Mathematical Programming
https://doi.org/10.1007/s10107-024-02081-4

FULL LENGTH PAPER

Series A

Finding global minima via kernel approximations

Alessandro Rudi[1] · **Ulysse Marteau-Ferey**[1] · **Francis Bach**[1]

Received: 30 December 2020 / Accepted: 7 March 2024

Abstract

We consider the global minimization of smooth functions based solely on function evaluations. Algorithms that achieve the optimal number of function evaluations for a given precision level typically rely on explicitly constructing an approximation of the function which is then minimized with algorithms that have exponential running-time complexity. In this paper, we consider an approach that jointly models the function to approximate and finds a global minimum. This is done by using infinite sums of square smooth functions and has strong links with polynomial sum-of-squares hierarchies. Leveraging recent representation properties of reproducing kernel Hilbert spaces, the infinite-dimensional optimization problem can be solved by subsampling in time polynomial in the number of function evaluations, and with theoretical guarantees on the obtained minimum. Given n samples, the computational cost is $O(n^{3.5})$ in time, $O(n^2)$ in space, and we achieve a convergence rate to the global optimum that is $O(n^{-m/d+1/2+3/d})$ where m is the degree of differentiability of the function and d the number of dimensions. The rate is nearly optimal in the case of Sobolev functions and more generally makes the proposed method particularly suitable for functions with many derivatives. Indeed, when m is in the order of d, the convergence rate to the global optimum does not suffer from the curse of dimensionality, which affects only the worst-case constants (that we track explicitly through the paper).

Keywords Global optimization · Polynomial optimization · Sum of squares · Semidefinite programming

Mathematics Subject Classification 90C26 · 47B32

✉ Alessandro Rudi
alessandro.rudi@inria.fr

Ulysse Marteau-Ferey
ulysse.marteau-ferey@inria.fr

Francis Bach
francis.bach@inria.fr

[1] INRIA - Département d'Informatique de l'École Normale Supérieure, PSL Research University, 2 rue Simone Iff, 75012 Paris, France

Published online: 04 April 2024

1 Introduction

We consider the general problem of unconstrained optimization. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a possibly non-convex function. Our goal is to solve the following problem

$$\min_{x \in \mathbb{R}^d} f(x). \tag{1.1}$$

In particular, we will consider the setting where (a) the function is smooth, that is, $f \in C^m(\mathbb{R}^d)$ with $m \in \mathbb{N}_+$ (f m-times continuously differentiable), and (b) we can evaluate it on given points, without the need of computing the gradient. For this class of problems there are known lower-bounds [1, 2] that show that it is not possible to achieve a global minimum with error ε with less than $O(\varepsilon^{-d/m})$ function evaluations. In this paper, we want to achieve this lower bound in terms of function evaluations, while having an optimization algorithm that has a running time that is polynomial in the underlying dimension and the number of function evaluations.

Several methods are available to solve this class of problems. For example, the function f can be approximated from its values at n sampled points, and the approximation of the function globally minimized instead of f. If the approximation is good enough, then this can be optimal in terms of n, but computationally infeasible. Optimal approximations can be obtained by multivariate polynomials [3] or functions in Sobolev spaces [4], with potentially adaptive ways of selecting points where the function is evaluated (see, e.g., [5] and references therein). Alternatively, when the function is itself a polynomial, algorithms based on the "sum-of-squares" paradigm can be used, but their computational complexity grows polynomially on $d^{r/2}$, where r is in the most favorable situations the order of the polynomial, but potentially larger when so-called hierarchies are used [6–8].

It turns out that the analysis of lower bounds on the number of function evaluations shows an intimate link between function interpolation and function minimization, i.e., the lower bounds of one problem are the same for the other problem. However, existing methods consider a two-step approach where (1) the function is approximated optimally, and (2) the approximation is minimized. In this paper, we consider a joint approach where approximation and optimization are done *jointly*.

We derive an algorithm that casts the possibly non-convex problem in Eq. (1.1) in terms of a simple convex problem based on a non-parametric representation of non-negative functions via positive definite operators [9]. As shown below, it can be considered as an infinite-dimensional counter-part to polynomial optimization with sums of squares, with two key differences: (1) the relaxation is always tight for the direct formulation, and (2) the computational cost does not depend on the dimension of the model (here infinite anyway), by using a subsampling algorithm and a computational trick common in statistics and machine learning.

The resulting algorithm with n sampled points will be able to achieve an error of $\varepsilon = O(n^{-m/d+3/d+1/2})$ as soon as $m \geq 3 + d/2$, with n function evaluations to reach the global minimum with precision ε, and a computational complexity of $O(n^{3.5} \log(1/\varepsilon))$ (with explicit constants). This is still not the optimal complexity in terms of the number of function evaluations (which is $\varepsilon = O(n^{-m/d})$), but this is

achieved with a polynomial-time algorithm in n. This is particularly interesting in the contexts where the function to be optimized is very smooth, i.e., $m \gg d$, possibly C^∞ or a polynomial. For example, if the function is differentiable at least $d + 3$ times, even if non-convex, the proposed algorithm finds the global minimum with error $O(n^{-1/2})$ and time $O(n^{3.5} \log n)$. Note that the (typically exponential) dependence on the dimensionality d is only in the constants and tracked explicitly in the rest of the paper.

Moreover, the algorithm is based on simple interior-point methods for semidefinite programming, directly implementable, and based only on function evaluations and matrix operations. It can thus leverage multiple GPU architectures to reach large values of n, which are needed when the dimension grows.

2 Outline of contributions

In this section, we present our framework, our algorithm and summarize the associated guarantees.

Denote by $\zeta \in \mathbb{R}^d$ a global minimizer of f and assume to know a bounded open region $\Omega \subset \mathbb{R}^d$ that contains ζ. We start with a straightforward and classical convex characterization of the problem in Eq. (1.1), with infinitely many constraints:

$$\max_{c \in \mathbb{R}} \; c \quad \text{such that} \quad \forall x \in \Omega, \; f(x) \geq c. \tag{2.1}$$

Note that the solution c_* of the problem above corresponds to $c_* = f(\zeta) = f_*$, the global minimum of f. The problem above is convex, but typically intractable to solve, due to the dense set of inequalities that c must satisfy.

To solve Eq. (2.1) our main idea is to represent the dense set of inequalities in terms of a dense set of *equalities* and then to approximate them by subsampling.

Tight relaxation. We start by introducing a quadratic form $\langle \phi(x), A\phi(x) \rangle$ with A a self-adjoint positive semidefinite operator from $\mathcal{H}$ to $\mathcal{H}$, for a suitable map $\phi : \Omega \to \mathcal{H}$ and an infinite-dimensional Hilbert space $\mathcal{H}$, to define the following problem

$$\max_{c \in \mathbb{R}, \; A \in \mathbb{S}_+(\mathcal{H})} \; c \quad \text{such that} \quad \forall x \in \Omega, \; f(x) - c = \langle \phi(x), A\phi(x) \rangle \,, \tag{2.2}$$

where $\mathbb{S}_+(\mathcal{H})$ is the set of bounded self-adjoint positive semi-definite operators on $\mathcal{H}$.

