjugate gradient method in the following way:*

$$||\tilde{H}^{1/2}(\beta_t - \beta_\infty)|| \leq Q(\tilde{H}, t) || \left(P_M C_n P_M + \lambda I\right)^{-1/2} P_M S_n^* b|| = Q(\tilde{H}, t) ||C_{\lambda n}^{1/2} \theta_*^{\lambda,M}|| \tag{B.3}$$

where $Q(\tilde{H}, t) = 2 \left(1 - \frac{2}{\sqrt{cond(\tilde{H})} + 1} \right)^t$

Appendix C

Probabilistic estimates

Throughout the report, we will adopt and discuss three different sub-sampling strategies which we can use to sample either the M points (especially when we consider the statistical case and under the assumption (M-2)) or the Q points.

- **Uniform sampling** : we select indices in $\{1, ..., n\}$ without replacement;
- **Sampling according to the weights** : we select indices in $\{1, ..., n\}$ according to the probability vector $p_i = \frac{w_i}{\sum_{j=1}^n w_j} = \frac{w_i}{n\overline{w}}$ where $\overline{w} = \frac{\sum_{j=1}^n w_j}{n}$
- **Sampling using (q, λ_0, δ)-approximate leverage scores**

  This method of sampling is meaningful only if we are in the statistical setting and not in the optimization one. Let $n \in \mathbb{N}$ and $\lambda > 0$. We adapt the definitions from [4]. In this context, the exact leverage scores are defined by
  $$\forall 1 \leq i \leq n,\ l_\lambda(i) = \left(W_n^{1/2} L_{nn} W_n^{1/2} (W_n^{1/2} L_{nn} W_n^{1/2} + \lambda I)^{-1}\right)_{ii}$$
  We then define approximate leverage scores:

  Definition 12 ((q, λ_0, δ)-approximate leverage scores)**.** *Let $\delta > 0$, $\lambda_0 > 0$ and $q > 1$. A (random) sequence $(\hat{l}_\lambda(i))_{1 \leq i \leq n}$ is denoted as (q, λ_0, δ)-approximate leverage scores, when the following holds with probability at least $1 - \delta$:*
  $$\forall \lambda \geq \lambda_0,\ \forall 1 \leq i \leq n,\ \frac{1}{q} l_\lambda(i) \leq \hat{l}_\lambda(i) \leq q l_\lambda(i)$$

Thus, if we are given a sequence of approximate leverage scores $(\hat{l}_\lambda(i))_{1 \leq i \leq n}$, we can sample indices in $\{1, ..., n\}$ according to the probability vector given by $p_i = \frac{\hat{l}_\lambda(i)}{\sum_{j=1}^n \hat{l}_\lambda(j)}$.

C.1 Optimization case

C.1.1 Bounds involving Q for the condition number

In the optimization case, the only sampling we can really consider for the Q points is sampling along weights.

Lemma 9. *Let Q be an integer, and suppose we sample $(\bar{x}_1, \bar{y}_1), ..., (\bar{x}_Q, \bar{y}_Q)$ from $(x_1, y_1), ..., (x_n, y_n)$ according to the probability vector $p_i := \frac{w_i}{\sum_{\tilde{i}=1}^n w_{\tilde{i}}}$, $1 \leq i \leq n$. Let $0 < \lambda \leq ||C_n||$ and $\mu > 0$.*

Then with probability at least $1 - \mu$,

$$||G_{\lambda Q}^{-1/2} \left(C_n - G_Q\right) G_{\lambda Q}^{-1/2}|| \leq \frac{t}{1-t} \quad \textit{with} \;\; t \leq \frac{2d(1 + \overline{w} N_\infty(\lambda))}{3Q} + \sqrt{\frac{2d\overline{w} N_\infty(\lambda)}{Q}} \tag{C.1}$$

where $d = \log \frac{8\overline{w}\kappa^2}{\lambda\mu}$ and $N_\infty(\lambda)$ is defined in definition 9. Recall $N_\infty(\lambda) \leq \overline{w}\frac{\kappa^2}{\lambda}$.

Proof.

- The aim is to bound the following quantity :

$$||G_{\lambda Q}^{-1/2} \left(C_n - G_Q\right) G_{\lambda Q}^{-1/2}|| \leq ||G_{\lambda Q}^{-1/2} C_{\lambda n}^{1/2}||^2 ||C_{\lambda n}^{-1/2} \left(C_n - G_Q\right) C_{\lambda n}^{-1/2}||$$

$$\leq \left(1-\lambda_{\max}(C_{\lambda n}^{-1/2}\left(C_n-G_Q\right)C_{\lambda n}^{-1/2})\right)^{-1}||C_{\lambda n}^{-1/2}\left(C_n-G_Q\right)C_{\lambda n}^{-1/2}||$$
$$\leq \frac{t}{1-t} \text{ where } t=||C_{\lambda n}^{-1/2}\left(C_n-G_Q\right)C_{\lambda n}^{-1/2}||$$

In this sequence we have used different inequalities to precise in [6]

- Define u to be the random variable distributed as follows:

$$u=\sqrt{\frac{w_j}{p_j n}}\theta_j=\sqrt{\overline{w}}\theta_j \qquad \text{with probability } p_j,\ \forall 1\leq j\leq n$$

It is easy to see that in our case, G_Q is of the form $\frac{1}{Q}\sum_{i=1}^{Q}u_i\otimes u_i$ where the u_i are independent and identically distributed following the law of u.

Let us now check the hypothesis for proposition 6 in the paper [6]

  - By definition of u:

$$\mathbb{E}[u\otimes u]=\sum_{j=1}^{n}p_j\left(\sqrt{\frac{w_j}{p_j n}}\theta_j\right)\otimes\left(\sqrt{\frac{w_j}{p_j n}}\theta_j\right)=\frac{1}{n}\sum_{j=1}^{n}w_j\theta_j\otimes\theta_j=C_n$$

  - Let us now show the bound on the eigenvalues :

$$\langle u, C_{\lambda n}^{-1}u\rangle \leq \sup_{1\leq j\leq n}\frac{||C_{\lambda n}^{-1/2}\sqrt{w_j}\theta_j||^2}{p_j n}=\frac{\sum_{j=1}^{n}w_i}{n}\sup_{1\leq j\leq n}||C_{\lambda n}^{-1/2}\theta_j||^2$$
$$=\overline{w}N_\infty(\lambda)\leq\overline{w}\frac{\kappa^2}{\lambda}$$

Therefore we can apply the inequality in [6] to find, as long as $0<\lambda\leq||C_n||$, so that for all $\mu\geq 0$, with probability at least $1-\mu$,

$$||C_{\lambda n}^{-1/2}\left(C_n-G_Q\right)C_{\lambda n}^{-1/2}||\leq\frac{2d(1+\overline{w}N_\infty(\lambda))}{3Q}+\sqrt{\frac{2d\overline{w}N_\infty(\lambda)}{Q}}$$

where $d=\log\frac{8\overline{w}}{\lambda\mu}$.

□

Lemma 10 (bound on the condition number)**.** *Let Q be an integer, and suppose we sample $(\bar{x}_1,\bar{y}_1),...,(\bar{x}_Q,\bar{y}_Q)$ from $(x_1,y_1),...,(x_n,y_n)$ according to the probability vector $p_i:=\frac{w_i}{\sum_{i=1}^{n}w_i}$, $1\leq i\leq n$. Let $0<\lambda\leq||C_n||$ $\eta>0$ and $\delta>0$.*

Then with probability at least $1-\delta$, if

$$Q\geq 8d\left[\overline{w}N_\infty(\lambda)\left(\left(\frac{1+\eta}{\eta}\right)^2+\frac{1+\eta}{3\eta}\right)+\frac{1+\eta}{3\eta}\right]$$

with $d=\log\frac{8\overline{w}\kappa^2}{\lambda\delta}$, then

$$Cond(\tilde{H})\leq 1+\eta$$

Proof. Using the notation from lemma C.1, we see that combining lemma 5 and C.1, we find that with probability $1-\delta$, since $0<\lambda<||C_n||$, for any integer Q,

$$\text{Cond}(\tilde{H})\leq\frac{1}{1-2t},\ \text{with } t\leq\frac{2d(1+\overline{w}N_\infty(\lambda))}{3Q}+\sqrt{\frac{2d\overline{w}N_\infty(\lambda)}{Q}}$$

Solving the previous inequality, we easily find that $\sqrt{Q}$ must be greater than the solution of an order 2 polynomial, and bounding this solution, we obtain the desired inequality.

□

C.1.2 Bounds involving M to measure the precision of the Nyström approximation in the sub-sampling case

In this section and only in this section, we will always make the assumption (M-2) wich we recall below:

$$S_M^* : \alpha \mapsto \sum_{j=1}^{M} \alpha_j \theta_{i_j} \text{ where } (i_j)_{1 \leq j \leq M} \in \{1, ..., n\}^M \tag{M-2}$$

This allows us to make a probabilistic analysis of the effect of P_M and is also the most natural way to reduce dimension. Let us define the different operators associated to this sub-sampling.

Definition 13 (M-points by sub-sampling)**.** *Suppose we make the assumption* (M-2)*. Suppose the* $(i_j)_{1 \leq j \leq M}$ *are iid samples from* $\{1, ..., n\}$ *according to the probability vector* $p = (p_i)_{1 \leq i \leq n}$*. We define*

- $W_M = Diag\left(w_{i_j}\right)_{1 \leq k \leq M} \in \mathbb{R}^{M \times M}$ *and* $D_M = Diag\left(\sqrt{\frac{1}{np_{i_j}}}\right)_{1 \leq j \leq M}$*;*
- $S_M : x \in \mathcal{H} \longmapsto \frac{1}{\sqrt{M}} \left(x \cdot v_{i_j}\right)_{1 \leq j \leq M} \ \in \mathbb{R}^M$
- $G_M = S_M^* D_M W_M D_M S_M = \frac{1}{M} \sum_{j=1}^{Q} \left(\sqrt{\frac{w_{i_j}}{np_{i_j}}} \theta_{i_j}\right) \otimes \left(\sqrt{\frac{w_{i_j}}{np_{i_j}}} \theta_{i_j}\right)$ *and* $G_{\lambda M} = G_M + \lambda I$

Lemma 11 (Nystrom approximation in the empirical setting, non-uniform)**.** *Suppose we sample Nystrom points by sampling* M *points from* $\{1, ..., n\}$ *along the probability vector* $p = (p_i)_{1 \leq i \leq n}$ *where* $p_i = \frac{w_i}{\sum_{\tilde{i}} w_{\tilde{i}}} = \frac{w_i}{n\overline{w}}$*. Suppose we have* $\delta > 0$, $\eta > 0$ *and* $0 < \lambda < ||C_n||$*. If* $M \geq d\left(2\overline{w} N_\infty(\lambda) \left(\frac{1+\eta}{\eta}\right)^2 + \frac{4}{3} \frac{1+\eta}{\eta}\right)$*, then with probability at least* $1 - \delta$*,*

$$||(I - P_M) C_{\lambda n}^{1/2}||^2 \leq \lambda(1 + \eta)$$

where $d = \log \frac{4\,Tr(C_n)}{\lambda\delta} \leq \log \frac{4\overline{w}\kappa^2}{\lambda\delta}$ *In particular, if* $\eta = 2$*, if* $M \geq d\left(\frac{9}{2}\overline{w} N_\infty(\lambda) + 2\right)$*, then with probability at least* $1 - \delta$*,*

$$||(I - P_M) C_{\lambda n}^{1/2}||^2 \leq 3\lambda$$

Proof. Recall that $G_M = \frac{1}{M} \sum_{j=1}^{M} \left(\sqrt{\frac{w_j}{p_j n}} \tilde{\theta}_j\right) \otimes \left(\sqrt{\frac{w_j}{p_j n}} \tilde{\theta}_j\right) = \frac{\overline{w}}{M} \sum_{j=1}^{M} \tilde{\theta}_j \otimes \tilde{\theta}_j$ which has the same range as P_M. Using proposition 3 and 8 of [4], we find that

$$||(I - P_M) C_{\lambda n}^{1/2}||^2 \leq \frac{\lambda}{1 - \lambda_{\max}\left(C_{\lambda n}^{-1/2} \left(C_n - G_M\right) C_{\lambda n}^{-1/2}\right)}$$

Proceeding as in lemma C.1, and applying proposition 6 of [6], we find that with probability at least $1 - \delta$,

$$\lambda_{\max}\left(C_{\lambda n}^{-1/2} \left(C_n - G_M\right) C_{\lambda n}^{-1/2}\right) \leq \frac{2d}{3M} + \sqrt{\frac{2d\overline{w}N_\infty(\lambda)}{M}}, \text{ where } d = \log \frac{4\mathrm{Tr}(C_n)}{\lambda\delta} \leq \log \frac{4\overline{w}\kappa^2}{\lambda\delta}$$

□

Following the same proof, we have the following lemma for uniform sampling.

Lemma 12 (Nystrom approximation in the empirical setting,uniform)**.** *Suppose we have a subset of* M *points of* $\{1, ..., n\}$ *sampled uniformly at random with replacement and* $||C_n|| > \lambda > 0$*. Take* $\delta > 0$, $\eta > 0$*. With probability at least* $1 - \delta$*, if* $M \geq d\left(2 \sup_{1 \leq i \leq n} w_i N_\infty(\lambda) \left(\frac{1+\eta}{\eta}\right)^2 + \frac{4}{3} \frac{1+\eta}{\eta}\right)$*, then with probability at least* $1 - \delta$*,*

$$||(I - P_M) C_{\lambda n}^{1/2}||^2 \leq \lambda(1 + \eta)$$

where $d = \log \frac{4\,Tr(C_n)}{\lambda\delta} \leq \log \frac{4\overline{w}\kappa^2}{\lambda\delta}$ *In particular, if* $\eta = 2$*, if* $M \geq d\left(\frac{9}{2} \sup_{1 \leq i \leq n} w_i N_\infty(\lambda) + 2\right)$*, then with probability at least* $1 - \delta$*,*

$$||(I - P_M) C_{\lambda n}^{1/2}||^2 \leq 3\lambda$$

Proof. We can obtain a bound with the "with replacement" version. However, in practice it is much better without replacement and one must use an equivalent of Berstein thing with operators without replacement.

□

C.2 Statistical case

C.2.1 Bounds involving Q for the condition number

n the statistical case, there are two possible ways of sampling the Q points which make sense : sampling according to the weights or sampling according to leverage scores.

Sampling according to the weights

Lemma 13. *Let Q be an integer, $\delta > 0$. Suppose we sample indexes $i_1, ..., i_Q$ from $\{1, ..., n\}$ according to the re-normalized weights. When $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{36\kappa^2||w||_\infty}{\delta}$, if $\frac{19\kappa^2||w||_\infty}{n} \log \frac{3n}{2\delta} \leq \lambda \leq ||C||$, with probability at least $1-\delta$,*

$$||\hat{G}_{Q\lambda}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{G}_{Q\lambda}^{-1/2}|| \leq \frac{t}{1-t} \text{ where } t \leq \frac{2d(1 + (||w||_1 + \epsilon||w||_\infty)\mathcal{N}_\infty(\lambda))}{3Q} + \sqrt{\frac{2d\,(||w||_1 + \epsilon||w||_\infty)\mathcal{N}_\infty(\lambda)}{Q}}$$

where $d = \log \frac{132\kappa^2||w||_1}{\lambda\delta}$ and $\epsilon = \sqrt{\frac{1}{2n}\log\frac{3}{\delta}}$

Proof.