The problem in Eq. (2.2) has a smaller optimized objective function than the problem in Eq. (2.1) because we constrain A to be positive semi-definite and any feasible point for Eq. (2.2) is feasible for Eq. (2.1). When f is a polynomial and $\phi(x)$ is composed of monomials of degree less than half the degree of f (and thus $\mathcal{H}$ finite-dimensional), then we recover the classical "sum-of-squares" relaxation of polynomial optimization. In that situation, the relaxation is tight only if $f - f_*$ is itself a sum-of-squares, which is known to not always be the case. Then, to make the relaxation tight in the limit,

several hierarchies of polynomial optimization problems have been considered using polynomials of increasing degrees [6–8].

In this paper, we consider a well-chosen infinite-dimensional space $\mathcal{H}$, and we prove that if f is smooth enough (i.e., m-times differentiable with $m > 3 + d/2$), under mild geometrical assumptions on f then there always exists a map ϕ, and a finite rank $A_* \in \mathbb{S}_+(\mathcal{H})$ for which the problem in Eq. (2.1) and the one above are equivalent, that is, the relaxation is tight.

Note that, the resulting ϕ, despite being infinite-dimensional, has an explicit and easy-to-compute ($O(d)$ in memory and time) inner product $k(x, x') = \langle \phi(x), \phi(x') \rangle$ that will be the only quantity required to run the algorithm. We will thus use Hilbert spaces $\mathcal{H}$ which are reproducing kernel Hilbert spaces [10], such as Sobolev spaces [11].

Subsampling. We approximate the problem above as follows. Given a finite set $\widehat{X} = \{x_1, \dots, x_n\}$ which is a subset of Ω, we restrict the equality in Eq. (2.2) to only $x_1, \dots, x_n$.

Unlike the case of polynomial optimization where subsampling is exact if n is large enough [12], in our case subsampling leads to an error that decreases in n and depends on the regularity of f and of the map $x \mapsto \langle \phi(x), A\phi(x) \rangle$. While f is smooth enough by assumption, we need to control the regularity of the map induced by A, to guarantee that the constraints subsampled on $\widehat{X}$ well approximate the whole set of constraints on Ω. Then we consider a penalization term based on the trace of A and solve the following problem

$$\begin{aligned} \max_{c \in \mathbb{R},\ A \in \mathbb{S}_+(\mathcal{H})} \quad & c - \lambda \mathrm{Tr}(A) \\ \text{such that} \quad & \forall i \in \{1, \dots, n\},\ f(x_i) - c = \langle \phi(x_i), A\phi(x_i) \rangle \,, \end{aligned} \tag{2.3}$$

for some positive λ (with the implicit assumption that we optimize over operators A with finite trace). We show in this paper that solving Eq. (2.3) leads to an approximate optimum of the original problem in Eq. (2.1), when n is large enough and λ small enough. Note that the value of c which we obtain after subsampling is not anymore a lower bound on the global minimum, but we can provide both a priori and a posteriori certificates of optimality (see Sect. 8.2).

Finite-dimensional algorithm. The problem in Eq. (2.3) is still formulated in an infinite-dimensional space. We can leverage the particular choice of penalty by the trace of A and the choice of Hilbert space to obtain a finite-dimensional algorithm. Indeed, for reproducing kernel Hilbert spaces, then, following [9], we only need to solve the problem in the finite-dimensional Hilbert space spanned by $\phi(x_1), \dots, \phi(x_n)$, that is, we only need to look at A of the form $A = \sum_{i,j=1}^n C_{ij} \phi(x_i) \otimes \phi(x_j)$ for some positive semi-definite *matrix* $C \in \mathbb{R}^{n \times n}$. We can then write $\mathrm{Tr}(A) = \mathrm{Tr}(CK)$, with $K \in \mathbb{R}^{n \times n}$ the matrix of dot-products with $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$, and $\langle \phi(x_i), A\phi(x_i) \rangle = (KCK)_{ii}$.

Algorithm 1 Global minimum. Given $f : \mathbb{R}^d \to \mathbb{R}$, Ω, $n \in \mathbb{N}_+$, $\lambda > 0$, $s > d/2$.

1: $\widehat{X} \leftarrow \{x_1, \ldots, x_n\}$ ▷ Sampled i.i.d. uniformly on Ω
2: $f_j \leftarrow f(x_j), \quad \forall j \in [n]$

Features computation

3: $K_{ij} \leftarrow k(x_i, x_j) \;\; i, j \in [n]$ ▷ k Sobolev kernel of smoothness s, Eq. (3.2)
4: $R \leftarrow \texttt{cholesky}(K)$ ▷ upper triangular Cholesky
5: $\Phi_j = j$-th column of R, $\quad \forall j \in [n]$

Solution of the approximate problem (use any algorithm in Sect. 6)

6: $\hat{c} \leftarrow \max_{c \in \mathbb{R}, B \in \mathbb{S}_+(\mathbb{R}^n)} \; c - \lambda \mathrm{Tr}(B) \quad$ such that $\quad \forall j \in [n], f_j - c = \Phi_j^\top B \Phi_j$
7: **return** $\hat{c}$

Consider the Cholesky decomposition of K as $K = R^\top R$, with $R \in \mathbb{R}^{n \times n}$ upper-triangular. We can directly solve for $B = RCR^\top$, noting that $KCK = R^\top BR$ and $\mathrm{Tr}(CK) = \mathrm{Tr}(B)$. We can thus use a representation in terms of finite-dimensional vectors $\Phi_1, \ldots, \Phi_n \in \mathbb{R}^n$ defined as the columns of R. We thus study the following problem,

$$\begin{aligned} \max_{c \in \mathbb{R}, \; B \in \mathbb{S}_+(\mathbb{R}^n)} \quad & c - \lambda \mathrm{Tr}(B) \\ \text{such that} \quad & \forall i \in \{1, \ldots, n\}, \quad f(x_i) - c = \Phi_i^\top B \Phi_i. \end{aligned} \tag{2.4}$$

From an algorithmic viewpoint, the problem above can be solved efficiently since this is a semi-definite program. We show in Sect. 6 how we can apply Newton method and classical interior-point algorithms, leading to a computational complexity of $O(n^{3.5} \log(1/\varepsilon))$ in time and $O(n^2)$ in space.

Note that in the context of sum-of-squares polynomials, the relationship with reproducing kernel Hilbert spaces had been explored for approximation purposes after a polynomial optimization algorithm is used [13]. In this paper, we propose to leverage kernel methods *within* the optimization algorithm.

Why not simply subsampling the inequality? One straightforward algorithm is to subsample the dense set of *inequalities* in Eq. (2.1). Doing this will simply lead to outputting $\min_{i \in \{1, \ldots, n\}} f(x_i)$. This last algorithm, while easy to implement and convergent, is very slow, with a rate of $O(n^{-2/d})$ (see the discussion in Sect. 11). Subsampling the dense set of *equalities* in Eq. (2.2) allows to use smooth interpolation tools. When $\lambda = 0$, the optimal value is also $\min_{i \in \{1, \ldots, n\}} f(x_i)$ (if the kernel matrix is invertible, see Sect. 6), but for $\lambda > 0$, we can leverage smoothness as shown below.