- Using techniques similar to the previous ones, for all $\mu > 0$ and $\lambda > 0$

$$||\hat{C}_{n\lambda}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{C}_{n\lambda}^{-1/2}|| \leq \frac{2d(1 + \overline{w}\mathcal{N}_\infty(\lambda))}{3Q} + \sqrt{\frac{2d\overline{w}\mathcal{N}_\infty(\lambda)}{Q}}$$

  where $d = \log \frac{4\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right)}{\mu}$

- Applying the bound from proposition 3, we find that for any $\mu > 0$, $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{12\kappa^2||w||_\infty}{\mu}$, if $\frac{19\kappa^2||w||_\infty}{n}\log\frac{n}{2\mu} \leq \lambda \leq ||C||$, then the following holds with probability at least $1-\mu$:

$$\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) \leq \frac{11\text{Tr}\,(C)}{\lambda} \leq \frac{11||w||_1\kappa^2}{\lambda} \text{ and } ||\hat{C}_{\lambda n}^{-1/2}C_\lambda^{1/2}||^2 \leq \frac{3}{2}$$

- Using the Hoeffding inequality, if $\epsilon = \sqrt{\frac{1}{2n}\log\frac{1}{\mu}}$, then with probability at least $1-\mu$, $\overline{w} \leq ||w||_1 + \epsilon||w||_\infty$

- Performing a union bound, we have for all $\mu > 0$, $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{12\kappa^2||w||_\infty}{\mu}$, if $\frac{19\kappa^2||w||_\infty}{n}\log\frac{n}{2\mu} \leq \lambda \leq ||C||$, with probability at least $1-3\mu$, for $\epsilon = \sqrt{\frac{1}{2n}\log\frac{1}{\mu}}$,

$$||\hat{C}_{n\lambda}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{C}_{n\lambda}^{-1/2}|| \leq \frac{2d(1 + (||w||_1 + \epsilon||w||_\infty)\mathcal{N}_\infty(\lambda))}{3Q} + \sqrt{\frac{2d\,(||w||_1 + \epsilon||w||_\infty)\mathcal{N}_\infty(\lambda)}{Q}}$$

  where $d = \log\frac{44\kappa^2||w||_1}{\lambda\mu}$

□

Lemma 14 (Sampling according to weights)**.** *Let Q be an integer, $\delta > 0$, $\eta > 0$. Suppose we sample indexes $i_1, ..., i_Q$ from $\{1, ..., n\}$ using $p_i = \frac{w_i}{n\overline{w}}$. When $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{36\kappa^2||w||_\infty}{\delta}$, if $\frac{19\kappa^2||w||_\infty}{n}\log\frac{3n}{2\delta} \leq \lambda \leq ||C||$, then if*

$$Q \geq 8d\,(||w||_1 + \epsilon||w||_\infty)\mathcal{N}_\infty(\lambda)\left(\frac{1+\eta}{3\eta} + \left(\frac{1+\eta}{\eta}\right)^2\right) + 8d\frac{1+\eta}{3\eta}$$

with probability at least $1-\delta$,

$$\mathit{Cond}(\tilde{H}) \leq 1 + \eta$$

where $d = \log\frac{132\,Tr(C)}{\lambda\delta} \leq \log\frac{132||w||_1\kappa^2}{\lambda\delta}$ and $\epsilon = \sqrt{\frac{1}{2n}\log\frac{3}{\delta}}$

Sampling according to leverage scores

Lemma 15. *Let Q be an integer, $\delta > 0$, $q \geq 1$, $\lambda_0 > 0$. Suppose we sample indexes $i_1, ..., i_Q$ from $\{1, ..., n\}$ using (q, λ_0, δ)-leverage scores. When $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{24\kappa^2||w||_\infty}{\delta}$, $\lambda_0 \vee \frac{19\kappa^2||w||_\infty}{n}\log\frac{n}{\delta} \leq \lambda \leq ||C||$, then with probability at least $1-\delta$,*

$$||\hat{G}_{Q\lambda}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{G}_{Q\lambda}^{-1/2}|| \leq \frac{t}{1-t} \quad \textit{with} \;\; t \leq \frac{2d(1+2.65q^2\mathcal{N}(\lambda))}{3Q} + \sqrt{\frac{5.3dq^2\mathcal{N}(\lambda)}{Q}}$$

where $d = \log\frac{88\,Tr(C)}{\lambda\delta} \leq \log\frac{88||w||_1\kappa^2}{\lambda\delta}$

Proof.

- As in a previous lemma :

$$||\hat{G}_{Q\lambda}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{G}_{Q\lambda}^{-1/2}|| \leq \frac{t}{1-t} \text{ where } t = ||\hat{C}_{n\lambda}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{C}_{n\lambda}^{-1/2}||$$

- Let us express the leverage scores using the operators we have introduced in section 3.

$$\begin{aligned}
l_\lambda(i) &= e_i^T W_n^{1/2}\hat{L}_{nn}W_n^{1/2}\left(W_n^{1/2}\hat{L}_{nn}W_n^{1/2} + \lambda n I\right)^{-1} e_i \\
&= e_i^T W_n^{1/2}\hat{S}_n\left(W_n^{1/2}\hat{S}_n\right)^*\left(W_n^{1/2}\hat{S}_n\left(W_n^{1/2}\hat{S}_n\right)^* + \lambda I\right)^{-1} e_i \\
&= e_i^T\left(W_n^{1/2}\hat{S}_n\right)\left(\left(W_n^{1/2}\hat{S}_n\right)^*\left(W_n^{1/2}\hat{S}_n\right) + \lambda I\right)^{-1}\left(W_n^{1/2}\hat{S}_n\right)^* e_i \\
&= \frac{1}{n}\langle\sqrt{w_i}\theta_i, \left(\hat{C}_n + \lambda I\right)^{-1}\sqrt{w_i}\theta_i\rangle_{\mathcal{H}} \\
&= \frac{1}{n}||\hat{C}_{\lambda n}^{-1/2}\sqrt{w_i}\theta_i||^2 = \frac{1}{n}\mathrm{Tr}\left(\left(\hat{C}_n + \lambda I\right)^{-1} w_i\theta_i \otimes \theta_i\right)
\end{aligned}$$

In particular, the last equation easily shows that $\sum_{j=1}^n l_\lambda(j) = \mathrm{Tr}\left(\left(\hat{C}_n + \lambda I\right)^{-1}\hat{C}_n\right) = \hat{\mathcal{N}}(\lambda)$

- Define v to be the random variable distributed as follows:

$$v = \frac{\sqrt{w_j}}{\sqrt{p_j n}}\theta_j \quad \text{ with probability } p_j,\ \forall 1 \leq j \leq n$$

It is easy to see that in our case, $\hat{G}_Q$ is of the form $\frac{1}{Q}\sum_{i=1}^Q v_i \otimes v_i$ where the v_i are independent and identically distributed following the law of v.

Let us now check the hypothesis for proposition 6 in the paper [6]

  - By definition of v:

$$\mathbb{E}[v \otimes v] = \sum_{j=1}^n p_j\left(\frac{\sqrt{w_j}}{\sqrt{p_j n}}\theta_j\right) \otimes \left(\frac{\sqrt{w_j}}{\sqrt{p_j n}}\theta_j\right) = \frac{1}{n}\sum_{j=1}^n w_j\theta_j \otimes \theta_j = \hat{C}_n$$

  - Let us now show the bound on the eigenvalues :

$$\begin{aligned}
\langle v, \hat{C}_{\lambda n}^{-1}v\rangle &\leq \sup_{1\leq j\leq n}\frac{||\hat{C}_{\lambda n}^{-1/2}\sqrt{w_j}K_{x_j}||^2}{p_j n} = \sup_{1\leq j\leq n}\frac{l_\lambda(j)}{p_j} = \sup_{1\leq j\leq n}\frac{l_\lambda(j)}{\hat{l}_\lambda(j)}\sum_{\tilde{j}}\hat{l}_\lambda(\tilde{j}) \\
&\leq q^2\frac{l_\lambda(j)}{l_\lambda(j)}\sum_{\tilde{j}} l_\lambda(\tilde{j}) = q^2\hat{\mathcal{N}}(\lambda)
\end{aligned}$$

Therefore we can apply proposition 8 in [4] to find, for all $0 < \lambda$, $\mu \geq 0$, with probability at least $1-\mu$,

$$||\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{C}_{\lambda n}^{-1/2}|| \leq \frac{2d(1+q^2\hat{\mathcal{N}}(\lambda))}{3Q} + \sqrt{\frac{2dq^2\hat{\mathcal{N}}(\lambda)}{Q}}$$

where $d = \log\frac{4\mathrm{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right)}{\mu}$.

- Applying the bound from proposition 3, we find that for any $\mu > 0$, $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{12\kappa^2||w||_\infty}{\mu}$, if $\frac{19\kappa^2||w||_\infty}{n} \log \frac{n}{2\mu} \leq \lambda \leq ||C||$, then the following holds with probability at least $1-\mu$:

$$\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) \leq \frac{11\text{Tr}\,(C)}{\lambda} \leq \frac{11||w||_1\kappa^2}{\lambda} \text{ and } \hat{\mathcal{N}}(\lambda) \leq 2.65\mathcal{N}(\lambda)$$

Therefore, applying a good union bound with the previous result, with probability at least $1-2\mu$,

$$||\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_Q\right)\hat{C}_{\lambda n}^{-1/2}|| \leq \frac{2d(1+2.65q^2\mathcal{N}(\lambda))}{3Q} + \sqrt{\frac{5.3dq^2\mathcal{N}(\lambda)}{Q}}$$

where $d = \log \frac{44\text{Tr}(C)}{\lambda\mu} \leq \log \frac{44||w||_1\kappa^2}{\lambda\mu}$

□

Lemma 16 (Bound on the condition number with Nystrom points)**.** *Let Q be an integer, $\delta > 0$, $q \geq 1$, $\lambda_0 > 0$, $\eta > 0$. Suppose we sample indexes $i_1, ..., i_Q$ from $\{1, ..., n\}$ using (q, λ_0, δ)-leverage scores. When $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{24\kappa^2||w||_\infty}{\delta}$, $\lambda_0 \vee \frac{19\kappa^2||w||_\infty}{n} \log \frac{n}{\delta} \leq \lambda \leq ||C||$, then if*

$$Q \geq 21.2dq^2\mathcal{N}(\lambda)\left(\frac{1+\eta}{3\eta} + \left(\frac{1+\eta}{\eta}\right)^2\right) + 8d\frac{1+\eta}{3\eta}$$

with probability at least $1-\delta$,

$$Cond(\tilde{H}) \leq 1+\eta$$

where $d = \log \frac{88\,Tr(C)}{\lambda\delta} \leq \log \frac{88||w||_1\kappa^2}{\lambda\delta}$

C.2.2 Bounds involving M to measure the precision of the Nyström approximation in the sub-sampling case

Lemma 17 (Nystrom approximation in the continuous setting, uniform)**.** *Suppose we have a subset of M points of $\{1, ..., n\}$ sampled uniformly at random amongst subsets of Mpoints of $\{1, ..., n\}$ and $\lambda > 0$. Take $\delta > 0$. With probability at least $1-\delta$, if $M \geq d\left(\frac{9}{2}\mathcal{N}_w(\lambda)\lambda + 2\right)$ then*

$$||(I-P_M)C_\lambda^{1/2}||^2 \leq 3\lambda$$

where $\mathcal{N}_w(\lambda) = \sup_{z\in\mathcal{Z}} ||C_\lambda^{-1/2}\sqrt{w(z)}\phi(x)||^2 \leq ||w||_\infty \frac{\kappa^2}{\lambda}$ and $d = \log \frac{4\,Tr(C)}{\lambda\delta} \leq \log \frac{4||w||_1\kappa^2}{\lambda\delta}$.

Proof. Recall that $\hat{C}_M = \frac{1}{M}\sum_{j=1}^M \left(\sqrt{w_j}\tilde{\theta}_j\right) \otimes \left(\sqrt{w_j}\tilde{\theta}_j\right)$ and $\hat{C}_M$ has range $\mathcal{H}_M$ as long as the w_j are stricly positive (positive weights hypothesis). Using proposition 3 of [4], we find that

$$||(I-P_M)C_\lambda^{1/2}|| \leq \lambda^{1/2}||\hat{C}_{\lambda M}^{-1/2}C_\lambda^{1/2}||$$

Then, using proposition 8 of [6], we have that

$$||(I-P_M)C_\lambda^{1/2}||^2 \leq \frac{\lambda}{1-\lambda_{\max}\left(C_\lambda^{-1/2}\left(C-\hat{C}_M\right)C_\lambda^{-1/2}\right)}$$

Let us apply proposition 6 of [6]. Define v to be the random variable $\sqrt{w(Z)}\phi(X)$. We see that with our hypotheses, $\hat{C}_M = \frac{1}{M}\sum_{j=1}^M v_j \otimes v_j$ where the v_j are i.i.d. and follow the law of v. Let us now check the hypotheses of proposition 6 in [6].

- By definition of v, we have $\mathbb{E}[v \otimes v] = C$.
- $\langle v, C_\lambda^{-1}v\rangle \leq \mathcal{N}_w(\lambda) := \sup_{z\in\mathcal{Z}} ||C_\lambda^{-1/2}\sqrt{w(z)}\phi(x)||^2 \leq ||w||_\infty \sup_{x\in\mathcal{X}} ||C_\lambda^{-1/2}\phi(x)||^2 \leq ||w||_\infty \frac{\kappa^2}{\lambda}$

We can therefore apply the proposition : with probability $1-\delta$, we have

$$\lambda_{\max}\left(C_\lambda^{-1/2}\left(C-\hat{C}_M\right)C_\lambda^{-1/2}\right) \leq \frac{2d}{3M} + \sqrt{\frac{2d\mathcal{N}_w(\lambda)}{M}}, \text{ where } d = \log\frac{4\text{Tr}(C)}{\lambda\delta} \leq \log\frac{4||w||_1\kappa^2}{\lambda\delta}$$

Then note that with the M we want, we systematically have $\lambda_{\max} < \frac{2}{3}$ which suffices. □

Lemma 18 (With Nystrom leverage scores)**.** *Let M be an integer, $\delta > 0$, $q \geq 1$, $\lambda_0 > 0$, and suppose the indices $i_1, ..., i_Q$ have been sampled using (q, λ_0, δ)-leverage scores. When $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{24\kappa^2||w||_\infty}{\delta}$, $\lambda_0 \vee \frac{19\kappa^2||w||_\infty}{n} \log \frac{n}{\delta} \leq \lambda \leq ||C||$, then if $M \geq \frac{8d}{3} + 21.2dq^2\mathcal{N}(\lambda)$, with probability at least $1 - \delta$,*

$$||(I - P_M)C_\lambda^{1/2}||^2 \leq 3\lambda$$

Proof. Recall that $\hat{G}_M = \frac{1}{M}\sum_{j=1}^M \left(\sqrt{\frac{w_j}{np_j}}\tilde{\theta}_j\right) \otimes \left(\sqrt{\frac{w_j}{np_j}}\tilde{\theta}_j\right)$ and $\hat{G}_M$ has range $\mathcal{H}_M$ as long as the w_j are stricly positive (positive weights hypothesis). Using proposition 3 of [4], we find that

$$||(I - P_M)C_\lambda^{1/2}||^2 \leq \lambda||\hat{G}_{\lambda M}^{-1/2}C_\lambda^{1/2}||^2 \leq \lambda||C_\lambda^{1/2}\hat{C}_{\lambda n}^{-1/2}||^2||\hat{G}_{\lambda M}^{-1/2}\hat{C}_{\lambda n}^{1/2}||^2$$