Theoretical guarantees. From a theoretical viewpoint, denoting by $\hat{c}$ the minimizer of Eq. (2.4), we provide upper bounds for $|f_* - \hat{c}|$ with explicit constants and that hold under mild geometrical assumptions on f. We prove that the bound depends on how the points in $\widehat{X} = \{x_1, \ldots, x_n\}$ are chosen. In particular, we prove that when they are chosen uniformly at random on Ω, the problem in Eq. (2.4) achieves the global minimum with error ε with a precise dependence on n.

The results in this paper hold under the following assumptions.

Assumption 1 (Geometric properties on Ω and f). *The following holds:*

(a) *Let $\Omega = \cup_{x \in S} B_r(x)$, where S is a bounded subset of $\mathbb{R}^d$ and $B_r(x)$ is the open ball of radius $r > 0$, centered in x.*
(b) *The function f is in $C^2(\mathbb{R}^d)$. Ω contains at least one global minimizer. The minimizers in Ω are isolated points with strictly positive Hessian and their number is finite. There is no minimizer on the boundary of Ω.*

Note that Assumption 1(a) can be easily relaxed to Ω having locally Lipschitz-continuous boundaries [11, Section 4.9]. Assumption 1(b) is satisfied if all global minimizers of f are in Ω, and are second-order strict local minimizers. Note that similar assumptions are made to show finite convergence for polynomial optimization hierarchies [14].

Theorem 1 (Main result, informal). *Let $\Omega \subset \mathbb{R}^d$ be a ball of radius $R > 0$. Let $s > d/2$ and let k be the Sobolev kernel of smoothness s (see Example 1). Let $f \in C^{s+3}(\mathbb{R}^d)$ and that satisfies Assumption 1(b). Let $\hat{c}$ be the result of Algorithm 1 executed with $n \in \mathbb{N}_+$ points chosen uniformly at random in Ω and $\lambda > 0$. Let $\delta > 0$. There exist $n_{s,d,\delta}, C_{s,d} > 0$ such that, when $n > n_{s,d,\delta}$, and*

$$\lambda \;\geq\; C_{s,d}\, n^{-s/d+1/2} \left(\log \tfrac{n}{\delta}\right)^{s/d-1/2},$$

then, with probability at least $1 - \delta$,

$$|\hat{c} - f_*| \;\leq\; 3\,\lambda \left(Tr(A_*) + |f|_{\Omega, \lceil s-d/2 \rceil}\right),$$

where A_* is any solution of Eq. (2.2).

Note that A_* exists since $f \in C^{s+3}(\mathbb{R}^d)$ and it satisfies the geometrical mild condition in Assumption 1(b) (as we prove in Sect. 4), and that all constants can be made explicit (see Theorem 6). From the result above, and with $m = s+3$, for $s > d/2$, we can achieve an error of order $n^{-s/d+1/2}$, which translates to $\varepsilon = O(n^{-m/d+3/d+1/2})$ as soon as $m > d/2 + 3$. We pay the additional exponent 3 since we construct the candidate matrix representing the solution by requiring that each component of the Hessian of f, which is $m-2$ times differentiable belongs to the RKHS. This accounts for the 2 term, the last 1 is paid simply since s can be not integer. The rate for the class of functions $C^m(\Omega)$ is sub-optimal by an exponent $1/2+3/d$. In the following remark, we are going to show that our algorithm achieves nearly-optimal convergence rates when the function to optimize is in a Sobolev space. Denote by $W_2^s(\Omega)$ the Sobolev space of squared-integrable functions of smoothness $s > 0$, i.e., the space of functions whose weak derivatives up to order s are square-integrable on Ω, (see [11]).

Remark 1 **(Nearly optimal rates for Sobolev spaces.).** If Ω satisfies Assumption 1(a), f satisfies Assumption 1(b) and $f \in W_2^s(\Omega)$, with $s > d/2 + 3$, then Algorithm 1 with Sobolev kernel of smoothnes $s - 3$ achieves the convergence rate

$$O\left(n^{-s/d+1/2+3/d}\right),$$

modulo logarithmic factors, as proven in Theorem 6. When d is large, then the error exponent is asymptotically optimal, since the term $3/d$ becomes negligible, leading to the optimal exponent $-s/d + 1/2$ (see, e.g., [4, Prop. 1.3.11]).

Finding the global minimizer. In Sect. 7 we derive an extension of the problem in Eq. (2.4), with the goal of finding the global minimizer. Under the additional assumption that the minimizer is unique, we obtain a similar rate as Theorem 5 for the localization of the global minimizer.

Warm restart scheme for linear rates. Applying a simple warm restart scheme, we prove, in Sect. 7.2, that when f has a unique global minimum, then it is possible to achieve it with error ε, with a number of observations that is only logarithmic in ε

$$n = O(C_{d,m} \log(1/\varepsilon)),$$

for some constant $C_{d,m}$ that can be exponential in d (note that the added assumption of unique minimizer makes this result not contradict the lower bound in $\varepsilon^{-d/m}$).

Relationship to polynomial optimization. When f is a polynomial of degree $2r$, then it is natural to consider $\phi(x)$ composed of all monomials of degree less than r, leading to a space $\mathcal{H}$ of dimension $\binom{d+r}{r}$. All polynomials can be represented as $f(x) = c + \phi(x)^\top A \phi(x)$ for some symmetric matrix A. When $A \succcurlyeq 0$, by using its eigendecomposition, we can see that the polynomial $x \mapsto \phi(x)^\top A \phi(x)$ is a sum-of-squares polynomial.

However, in general A may not be positive semi-definite, as non-negative polynomials are not all sum-of-squares. Moreover, even when there exists a matrix $A \succcurlyeq 0$, the corresponding c may not be the minimum of f (it only needs to be a lower bound)—see, e.g., [6] and references therein.

If $f(x) - f_*$ is a sum of squares, then, with $\lambda = 0$ and $n = \binom{d+2r}{2r}$ points (to ensure that subsampling is exact), we exactly get the minimum of f, as we are solving *exactly* the usual optimization problem.

When $f(x) - f_*$ is not a sum of squares, then a variety of hierarchies have been designed when optimization is performed on a compact constraint set described with polynomial inequalities (such as taking an ℓ_2-ball for Ω), that augment the problem dimensionality to reach global convergence [6–8]. In Sect. 9, we show how our framework fits with one these hierarchies, and also can provide computational gains.

Note that our framework, by looking directly at an infinite-dimensional space circumvents the need for hierarchies, and solves a single optimization problem. The difficulty is that it requires sampling. Moreover by using only kernel evaluations, we circumvent the explicit construction of a basis for $\mathcal{H}$, which is computationally cumbersome when d grows.

Organization of the paper. The paper is organized as follows: in Sect. 3, we present the kernel setting our paper relies on; then, in Sect. 4, we analyze the infinite-dimensional problem and show its equivalence with global minimization. Then, in Sect. 5, we present our theoretical guarantee for the finite-dimensional algorithm, as summarized in Theorem 1. In Sect. 6 we present the dual algorithm based on self-concordant barriers and the damped Newton algorithm. In Sect. 7, we present our extension to find the global minimizer, while in Sect. 8, we provide certificates of optimality for potentially inexactly solved problems. In Sect. 9, we discuss further relationships with polynomial hierarchies, and provide illustrative experiments in Sect. 10. We conclude in Sect. 11 with a discussion opening up to many future problems.