Then, using proposition 8 of [6], we have that

$$||(I - P_M)C_\lambda^{1/2}||^2 \leq \frac{\lambda}{\left(1 - \lambda_{\max}\left(C_\lambda^{-1/2}\left(C - \hat{C}_n\right)C_\lambda^{-1/2}\right)\right)\left(1 - \lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right)\right)}$$

Let us apply proposition 6 of [6] : with probability $1 - \tau$ and all $\lambda > 0$, we have

$$\lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right) \leq \frac{2d}{3M} + \sqrt{\frac{2dq^2\hat{\mathcal{N}}(\lambda)}{M}}, \text{ where } d = \log\frac{2\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right)}{\tau}$$

For any $\tau > 0$, $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{12\kappa^2||w||_\infty}{\tau}$, if $\frac{19\kappa^2||w||_\infty}{n} \log \frac{n}{2\tau} \leq \lambda \leq ||C||$, then the following holds with probability at least $1 - \tau$:

$$\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) \leq \frac{11\text{Tr}\,(C)}{\lambda} \leq \frac{11||w||_1\kappa^2}{\lambda},\ ||B_n|| \leq \frac{1}{3},\ \hat{\mathcal{N}}(\lambda) \leq 2.65\mathcal{N}(\lambda)$$

Taking a union bound and $\tau = \frac{\delta}{2}$, we see that for any $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{24\kappa^2||w||_\infty}{\delta}$, if $\frac{19\kappa^2||w||_\infty}{n} \log \frac{n}{\delta} \leq \lambda \leq ||C||$

$$||(I - P_M)C_\lambda^{1/2}||^2 \leq \frac{3\lambda}{2\left(1 - \lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right)\right)}$$

and

$$\lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right) \leq \frac{2d}{3M} + \sqrt{\frac{5.3dq^2\mathcal{N}(\lambda)}{M}}, \text{ where } d = \log\frac{44\text{Tr}\,(C)}{\lambda\delta}$$

The right hand side quantity is upper bounded by $\frac{1}{2}$ as soon as $M \geq \frac{8d}{3} + 21.2dq^2\mathcal{N}(\lambda)$ hence the result. □

Lemma 19 (With weights)**.** *Let M be an integer, $\delta > 0$, and suppose the indices $i_1, ..., i_M$ have been sampled using the weights. When $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{36\kappa^2||w||_\infty}{\delta}$, $\frac{19\kappa^2||w||_\infty}{n} \log \frac{3n}{2\delta} \leq \lambda \leq ||C||$, then if $M \geq \frac{8d}{3} + 21.2d\left(||w||_1 + \epsilon||w||_\infty\right)\mathcal{N}_\infty(\lambda)$, with probability at least $1 - \delta$,*

$$||(I - P_M)C_\lambda^{1/2}||^2 \leq 3\lambda$$

where $d = \log \frac{66\,Tr(C)}{\lambda\delta}$ and $\epsilon = \sqrt{\frac{1}{2n}\log\frac{3}{\delta}}$

Proof. As previously

$$||(I - P_M)C_\lambda^{1/2}||^2 \leq \frac{\lambda}{\left(1 - \lambda_{\max}\left(C_\lambda^{-1/2}\left(C - \hat{C}_n\right)C_\lambda^{-1/2}\right)\right)\left(1 - \lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right)\right)}$$

Let us apply proposition 6 of [6] : with probability $1 - \tau$ and all $\lambda > 0$, we have

$$\lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right) \leq \frac{2d}{3M} + \sqrt{\frac{2d\overline{w}\hat{\mathcal{N}}_\infty(\lambda)}{M}}, \text{ where } d = \log\frac{2\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right)}{\tau}$$

For any $\tau > 0$, $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log \frac{12\kappa^2||w||_\infty}{\tau}$, if $\frac{19\kappa^2||w||_\infty}{n} \log \frac{n}{2\tau} \leq \lambda \leq ||C||$, then the following holds with probability at least $1 - \tau$:

$$\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) \leq \frac{11\text{Tr}\,(C)}{\lambda} \leq \frac{11||w||_1\kappa^2}{\lambda},\ ||B_n|| \leq \frac{1}{3}$$

Using the Hoeffding inequality, if $\epsilon = \sqrt{\frac{1}{2n}\log\frac{1}{\tau}}$, then with probability at least $1-\tau$, $\overline{w} \leq ||w||_1 + \epsilon||w||_\infty$

Taking a union bound and $\tau = \frac{\delta}{3}$, we see that for any $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{36\kappa^2||w||_\infty}{\delta}$, if $\frac{19\kappa^2||w||_\infty}{n}\log\frac{3n}{2\delta} \leq \lambda \leq ||C||$

$$||(I-P_M)C_\lambda^{1/2}||^2 \leq \frac{3\lambda}{2\left(1-\lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right)\right)}$$

and

$$\lambda_{\max}\left(\hat{C}_{\lambda n}^{-1/2}\left(\hat{C}_n - \hat{G}_M\right)\hat{C}_{\lambda n}^{-1/2}\right) \leq \frac{2d}{3M} + \sqrt{\frac{5.3d\left(||w||_1 + \epsilon||w||_\infty\right)\mathcal{N}_\infty(\lambda)}{M}}$$

where $d = \log\frac{66\mathrm{Tr}(C)}{\lambda\delta}$ and $\epsilon = \sqrt{\frac{1}{2n}\log\frac{3}{\delta}}$

The right hand side quantity is upper bounded by $\frac{1}{2}$ as soon as $M \geq \frac{8d}{3} + 21.2\left(||w||_1 + \epsilon||w||_\infty\right)\mathcal{N}_\infty(\lambda)$ hence the result. □

C.3 Bounds for the intrinsic dimension

We first take this bound from proposition 1 in [4], easily adapted to the case with weights.

Lemma 20 (Empirical effective dimension)**.** *Recall that* $\hat{\mathcal{N}}(\lambda) = Tr\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right)$. *For any* $\delta > 0$, $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{6\kappa^2||w||_\infty}{\delta}$, *if* $\frac{19\kappa^2||w||_\infty}{n}\log\frac{n}{4\delta} \leq \lambda \leq ||C||$, *then the following holds with probability at least* $1-\delta$,

$$|\hat{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq 1.65\mathcal{N}(\lambda)$$

Remark 1. *Under the same conditions, with probability* $1-2\delta$ *:*

$$|\hat{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq 1.65\mathcal{N}(\lambda) \text{ and } ||B_n|| \leq \frac{1}{3}$$

where $B_n = C_\lambda^{-1/2}\left(C - \hat{C}_n\right)C_\lambda^{-1/2}$.

Proposition 3 (Bound on the intrinsic dimension)**.** *Recall the definition of the intrinsic dimension of a trace-class operator* T *:* $intdim\,(T) = \frac{Tr(T)}{||T||}$. *For any* $\delta > 0$, $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{12\kappa^2||w||_\infty}{\delta}$, *if* $\frac{19\kappa^2||w||_\infty}{n}\log\frac{n}{2\delta} \leq \lambda \leq ||C||$, *then the following holds with probability at least* $1-\delta$ *:*

$$intdim\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) \leq 2.65\left(1+\frac{\lambda}{||C||}\right)intdim\left(C_\lambda^{-1}C\right) \leq 5.3 intdim\left(C_\lambda^{-1}C\right)$$

and

$$intdim\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) \leq \frac{5.3\,Tr(C)}{||C||} + \frac{5.3\,Tr(C)}{\lambda} \leq \frac{11\,Tr(C)}{\lambda} \leq \frac{11||w||_1\kappa^2}{\lambda}$$

Proof. Let $\delta > 0$ Using the previous, under the conditions of the theorem, we have with probability at least $1-\delta$:

$$|\hat{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq 1.65\mathcal{N}(\lambda) \text{ and } ||B_n|| \leq \frac{1}{3}$$

where $B_n = C_\lambda^{-1/2}\left(C - \hat{C}_n\right)C_\lambda^{-1/2}$. Note that

$$\hat{C}_{\lambda n}^{-1}\hat{C}_n - C_\lambda^{-1}C = -\lambda C_\lambda^{-1/2}(I-B_n)^{-1}B_n C_\lambda^{-1/2}$$

Thus,

$$||\hat{C}_{\lambda n}^{-1}\hat{C}_n - C_\lambda^{-1}C|| \leq \lambda C_\lambda^{-1}||(I-B_n)^{-1}B_n|| \leq \frac{\lambda}{||C||}||C_\lambda^{-1}C||\frac{||B_n||}{1-||B_n||} \leq \frac{\lambda}{2||C||}||C_\lambda^{-1}C||$$

Noting that $\mathrm{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) = \frac{\hat{\mathcal{N}}(\lambda)}{||\hat{C}_{\lambda n}^{-1}\hat{C}_n||}$, applying a simple bound, we get

$$\mathrm{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) = \frac{\hat{\mathcal{N}}(\lambda)}{||\hat{C}_{\lambda n}^{-1}\hat{C}_n||} \leq \frac{2.65}{1-\frac{\lambda}{2||C||}}\frac{\mathcal{N}(\lambda)}{||C_\lambda^{-1}C||} = \frac{2.65}{1-\frac{\lambda}{2||C||}}\mathrm{intdim}\left(C_\lambda^{-1}C\right)$$

Then, using the fact that if $\lambda \leq ||C||$, $\frac{1}{1-\frac{\lambda}{2||C||}} \leq 1 + \frac{\lambda}{||C||}$ and intdim $\left(C_\lambda^{-1}C\right) \leq \frac{2\mathrm{Tr}(C)}{\lambda}$,

$$\text{intdim}\left(\hat{C}_{\lambda n}^{-1}\hat{C}_n\right) \leq \frac{5.3\mathrm{Tr}\,(C)}{||C||} + \frac{5.3\mathrm{Tr}\,(C)}{\lambda} \leq \frac{11\mathrm{Tr}\,(C)}{\lambda} \leq \frac{11||w||_1\kappa^2}{\lambda}$$

□

C.4 Bounds for the sampling error in the statistical case

In this section, we aim to bound the sampling error, that is the quantity $\mathcal{S}(\lambda, n) = \left\|C_\lambda^{-1/2}\left(\hat{S}_n^*\hat{b}_n - \hat{C}_n\theta_*\right)\right\| = \left\|C_\lambda^{-1/2}\left(\hat{S}_n^* W_n \hat{r}_n - \hat{C}_n\theta_*\right)\right\|$.

We will consider two different technical hypotheses depending on the point of view and precision we need. Assume that hypothesis (S-4) is satisfied.

$$\exists\sigma > 0,\ L > 0,\ \forall p \geq 2,\ \mathbb{E}\left[|b(Z) - w(Z)\langle\phi(X), \theta_*\rangle_{\mathcal{H}}|^p\right] \leq \frac{1}{2}p!\sigma^2 L^{p-2} \tag{S-5}$$

$$\exists\sigma > 0,\ L > 0,\ \forall p \geq 2,\ \forall z \in \mathcal{Z},\ \left|r(z) - \langle\theta_*, \phi(x)\rangle_{\mathcal{H}}\right|^p w(z) \leq \frac{1}{2}p!\sigma^2 L^{p-2} \tag{S-5b}$$

We now state two lemmas whose proof is straightforward using proposition 11 in [4].

Lemma 21. *Assume* (S-5). *Then for any $\tau > 0$, with probability* $1 - \tau$,

$$\mathcal{S}(\lambda, n) \leq \frac{2\sqrt{\mathcal{N}_\infty(\lambda)}L\log\frac{2}{\tau}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}_\infty(\lambda)\log\frac{2}{\tau}}{n}}$$

Lemma 22. *Assume* (S-5b). *Then for any $\tau > 0$, with probability* $1 - \tau$,

$$\mathcal{S}(\lambda, n) \leq \frac{2\sqrt{\mathcal{N}_\infty(\lambda)}L\log\frac{2}{\tau}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)\log\frac{2}{\tau}}{n}}$$

Remark 2. *We have the two following particular cases.*

- *If $||b||_\infty < +\infty$, if we assume* (S-1), (S-2) *and* (S-4), *then* (S-5) *and* (S-3 bis) *are satisfied, with $\sigma = L = (||b||_\infty + \kappa||w||_\infty||\theta_*||_{\mathcal{H}})$;*
- *if $||r||_\infty < \infty$, if we assume* (S-1), (S-2) *and* (S-4), *then* (S-5b) *and* (S-3) *are satisfied with $\sigma = (||r||_\infty + \kappa||\theta_*||_{\mathcal{H}})\,||w||_\infty^{1/2}$ and $L = (||r||_\infty + \kappa||\theta_*||_{\mathcal{H}})$.*

Lemma 23. *Assume* (S-1), (S-2) *and* (S-4).

- *If* (S-5) *is satisfied, then with probability at least* $1 - \tau$,

$$\left\|\hat{C}_{\lambda n}^{1/2}\hat{\theta}_*^{\lambda,M}\right\| \leq \lambda^{-1/2}\kappa\left(\sqrt{\frac{2\sigma^2\log\frac{1}{\tau}}{n}} + \frac{L\log\frac{1}{\tau}}{n} + \sigma + \kappa||\theta_*||_{\mathcal{H}}||w||_1\right) \tag{C.2}$$

- *If* (S-5b) *is satisfied, then with probability at least* $1 - \tau$,

$$\left\|\hat{C}_{\lambda n}^{1/2}\hat{\theta}_*^{\lambda,M}\right\| \leq \sqrt{\sigma^2 + \epsilon_1} + \sqrt{||w||_1 + \epsilon_2||w||_\infty}\kappa||\theta_*|| \tag{C.3}$$

where $\epsilon_1 = \frac{4L^2||w||_\infty\log\frac{2}{\tau}}{n} + \sqrt{\frac{2(4L\sigma)^2||w||_1\log\frac{2}{\tau}}{n}}$ *and* $\epsilon_2 = \sqrt{\frac{\log\frac{2}{\tau}}{2n}}$

Proof. Recall that

$$\left\|\hat{C}_{\lambda n}^{1/2}\hat{\theta}_*^{\lambda,M}\right\| = \left\|\left(P_M\hat{C}_nP_M + \lambda I\right)^{-1/2}P_M\hat{S}_n^*\hat{b}_n\right\| = \left\|\left(P_M\hat{C}_nP_M + \lambda I\right)^{-1/2}P_M\hat{S}_n^*W_n\hat{r}_n\right\|$$

- Assume (S-5). Using the first part of the equality, we easily get :

$$\left\|\hat{C}_{\lambda n}^{1/2}\hat{\theta}_*^{\lambda,M}\right\| = \left\|\left(P_M\hat{C}_nP_M + \lambda I\right)^{-1/2}P_M\hat{S}_n^*\hat{b}_n\right\|$$

$$\begin{aligned}
&\le \lambda^{-1/2} \left\| \frac{1}{n} \sum_{i=1}^{n} b_i \phi(x_i) \right\| \\
&\le \lambda^{-1/2} \kappa \left(\frac{1}{n} \sum_{i=1}^{n} |b_i - w_i \phi(x_i) \cdot \theta_*| + \frac{1}{n} \sum_{i=1}^{n} |w_i \phi(x_i) \cdot \theta_*| \right) \\
&\le \lambda^{-1/2} \kappa \left(\frac{1}{n} \sum_{i=1}^{n} |b_i - w_i \phi(x_i) \cdot \theta_*| + \kappa ||\theta_*||_{\mathcal{H}} ||w||_1 \right)
\end{aligned}$$