3 Setting

In this section, we first introduce some definitions and notation about *reproducing Kernel Hilbert spaces* in Sect. 3.1 (for more details, see [15, 16]), and present our detailed assumptions in Sect. 3.2. In Sect. 4 we show how our infinite-dimensional sum-of-squares representation can be built, and in Sect. 5 we provide guarantees on subsampling.

3.1 Definitions and notation

In this section we denote by $u \cdot v$, $a \circ v$ respectively the pointwise multiplication between the functions u and v, and the composition between the functions a and v. We denote by $\mathbb{N}$ the set of natural numbers including 0, by $\mathbb{N}_+$ the set $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$ and $[n]$ the set $\{1, \dots, n\}$ for $n \in \mathbb{N}_+$. We will always consider $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$ if not specified otherwise. Moreover we denote by $B_r(z)$ the open ball $B_r(z) = \{x \in \mathbb{R}^d \mid \|x - z\| < r\}$. Let $\Omega \subseteq \mathbb{R}^d$ be an open set. Let $\alpha \in \mathbb{N}^d$. We introduce the following *multi-index notation* $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $\partial_x^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ [11]. For $m \in \mathbb{N}$, and Ω an open set of $\mathbb{R}^d$, denote by $C^m(\Omega)$ the set of m-times differentiable functions on Ω with continuous m-th derivatives. For any function u defined on a superset of Ω and m times differentiable on Ω, define the following semi norm.

$$|u|_{\Omega,m} = \max_{|\alpha|=m} \sup_{x \in \Omega} \left|\partial^\alpha u(x)\right|. \tag{3.1}$$

Positive definite matrices and operators. Let $\mathcal{H}$ be a Hilbert space, endowed with the inner product $\langle \cdot, \cdot \rangle$. Let $A : \mathcal{H} \to \mathcal{H}$ be a linear operator and denote by A^* the adjoint operator, by $\mathrm{Tr}(A)$ the trace of A and by $\|\cdot\|_F$ the Hilbert-Schmidt norm $\|A\|_F^2 = \mathrm{Tr}(A^*A)$. We always endow $\mathbb{R}^p$ with the standard inner product $x^\top y = \sum_{i=1}^p x_i y_i$ for any $x, y \in \mathbb{R}^p$. In the case $\mathcal{H} = \mathbb{R}^p$, with the standard inner product, then $A \in \mathbb{R}^{p \times p}$ is a matrix and the Hilbert-Schmidt norm corresponds to the Frobenius norm. We say that $A \succeq 0$ or A is a *positive operator* (positive matrix if $\mathcal{H}$ is finite dimensional),

when A is bounded, self-adjoint, and $\langle u, Au\rangle \geq 0,\ \forall u \in \mathcal{H}$. We denote by $\mathbb{S}_+(\mathcal{H})$ the space of positive operators on $\mathcal{H}$. Moreover, we denote by $A \succ 0$, or A strictly positive operator, the case $\langle u, Au\rangle > 0$ for all $u \in \mathcal{H}$ such that $u \neq 0$.

Kernels and reproducing kernel Hilbert spaces. For this section we refer to [15–17], for more details (see also Appendix A.3). Let Ω be a set. A function $k : \Omega \times \Omega \to \mathbb{R}$ is called a *positive definite kernel* if all matrices of pairwise evaluations are positive semi-definite, that is, if it satisfies the following equation

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0, \qquad \forall n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in \Omega.$$

Given a kernel k, the *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$, with the associated inner product $\langle \cdot, \cdot \rangle$, is a space of real functions with domain Ω, with the following properties.

(a) The function $k_x = k(x, \cdot)$ satisfies $k_x \in \mathcal{H}$ for any $x \in \Omega$.
(b) The inner product satisfies $\langle f, k_x\rangle = f(x)$ for all $f \in \mathcal{H}$, $x \in \Omega$. In particular $\langle k_{x'}, k_x\rangle = k(x', x)$ for all $x, x' \in \Omega$.

In other words, function evaluations are uniformly bounded and continuous linear forms and the k_x are the evaluation functionals. The norm associated to $\mathcal{H}$ is the one induced by the inner product, i.e., $\|f\|^2 = \langle f, f\rangle$. We remark that given a kernel on Ω there exists a unique associated RKHS on Ω [10]. Moreover, the kernel admits a characterization in terms of a *feature map* ϕ,

$$\phi : \Omega \to \mathcal{H}, \quad \text{defined as} \quad \phi(x) = k(x, \cdot) = k_x, \quad \forall x \in \Omega.$$

Indeed according to the point (b) above, we have $k(x, x') = \big\langle \phi(x), \phi(x') \big\rangle$ for all $x, x' \in \Omega$. We will conclude the section with an example of RKHS that will be useful in the rest of the paper.

Example 1 (Sobolev kernel [18]). Let $s > d/2$, with $d \in \mathbb{N}_+$, and Ω be a bounded open set. Let

$$k_s(x, x') = c_s \|x - x'\|^{s-d/2} \mathcal{K}_{s-d/2}(\|x - x'\|), \quad \forall x, x' \in \Omega, \tag{3.2}$$

where $\mathcal{K} : \mathbb{R}_+ \to \mathbb{R}$ the Bessel function of the second kind (see, e.g., 5.10 in [18]) and $c_s = \frac{2^{1+d/2-s}}{\Gamma(s-d/2)}$. The constant c_s is chosen such that $k_s(x, x) = 1$ for any $x \in \Omega$. In the particular case of $s = d/2+1/2$, we have $k(x, x') = \exp(-\|x-x'\|)$. Note that a scale factor is often added as $k(x, x') = \exp(-\|x - x'\|/\sigma)$ in this last example. In such case, all bounds that we derive in this paper would then have extra factors proportional to powers of σ. To conclude, when Ω has locally Lipschitz boundary (a sufficient condition is Assumption 1(a)) then $\mathcal{H} = W_2^s(\Omega)$, where $W_2^s(\Omega)$ is the Sobolev space of functions whose weak-derivatives up to order s are square-integrable [11]. Moreover, in this case $\|\cdot\|_{\mathcal{H}}$ is equivalent to $\|\cdot\|_{W_2^s(\Omega)}$.

Reproducing kernel Hilbert spaces are classically used in fitting problems, such as appearing in statistics and machine learning, because of function evaluations $f \mapsto f(x)$ are bounded operators for any x, and optimization problems involving f only through function evaluations at a finite number of points $x_1, \ldots, x_n$, and penalized with the norm $\|f\|$, can be solved by looking only a f of the form $f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$ [15, 16]. We will use an extension of this classical "representer theorem" to operators and spectral norms in Sect. 5.

3.2 Precise assumptions on reproducing kernel Hilbert space

On top of Assumption 1 (made on the function f and the set Ω), we make the following assumptions on the space $\mathcal{H}$ and the associated kernel k.

Assumption 2 (Properties of the space $\mathcal{H}$). *Given a bounded open set $\Omega \subset \mathbb{R}^d$, let $\mathcal{H}$ be a space of functions on Ω with norm $\|\cdot\|_{\mathcal{H}}$, satisfying the following conditions*

(a) $w|_\Omega \in \mathcal{H}$, $\forall w \in C^\infty(\mathbb{R}^d)$. *Moreover there exists* $\mathsf{M} \geq 1$ *such that*

$$\|u \cdot v\|_{\mathcal{H}} \leq \mathsf{M} \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}, \quad \forall u, v \in \mathcal{H}.$$

(b) *$a \circ v \in \mathcal{H}$, for any $a \in C^\infty(\mathbb{R}^p)$, $v = (v_1, \ldots, v_p)$, $v_j \in \mathcal{H}$, $j \in [p]$.*

(c) *Let $z \in \mathbb{R}^d$, $r > 0$ s.t. the ball $B_r(z)$ is in Ω. For any $u \in \mathcal{H}$, there exists $g_{r,z} \in \mathcal{H}$ s.t.*

$$g_{r,z}(x) = \int_0^1 (1-t) u(z + t(x-z)) dt, \quad \forall x \in B_r(z).$$

(d) *$\mathcal{H}$ is a RKHS with associated kernel k. For some $m \in \mathbb{N}_+$ and some $\mathsf{D}_m \geq 1$, the kernel k satisfies*

$$\max_{|\alpha|=m} \sup_{x,y \in \Omega} |\partial_x^\alpha \partial_y^\alpha k(x, y)| \leq \mathsf{D}_m^2 < \infty.$$

Assumptions 2(a) to 2(c) above require essentially that functions in $\mathcal{H}$ (a) can be multiplied by other functions in $\mathcal{H}$, or by infinitely smooth functions, and still be in $\mathcal{H}$; (b) that can be composed with infinitely smooth functions, or (c) integrated, and still be in $\mathcal{H}$. Moreover Assumption 2(d) requires that $\mathcal{H}$ is a RKHS with a kernel that is m-times differentiable. An interesting consequence of Assumption 2(d) is the following remark (for more details, see, e.g., [17, Corollary 4.36]).

Remark 2 Assumption 2(d) guarantees that $\mathcal{H} \subseteq C^m(\Omega)$ and $|u|_{\Omega,m} \leq \mathsf{D}_m \|u\|_{\mathcal{H}}$.

Note that Assumptions 2(a) to 2(c) are the only required in Sect. 4 to prove the crucial decomposition in Theorem 2 and are satisfied by notable spaces (that are not necessarily RKHS) like $C^s(\Omega)$ or Sobolev spaces $W_p^s(\Omega)$ with $s > d/p$ and $p \in [1, \infty]$. Instead, Assumption 2(d) is required for the analysis of the finite-dimensional problem and in particular Theorems 4 and 5. In the following proposition we show that $W_2^s(\Omega)$ with $s > d/2$ and Ω satisfying Assumption 1(a) satisfy the whole of Assumption 2.

Proposition 1 (Sobolev kernels satisfy Assumption 2). *Let Ω be a bounded open set of $\mathbb{R}^d$. The Sobolev kernel with $s > d/2$ recalled in Example 1 satisfies Assumption 2 for any $m \in \mathbb{N}_+$, $m < s - \frac{d}{2}$ and*

$$\mathsf{M} = (2\pi)^{d/2} 2^{s+1/2}, \qquad \mathsf{D}_m = (2\pi)^{d/4} \sqrt{\frac{\Gamma(m + d/2)\Gamma(s - d/2 - m)}{\Gamma(s - d/2)\Gamma(d/2)}}.$$

The proof of proposition above is in Appendix D.2. We make a last assumption regarding the differentiability of f, namely that f and its second-derivatives are in $\mathcal{H}$.

Assumption 3 (Analytic properties of f). *The function f satisfies $f|_\Omega \in C^2(\Omega) \cap \mathcal{H}$ and $\frac{\partial^2 f}{\partial x_i \partial x_j}|_\Omega \in \mathcal{H}$ for all $i, j \in [d]$.*

4 Equivalence of the infinite-dimensional problem

In Theorem 2 and Corollary 1, we provide a representation of $f - f_*$ in terms of an infinite-dimensional, *but finite-rank*, positive operator, under basic geometric conditions on f and algebraic properties of $\mathcal{H}$. In Theorem 3 we use this operator to prove that Eq. (2.2) achieves the global minimum of f. In this section we analyze the conditions under which the problem in (2.2) has the same solution as the one in Eq. (2.1).

The proof follows by explicitly constructing a bounded positive operator A_* (which will have finite trace) that satisfy $f(x) - f_* = \langle \phi(x), A_* \phi(x) \rangle$ for all $x \in \Omega$. Note that, by construction $f - f_*$ is a non-negative function. If $w := \sqrt{f - f_*} \in \mathcal{H}$ then $A_* = w \otimes w$ would suffice. However, denoting by $\zeta \in \Omega$ a global minimizer, note that $f(\zeta) - f_* = 0$ and the smoothness of $\sqrt{f - f_*}$ may degrade around ζ, making $\sqrt{f - f_*} \notin \mathcal{H}$ even if $f - f_* \in \mathcal{H}$.

Here we follow a different approach. In Lemma 1 we provide a decomposition that represents the function $f - f_*$ locally around each global optimum using the fact that it is locally strongly convex around the minimizers. In the proof of Theorem 2 we provide a decomposition of the function far from the optimal points; we then glue these different decompositions via bump functions.

Lemma 1 *Let $\mathcal{H}$ be a space of functions on Ω that satisfy Assumptions 2(a) to 2(c). Let $\zeta \in \Omega$ and $r, \gamma > 0$. Let $B_r(\zeta) \subset \Omega$ be a ball centered in ζ of radius r and $g \in C^2(\Omega)$ satisfy $g(\zeta) = 0$, $\nabla^2 g(x) \succcurlyeq \gamma I$ for $x \in B_r(\zeta)$ and $\frac{\partial^2}{\partial x_i \partial x_j} g \in \mathcal{H}$ for $i, j \in [d]$. Then, there exists $w_j \in \mathcal{H}$, $j \in [d]$ such that*

$$g(x) = \sum_{j=1}^{d} w_j(x)^2, \quad \forall x \in B_r(\zeta). \tag{4.1}$$

Proof Let $x \in B_r(\zeta)$ and consider the function $h(t) = g(\zeta + t(x - \zeta))$ on $[0, 1]$. Note that $h(0) = g(\zeta)$ and $h(1) = g(x)$. Taking the Taylor expansion of h of order

1, we have $h(1) = h(0) + h'(0) + \int_0^1 (1-t)h''(t)dt$, with $h(0) = g(\zeta)$, $h'(0) = (x-\zeta)^\top \nabla g(\zeta)$ and $h''(t) = (x-\zeta)^\top \nabla^2 g(\zeta + t(x-\zeta))(x-\zeta)$. Since $g(\zeta) = 0$ by construction and $\nabla g(\zeta) = 0$ since ζ is a local minimizer of g, we have $h(0) = h'(0) = 0$ leading to

$$g(x) = (x-\zeta)^\top R(x)(x-\zeta), \quad R(x) = \int_0^1 (1-t)\nabla^2 g(\zeta + t(x-\zeta))dt. \quad (4.2)$$