Thus, using a basic Bernstein inequality, we find that with probability at least $1 - \tau$,

$$\left\| \hat{C}_{\lambda n}^{1/2} \hat{\theta}_*^{\lambda, M} \right\| \le \lambda^{-1/2} \kappa \left(\sqrt{\frac{2\sigma^2 \log \frac{1}{\tau}}{n}} + \frac{L \log \frac{1}{\tau}}{n} + \sigma + \kappa ||\theta_*||_{\mathcal{H}} ||w||_1 \right)$$

- Assume (S-5b). Using the second part of the equality, we get

$$\left\| \hat{C}_{\lambda n}^{1/2} \hat{\theta}_*^{\lambda, M} \right\| \le \underbrace{\left\| \left(P_M \hat{C}_n P_M + \lambda I \right)^{-1/2} P_M \hat{S}_n^* W_n^{1/2} \right\|} \le 1 \left\| W_n^{1/2} \hat{r}_n \right\|$$

Let us now bound the term $\left\| W_n^{1/2} \hat{r}_n \right\|$. Separating it in two, we get

$$\left\| W_n^{1/2} \hat{r}_n \right\| = \sqrt{\frac{\sum_{i=1}^{n} w_i |r_i|^2}{n}} \le \sqrt{\frac{\sum_{i=1}^{n} w_i |r_i - \theta_* \cdot \phi(x_i)|^2}{n}} + \sqrt{\frac{\sum_{i=1}^{n} w_i |\theta_* \cdot \phi(x_i)|^2}{n}}$$

The second term is easily bounded by $\sqrt{\overline{w}} \kappa ||\theta_*||_{\mathcal{H}}$. To bound the first term, we set $\zeta = w(Z) \left| r(Z) - \theta_* \cdot \phi(X) \right|^2$ and apply a Bernstein inequality. Indeed, let $s := \mathbb{E}[\zeta]$. Using (S-5b), we get that for all $p \ge 2$,

$$\begin{aligned}
\mathbb{E}[|\zeta|^p] &\le \frac{1}{2} (2p)! \sigma^2 L^{2p-2} ||w||_\infty^{p-2} ||w||_1 \\
&\le \frac{1}{2} p! 4^2 \sigma^2 L^2 ||w||_1 (4L^2 ||w||_\infty)^{p-2}
\end{aligned}$$

Setting $\tilde{L} = 4L^2 ||w||_\infty$ and $\tilde{\sigma} = 4\sigma L ||w||_1$, a $\forall p \ge 2,\ \mathbb{E}[|\zeta|^p] \le \frac{1}{2} p! \tilde{\sigma}^2 \tilde{L}^{p-2}$. Thus, with probability at least $1 - \tau$,

$$\frac{1}{n} \sum_{i=1}^{n} \zeta_i - \mathbb{E}[\zeta] \le \frac{\tilde{L} \log \frac{1}{\tau}}{n} + \sqrt{\frac{2\tilde{\sigma} \log \frac{1}{\tau}}{n}}$$

Finally, since $\mathbb{E}[\zeta] \le \sigma^2$, we get our final bound, with probability at least $1 - \tau$:

$$\frac{1}{n} \sum_{i=1}^{n} w_i |r_i - \theta_* \cdot \phi(x_i)|^2 \le \sigma^2 + \frac{\tilde{L} \log \frac{1}{\tau}}{n} + \sqrt{\frac{2\tilde{\sigma} \log \frac{1}{\tau}}{n}}$$

□

Appendix D

Final bounds for FALKON with weights

D.1 Optimization setting

Since FALKON approximates $\theta_*^{\lambda,M}$ i.e. the solution to problem $(\mathcal{Q}_M^\lambda)$ using a conjugate gradient algorithm, our first step is to estimate the distance between our FALKON estimator $\theta_*^{\lambda,M,Q,t}$ and this quantity, which is the purpose of the following lemma:

Lemma 24 (performance of the FALKON wrt $\theta_*^{\lambda,M}$)**.** *Let Q be an integer, and suppose we sample the Q points according to the probability vector $p_i := \frac{w_i}{\sum_{i=1}^n w_i}$, $1 \leq i \leq n$. Let $0 < \lambda \leq ||C_n||$ and $\delta > 0$.*

For all $\nu > 0$, if $Q \geq 8d \cosh^2\left(\frac{\nu}{2}\right)\left[\left(\frac{1}{3} + \cosh^2\left(\frac{\nu}{2}\right)\right)\overline{w}N_\infty(\lambda) + \frac{1}{3}\right]$, then, with probability at least $1-\delta$,

$$\forall t \geq 0, ||C_{\lambda n}^{1/2}(\theta_*^{\lambda,M,Q,t} - \theta_*^{\lambda,M})|| \leq 2e^{-\nu t}||C_{\lambda n}^{1/2}\theta_*^{\lambda,M}||$$

where $d = \log\frac{8\overline{w}\kappa^2}{\lambda\delta}$.

In particular, if $Q \geq d\,(16.4\overline{w}N_\infty(\lambda) + 3.4)$, then with probability at least $1-\delta$,

$$\forall t \geq 0, ||C_{\lambda n}^{1/2}(\theta_*^{\lambda,M,Q,t} - \theta_*^{\lambda,M})|| \leq 2e^{-t}||C_{\lambda n}^{1/2}\theta_*^{\lambda,M}||$$

which is equivalent to saying

$$\forall t \geq 0, \tilde{E}_M^\lambda(\theta_*^{\lambda,M,Q,t}) - \tilde{E}_M^\lambda(\theta_*^{\lambda,M}) \leq 4e^{-2t}\tilde{E}_M^\lambda(\theta_*^{\lambda,M})$$

Proof.

- By lemma 2, $\tilde{H} = B^T S_M\,(C_n + \lambda I)\,S_M^* B$. Using lemma 6,

$$||\tilde{H}^{1/2}(\beta_*^{\lambda,M,Q,t} - \beta_*^{\lambda,M,Q})||^2 = ||C_{\lambda n}^{1/2}(\theta_*^{\lambda,M,Q,t} - \theta_*^{\lambda,M})||^2$$

  Since $||\tilde{H}^{1/2}(\beta_*^{\lambda,M,Q,t} - \beta_*^{\lambda,M,Q})||^2 \leq Q(\tilde{H},t)||\tilde{H}^{1/2}\beta_*^{\lambda,M,Q}||$, combining the previous calculation with lemma 8, we have

$$||C_{\lambda n}^{1/2}(\theta_*^{\lambda,M,Q,t} - \theta_*^{\lambda,M})||^2 \leq Q(\tilde{H},t)||C_{\lambda n}^{1/2}\theta_*^{\lambda,M}||$$

- Bounding the term $Q(\tilde{H},t)$ is equivalent to bounding the condition number with the right constant, and then use lemma 10.

□

Note: If we use random sub-sampling instead of weights, we only need to replace $\overline{w}$ by $\sup_{1\leq i\leq n} w_i$.

D.2 Statistical setting

Recall that we always make the (S-4) assumption on the existence of a solution to problem $(\mathcal{Q})$. In this section, we study the problem under the more general hypotheses (the so-called source condition)

$$\exists s \in [0,1/2],\ C^{-s}\theta_* \text{ exists and } \exists R \geq 1,\ ||C^{-s}\theta_*|| \leq R \tag{S-4b}$$

Let us now decompose the risk between $\hat{\theta}_*^{\lambda,M}$ and θ_*, as is done in Theorem 2 of [4].

Lemma 25. *Let s be as in assumption* (S-4b) *and $\hat{\theta}_*^{\lambda,M}$ be the Nystrom estimator. Then for any $\lambda, M > 0$, we have the following bound on the objective:*

$$|\mathcal{E}(\hat{\theta}_*^{\lambda,M}) - \mathcal{E}(\theta_*)|^{1/2} \leq k_2^2 \mathcal{S}(\lambda, n) + R\left[(1 + k_1 k_2)\mathcal{C}(M)^{1/2+s} + k_2^2 \lambda^{1/2+s}\right] \tag{D.1}$$

where $k_1 = ||\hat{C}_{\lambda n}^{1/2} C_\lambda^{-1/2}||$, $k_2 = ||C_\lambda^{1/2} \hat{C}_{\lambda n}^{-1/2}||$, $\mathcal{S}(\lambda, n) = ||C_\lambda^{-1/2} \hat{S}_n^*(\hat{b}_n - W_n \hat{S}_n \theta_*)||$ and $\mathcal{C}(M) = ||C_\lambda^{1/2}(I - P_M)||^2$.

Proof.

$$\begin{aligned} |\mathcal{E}(\hat{\theta}_*^{\lambda,M}) - \mathcal{E}(\theta_*)|^{1/2} &= ||C^{1/2}\left(\hat{\theta}_*^{\lambda,M} - \theta_*\right)|| \\ &\leq \underbrace{||C^{1/2} g_{\lambda,M}(\hat{C}_n)\hat{S}_n^*(\hat{b}_n - W_n \hat{S}_n \theta_*)||}_{A} + \underbrace{||C^{1/2}\left(I - g_{\lambda,M}(\hat{C}_n)\hat{C}_n\right)\theta_*||}_{B} \end{aligned}$$

where $g_{\lambda,M}(\hat{C}_n) = V\left(V^*\hat{C}_n V + \lambda I\right)^{-1} V^*$

- Let us first bound term A.

$$A \leq ||C^{1/2}\hat{C}_{\lambda n}^{-1/2}||\ ||\hat{C}_{\lambda n}^{1/2} g_{\lambda,M}(\hat{C}_n)\hat{C}_{\lambda n}^{1/2}||\ ||\hat{C}_{\lambda n}^{-1/2} C_\lambda^{1/2}||\ ||C_\lambda^{-1/2}\hat{S}_n^*(\hat{b}_n - W_n \hat{S}_n \theta_*)||$$

  And thus, using techniques from [4]

$$A \leq k_2^2 \mathcal{S}(\lambda, n)$$

- Dealing with term B is done exactly as in the proof of Theorem 2 in [4].

□

Theorem 3. *Let $\delta > 0$. Assume* (S-1),(S-2),(S-4b). *Assume that n, λ satisfy $n \geq 405\kappa^2||w||_\infty \vee 67\kappa^2||w||_\infty \log\frac{72||w||_\infty \kappa^2}{\delta} \vee 100\max(1, ||w||_\infty)\log\frac{6}{\delta}$, $\frac{19\kappa^2||w||_\infty}{n}\log\frac{3n}{\delta} \vee \lambda_0 \leq \lambda \leq ||C||$. With probability at least $1 - \delta$,*

$$\forall t \geq t_{\min},\ \mathcal{R}\left(\hat{\theta}_*^{\lambda,M,t}\right)^{1/2} \leq \frac{4}{3}\mathcal{S}(\lambda, n) + 10R\lambda^{1/2+s} \tag{D.2}$$

Where, assuming (S-5),

$$t_{\min} = \log\frac{\kappa\,(1.2\sigma + 0.01L + \kappa||\theta_*||_{\mathcal{H}}||w||_1)}{8R\lambda^{1+s}} \text{ and } \mathcal{S}(\lambda, n) = \frac{2\sqrt{\mathcal{N}_\infty(\lambda)}L\log\frac{6}{\delta}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}_\infty(\lambda)\log\frac{6}{\delta}}{n}}$$

and assuming (S-5b),

$$t_{\min} = \log\frac{\sigma + 0.3L + \sqrt{\overline{||w||_1}}\kappa||\theta_*||_{\mathcal{H}}}{8R\lambda^{1/2+s}} \text{ and } \mathcal{S}(\lambda, n) = \frac{2\sqrt{\mathcal{N}_\infty(\lambda)}L\log\frac{6}{\delta}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)\log\frac{6}{\delta}}{n}}$$

The bounds needed for M are the following depending on the different sampling schemes, setting $d = \log\frac{132\,Tr(C)}{\lambda\delta} \leq \log\frac{132||w||_1\kappa^2}{\lambda\delta}$:

- *In the random case, $M \geq (d - \log\frac{22}{4})(4.5\mathcal{N}_w(\lambda) + 2)$*
- *In the Nystrom case $M \geq \frac{8d}{3} + 21.2dq^2\mathcal{N}(\lambda)$*
- *in the weighted case, $M \geq \frac{8d}{3} + 21.2d\overline{||w||_1}\mathcal{N}_\infty(\lambda)$*

And on Q, setting $\tilde{d} = \log\frac{264\,Tr(C)}{\lambda\delta} \leq \log\frac{234||w||_1\kappa^2}{\lambda\delta}$:

- *In the weighted case $Q \geq \tilde{d}\left(16.4\overline{||w||_1}\mathcal{N}_\infty(\lambda) + 3.4\right)$*
- *In the Nystrom case $Q \geq \tilde{d}\left(44q^2\mathcal{N}(\lambda) + 3.4\right)$*

where $\overline{||w||_1} = ||w||_1 + 0.08||w||_\infty$.

Proof. First note that

$$\mathcal{R}\left(\hat{\theta}_*^{\lambda,M,t}\right)^{1/2} \leq \left\|C^{1/2}\left(\hat{\theta}_*^{\lambda,M} - \hat{\theta}_*^{\lambda,M,t}\right)\right\| + \left\|C^{1/2}\left(\hat{\theta}_*^{\lambda,M} - \theta_*\right)\right\|$$

Using the definitions in the previous lemma, we have

$$\left\|C^{1/2}\left(\hat{\theta}_*^{\lambda,M} - \theta_*\right)\right\| \leq k_2^2\mathcal{S}(\lambda,n) + R\left[(1+k_1k_2)\mathcal{C}(M)^{1/2+s} + k_2^2\lambda^{1/2+s}\right]$$

and

$$\left\|C^{1/2}\left(\hat{\theta}_*^{\lambda,M} - \hat{\theta}_*^{\lambda,M,t}\right)\right\| \leq 2k_2Q(\tilde{H},t)\left\|\hat{C}_{\lambda n}^{1/2}\hat{\theta}_*^{\lambda,M}\right\|$$

Then we combine the following elements:

- we bound k_1 and k_2 using section C.3, which is essentially asking that n be greater than a certain constant;
- we bound $\mathcal{S}(\lambda,n)$ using the bounds we obtain in section C.4;
- we bound $\mathcal{C}(M)$ by 3λ under the conditions given in section C.2.2, depending on the sub-sampling strategy;

These three elements allow us to bound $\left\|C^{1/2}\left(\hat{\theta}_*^{\lambda,M} - \theta_*\right)\right\| = \mathcal{R}(\hat{\theta}_*^{\lambda,M})^{1/2}$ by

$$\frac{4}{3}\mathcal{S}(\lambda,n) + 8.5R\lambda^{1/2+s}$$

To bound $\left\|C^{1/2}\left(\hat{\theta}_*^{\lambda,M} - \hat{\theta}_*^{\lambda,M,t}\right)\right\|$, we use

- once again the bound on k_2;
- the bound on $\left\|\hat{C}_{\lambda n}^{1/2}\hat{\theta}_*^{\lambda,M}\right\|$ provided in section C.4
- the bound on $Q(\tilde{H},t)$ given in section C.2.1 depending on the sampling strategy for Q.

□

We can now apply this result to obtain the generalization bounds, only supposing that θ_* exists.