Note that for $x \in B_r(\zeta)$ we have $\nabla^2 g(x) \succcurlyeq \gamma I$ and so $R(x) \succcurlyeq \gamma I$. In particular, this implies that for any $x \in B_r(\zeta)$, $S(x) = \sqrt{R(x)}$ is well defined ($\sqrt{\cdot} : \mathbb{S}_+(\mathbb{R}^d) \to \mathbb{S}_+(\mathbb{R}^d)$ is the spectral square root, where for any $M \in \mathbb{S}_+(\mathbb{R}^d)$ and any eigen-decomposition $M = \sum_{j=1}^d \lambda_j u_j u_j^\top$, $\sqrt{M} = \sum_{j=1}^d \sqrt{\lambda_j} u_j u_j^\top$). Thus,

$$\forall x \in B_r(\zeta),\ g(x) = (x-\zeta)^\top S(x)S(x)(x-\zeta) = \sum_{i=1}^d \left(e_i^\top S(x)(x-\zeta)\right)^2.$$

The following steps prove the existence of $w_i \in \mathcal{H}$ such that $w_i|_{B_r(\zeta)} = e_i^\top S(\cdot)(\cdot - \zeta)$. Let $(e_1, ..., e_d)$ be the canonical basis of $\mathbb{R}^d$ and $\mathbb{S}(\mathbb{R}^d)$ be the set of symmetric matrices on $\mathbb{R}^d$ endowed with Frobenius norm, in the rest of the proof we identify it with the isometric space $\mathbb{R}^{d(d+1)/2}$ (corresponding of taking the upper triangular part of the matrix and reshaping it in form of a vector).

Step 1. *There exists a function $\overline{R} : \Omega \to \mathbb{S}(\mathbb{R}^d)$, such that*

$$\forall i, j \in [d],\ e_i^\top \overline{R} e_j \in \mathcal{H} \text{ and } \overline{R}|_{B_r(\zeta)} = R.$$

This is a direct consequence of the fact that $\frac{\partial^2}{\partial x_i \partial x_j} g \in \mathcal{H}$ for all $i \le j \in [d]$, of Assumption 2(c) and of the definition of R in Eq. (4.2).

Step 2. *There exists a function $\overline{S} : \Omega \to \mathbb{S}(\mathbb{R}^d)$ such that*

$$\forall i, j \in [d],\ e_i^\top \overline{S} e_j \in \mathcal{H} \text{ and } \forall x \in B_r(\zeta),\ \overline{S}(x) = \sqrt{R}(x).$$

Let $\tau := \sup_{x \in B_r(\zeta)} \|R(x)\|_{\mathrm{op}} = \|\overline{R}(x)\|_{\mathrm{op}}$, which is well defined because R is continuous since $g \in C^2(\Omega)$. Define the compact set $K = \{T \in \mathbb{S}(\mathbb{R}^d) \mid \gamma I \preceq T \preceq \tau I\}$ and the open set $U = \{T \in \mathbb{S}(\mathbb{R}^d) \mid \frac{\gamma}{2} I \prec T \prec 2\tau I\}$. Note that $K \subset U \subset \mathbb{S}(\mathbb{R}^d)$.

Fix $i, j \in [d]$ and consider the function $\theta_{i,j} : U \to \mathbb{R}$ defined by $\theta_{i,j}(M) = e_i^\top \sqrt{M} e_j$. Since the square root $\sqrt{\cdot} : \mathbb{S}_+(\mathbb{R}^d) \to \mathbb{S}_+(\mathbb{R}^d)$ is infinitely differentiable (see e.g. the explicit construction in [19] Thm. 1.1) and $U \subset \mathbb{S}_+(\mathbb{R}^d)$ then $\theta_{i,j}$ is infinitely differentiable on U, i.e., $\theta_{i,j} \in C^\infty(U)$. By Proposition 7, since K is a compact set in U, there exists $\overline{\theta}_{i,j} \in C_0^\infty(\mathbb{S}(\mathbb{R}^d))$ such that $\forall T \in K$, $\overline{\theta}_{i,j}(T) = \theta_{i,j}(T)$.

Define $\overline{S}(x) = \sum_{i,j \in [d]} (\overline{\theta}_{i,j} \circ \overline{R})(x) e_i e_j^\top$ for any $x \in \Omega$. Applying Assumption 2(b), $e_i^\top \overline{S} e_j = \overline{\theta}_{i,j} \circ \overline{R} \in \mathcal{H}$ since the $\overline{R}_{k,l} \in \mathcal{H}$, $k, l \in [d]$ and $\overline{\theta}_{i,j}$ is in

$C_0^\infty(\mathbb{S}(\mathbb{R}^d))$. Moreover, by construction, for any $x \in B_r(\zeta)$, we have $\overline{R}(x) = R(x) \in K$ and so

$$\overline{S}_{i,j}(x) = \overline{\theta}_{i,j}(\overline{R}(x)) = \theta_{i,j}(R(x)) = e_i^\top \sqrt{R(x)} e_j.$$

Note that here, we have applied Proposition 7 and Assumption 2(b) to $\mathbb{S}(\mathbb{R}^d)$ and not to $\mathbb{R}^{d(d+1)/2}$; this can be made formal by using the linear isomorphism between $\mathbb{S}(\mathbb{R}^d)$ endowed with the Frobenius norm and $\mathbb{R}^{d(d+1)/2}$ endowed with the Euclidean norm.

Step 3. *There exists a function $\overline{h} = (\overline{h}_j)_{j\in[d]} : \Omega \to \mathbb{R}^d$ such that*

$$\forall j \in [d], \overline{h}_j \in \mathcal{H} \text{ and } \forall x \in B_r(\zeta),\ \overline{h}(x) = x - \zeta.$$

Fix $j \in [n]$. Define $\overline{B}_r(\zeta) = K \subset U = B_{2r}(\zeta)$ and apply Proposition 7 to $x \in U \mapsto e_j^\top(x - \zeta)$ to get $h_j \in C_0^\infty(\mathbb{R}^d)$ which coincides with $e_j^\top(\cdot - \zeta)$ on K hence on $B_r(\zeta)$. Applying Assumption 2(a), the restriction $\overline{h}_j = h_j|_\Omega$ is in $\mathcal{H}$, and hence $\overline{h} = \sum_{j\in[d]} \overline{h}_j e_j$ satisfies the desired property.

Step 4. *The $w_i = e_i^\top \overline{S}\, \overline{h},\ i \in [d]$ have the desired property.*

It is clear that the w_i are in $\mathcal{H}$ as a linear combination of products of functions in $\mathcal{H}$ (see Assumption 2(a)), since $w_i = \sum_{j\in[d]} \overline{S}_{ij}(x)\overline{h}_j(x)$ for any $x \in \Omega$. Moreover,

$$\sum_{i\in[d]} w_i^2 = \overline{h}^\top \overline{S}^\top \left(\sum_{i=1}^d e_i e_i^\top\right) \overline{S}\, \overline{h} = \overline{h}^\top \overline{S}^2 \overline{h}.$$

Using the previous points,

$$\forall x \in B_r(\zeta),\ \sum_{i\in[d]} w_i^2(x) = \overline{h}^\top(x)\overline{S}^2(x)\overline{h}(x) = (x-\zeta)^\top R(x)(x-\zeta) = g(x).$$

□

Now we are going to use the local representations provided by the lemma above to build a global representation in terms of a finite-rank positive operator. Indeed far from the global optima the function $f - f_*$ is strictly positive and so we can take a smooth extension of the square root to represent it and glue it with the local representations around the global optima via bump functions as follows.

Theorem 2 *Let Ω be a bounded open set and let $\mathcal{H}$ be a space of functions on Ω that satisfy Assumptions 2(a) to 2(c). Let f satisfy Assumptions 1(b) and 3. Then there exist $w_1, \ldots, w_q \in \mathcal{H}$ with $q \le dp + 1$ and $p \in \mathbb{N}_+$ the number of minimizers in Ω, such that*

$$f(x) - f_* = \sum_{j\in[q]} w_j(x)^2, \qquad \forall\, x \in \Omega. \tag{4.3}$$

Proof Let $Z = \{\zeta_1, \ldots, \zeta_p\}$, $p \in \mathbb{N}_+$ be the *non-empty* set of global minima of f, according to Assumption 1(b). Denote by $f_* = \min_{x\in\Omega} f(x)$ the global minimum of f, and by $g : \Omega \to \mathbb{R}$ the function $g = f|_\Omega - f_*\mathbf{1}|_\Omega$ where $\mathbf{1}$ is the function $\mathbf{1}(x) = 1$ for any $x \in \mathbb{R}^d$. Assumption 3 implies that $\nabla^2 g = \nabla^2 f|_\Omega$ is continuous, an that $\frac{\partial^2 g}{\partial x_i \partial x_j} \in \mathcal{H}$ for any $i, j \in [d]$. Moreover, $g \in \mathcal{H}$. Indeed, by construction $f_*\mathbf{1}$ is in $C^\infty(\mathbb{R}^d)$, and since $\mathcal{H}$ satisfies Assumption 2(a), $f_*\mathbf{1}|_\Omega \in \mathcal{H}$. Since $f|_\Omega \in \mathcal{H}$ by Assumption 3, then $g \in \mathcal{H}$.

Step 1. *There exists $r > 0$ and $\alpha > 0$ such that (i) the $B_r(\zeta_l)$, $l \in [p]$ are included in Ω and (ii) for any $x \in \bigcup_{l\in[p]} B_r(\zeta_l)$, it holds $\nabla^2 g(x) \succeq \alpha I$.*

By Assumption 1(b), for all $\zeta \in Z$, $\nabla^2 g(\zeta) \succ 0$. Since $\nabla^2 g$ is continuous, Z is a finite set, and Ω is an open set, there exists a radius $r > 0$ and $\alpha > 0$ such that for all $l \in [p]$, $B_r(\zeta_l) \subset \Omega$ and $\nabla^2 g|_{B_r(\zeta_l)} \succeq \alpha I$. For the rest of the proof, fix r, α satisfying this property. For any $X \subset \Omega$ denote with $\mathbf{1}_X$ the indicator function of a X in Ω. We define $\chi_0 = \mathbf{1}_{\Omega\setminus\bigcup_{l\in[p]} B_{r/2}(\zeta_l)}$, and $\chi_l = \mathbf{1}_{B_r(\zeta_l)}$, $l \in [p]$.

Step 2. *There exists $w_0 \in \mathcal{H}$ s.t. $w_0^2\chi_0 = g\chi_0$.*

Ω is bounded and by Assumption 1(b), the set of global minimizers of f included in Ω is finite and there is no mimimizer of f on the boundary, i.e., there exists $m_1 > 0$ and a compact $K \subset \Omega$ such that $\forall x \in \Omega \setminus K$, $g(x) \geq m_1$.

Moreover, f has no global optima on the compact $K \setminus \bigcup_{\zeta\in Z} B_{r/2}(\zeta)$ since the set of global optima is Z, hence the existence of $m_2 > 0$ such that $\forall x \in K \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l)$, $g(x) \geq m_2$. Taking $m = \min(m_1, m_2)$, it holds $\forall x \in \Omega \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l)$, $g(x) \geq m > 0$. Since $f \in C^2(\Omega)$, f is also bounded above on Ω hence the existence of $M > 0$ such that $g \leq M$. Thus

$$\forall x \in \Omega \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l),\ g(x) \in I \subset (m/2, 2M), \qquad I = [m, M].$$

Since $\sqrt{\cdot} \in C^\infty((m/2, 2M))$, $(m/2, 2M)$ is an open subset of $\mathbb{R}$ and I is compact, applying Proposition 7, there exists a smooth extension $s_I \in C_0^\infty(\mathbb{R})$ such that $s_I(t) = \sqrt{t}$ for any $t \in I$. Now since $g \in \mathcal{H}$ and $s_I \in C_0^\infty(\mathbb{R})$, by Assumption 2(b), $w_0 := s_I \circ g \in \mathcal{H}$. Since $\forall x \in \Omega \setminus \bigcup_{l\in[p]} B_{r/2}(\zeta_l)$, $g \in I$, this shows $g\chi_0 = w_0^2\chi_0$.

Step 3. *For all $l \in [p]$, there exists $(w_{l,j})_{j\in[d]} \in \mathcal{H}^d$ s.t. $g(x)\chi_l = \sum_{j=1}^d w_{l,j}^2\ \chi_l$.*

This is an immediate consequence of Lemma 1 since $\nabla^2 g(x) \succeq \alpha I$ on $B_r(\zeta_l)$.

Step 4. *There exists $b_l \in C^\infty(\mathbb{R}^d)$ s.t. $b_l = b_l\ \chi_l$ for all $l \in \{0, 1, \ldots, p\}$ and $\sum_{l=0}^p b_l^2 = 1$.* This corresponds to Lemma 7, Appendix A.4 applied to the balls $B_r(\zeta_l)$, $l \in [p]$.

Step 5. Using all the previous steps

$$\begin{aligned} g &= \sum_{l=0}^p g\ b_l^2 = \sum_{l=0}^p g(\chi_l\ b_l)^2 = \sum_{l=0}^p (\chi_l\ g)\ (\chi_l b_l^2) \\ &= (\chi_0 w_0^2)\ (\chi_0\ b_0^2) + \sum_{l=1}^p \left(\chi_l \sum_{j=1}^d w_{l,j}^2\right)\ \chi_l\ b_l^2 \end{aligned}$$

$$= ([b_0\ \chi_0]\ w_0)^2 + \sum_{l=1}^{p}\sum_{j=1}^{d}([b_l\ \chi_l]\ w_{l,j})^2 = (b_0\ w_0)^2 + \sum_{l=1}^{p}\sum_{j=1}^{d}(b_l\ w_{l,j})^2.$$

Applying Assumption 2(a) to each function inside the squares in the previous expressions yields the result. □

A direct corollary of the theorem above is the existence of $A_* \in \mathbb{S}_+(\mathcal{H})$ when $\mathcal{H}$ is a reproducing kernel Hilbert space satisfying the assumptions of Theorem 2.

Corollary 1 *Let k be a kernel whose associated RKHS $\mathcal{H}$ satisfies Assumptions 2(a) to 2(c) and let f satisfy Assumptions 1(b) and 3, then there exists $A_* \in \mathbb{S}_+(\mathcal{H})$ with $rank(A_*) \leq d|Z| + 1$ such that $f(x) - f^* = \langle \phi(x), A_*\phi(x)\rangle$ for all $x \in \Omega$.*

Proof By Theorem 2 we know that if f satisfies Assumptions 1(b) and 3 w.r.t. a space $\mathcal{H}$ that satisfies Assumptions 2(a) to 2(c), there exists $w_1, \ldots, w_q \in \mathcal{H}$ with $q \leq d|Z| + 1$ such that $f(x) - f^* = \sum_{j\in[q]} w_j^2(x)$ for any $x \in \Omega$. Since $\mathcal{H}$ is a reproducing kernel Hilbert space, for any $h \in \mathcal{H},\ x \in \Omega$ we have $h(x) = \langle \phi(x), h\rangle_{\mathcal{H}}$. Moreover, by the properties of the outer product in Hilbert spaces, for any $h, v \in \mathcal{H}$, it holds $(\langle h, v\rangle_{\mathcal{H}})^2 = \langle h, (v \otimes_{\mathcal{H}} v)h\rangle$.

Thus, for any $x \in \Omega$, $j \in [q]$, it holds $w_j(x)^2 = \left\langle \phi(x), (w_j \otimes w_j)\phi(x)\right\rangle$ and hence

$$\forall x \in \Omega,\ f(x) - f^* = \langle \phi(x), A_*\phi(x)\rangle\,, \qquad A_* = \sum_{j\in[q]} w_j \otimes w_j.$$

□

The following corollary corresponds to the application of Theorem 2 where Assumptions 2(a) to 2(c) are satisfied by requiring only that $f \in C^{s+2}(\mathbb{R}^d)$ and shows that any non-negative $f \in C^{s+2}(\mathbb{R}^d)$ satisfying the geometric conditions admits a sum-of-squares decomposition (more details in Appendix G.2).

Corollary 2 *Let Ω be a bounded open set and $f \in C^{s+2}(\mathbb{R}^d)$, $s \in \mathbb{N}$, satisfying Assumption 1(b). Then there exist $w_1, \ldots, w_p \in C^s(\mathbb{R}^d)$, $p \in \mathbb{N}_+$, such that*

$$\forall x \in \Omega,\ f(x) - f_* = \sum_{j\in[p]} w_j^2(x).$$

To conclude the section we prove the problem in Eq. (2.2) admits a maximizer whose non-negative operator is of rank at most $d|Z| + 1$.

Theorem 3 *Let $\Omega \subset \mathbb{R}^d$ be an open set, k be a kernel, $\mathcal{H}$ the associated RKHS, and $f : \mathbb{R}^d \to \mathbb{R}$. Under Assumptions 1 to 3, the problem in Eq. (2.2) admits an optimal solution (c_*, A_*) with $c_* = f_*$, and A_* a positive operator on $\mathcal{H}$ with rank at most $d|Z| + 1$.*

Proof Let p_0 be the maximum of Eq. (2.1). Since $A \succeq 0$ implies $\langle \phi(x), A\phi(x)\rangle \geq 0$ for all $x \in \Omega$, the problem in Eq. (2.1) is a relaxation of Eq. (2.2), where the constraint $f(x) - c = \langle \phi(x), A\phi(x)\rangle$ is substituted by $f(x) - c \geq 0, \forall x \in \Omega$. Then $p_0 \geq p_*$ if a maximum p_* exists for Eq. (2.2). Moreover if there exists A that satisfies the constraints in Eq. (2.2) for the value $c_* = f_*$, then $p_0 = p_*$ and (c_*, A) is a maximizer of Eq. (2.2). The proof is concluded by applying Corollary 1 that shows that there exists A satisfying the constraints in Eq. (2.2) for the value $c = f_*$. □

In Corollary 1 and Theorem 3 we proved the existence of an infinite-dimensional trace-class positive operator A_* that satisfies $\langle \phi(x), A_*\phi(x)\rangle = f(x) - f_*$ for any $x \in \Omega$ and maximizing Eq. (2.2). The proof is quite general, requiring some geometric properties on f, the fact that f and its second derivatives belong to $\mathcal{H}$ and some algebraic properties of the space $\mathcal{H}$, in particular to be closed to multiplication with a C^∞ function, to integration, and to composition with a C^∞ map. The generality of the proof does not allow to derive an easy characterization of the trace of A_*.

5 Properties of the finite-dimensional problem

In the previous section we proved that there exists a finite rank positive operator A_* minimizing Eq. (2.2). In this section we study the effect of the discretization of Eq. (2.2) on a given a set of distinct points $\widehat{X} = \{x_1, \ldots, x_n\}$. First, we derive Theorem 4 which is fundamental to prove Theorem 5, and is our main technical result (we believe it can have a broader impact beyond the use in this paper as discussed in Sect. 11). Given a smooth function g on Ω, in Theorem 4 we prove that if there exists a matrix $B \in \mathbb{S}_+(\mathbb{R}^n)$ such that $g(x_i) = \Phi_i^\top B \Phi_i$ for $i \in [n]$ (the vectors $\Phi_j \in \mathbb{R}^n$ are defined before Eq. (2.4)), then the inequality $g(x) \geq -\varepsilon$ holds for any $x \in \Omega$ for an ε depending on the smoothness of the kernel, the smoothness of g and how well the points in $\widehat{X}$ cover Ω. We denote by $h_{\widehat{X},\Omega}$ the *fill distance* [18],

$$h_{\widehat{X},\Omega} = \sup_{x\in\Omega} \min_{i\in[n]} \|x - x_i\|, \tag{5.1}$$

corresponding to the maximum distance between a point in Ω and the set $\widehat{X}$. In particular, if the kernel and g are m-times differentiable, Theorem 4 proves that $g(x) \geq -\varepsilon$ holds with $\varepsilon = O(h^m_{\widehat{X},\Omega})$ which is an improvement when $m \gg 2$ with respect to standard discretization results that guarantee exponents of only 1 or 2. Then in Lemma 3 we show that there exists a finite-dimensional positive definite matrix $B \in \mathbb{S}_+(\mathbb{R}^n)$ such that $\mathrm{Tr}(B) \leq \mathrm{Tr}(A_*)$ and $\Phi_i^\top B \Phi_i = \langle \phi(x_i), A_*\phi(x_i)\rangle$ for all $i \in [n]$. Finally, in Theorem 5, we combine Lemma 3 with Theorem 4, to show that the problem in Eq. (2.4) provides a solution that is only $O(h^m_{\widehat{X},\Omega})$ distant from the solution of the infinite dimensional problem in Eq. (2.2).

To start we recall some basic properties of Φ_i and $\phi(x_i)$, for $i \in [n]$, already sketched in Sect. 2. In particular, the next proposition shows that, by construction, $\Phi_i^\top \Phi_j = \phi(x_i)^\top \phi(x_j)$ for any $i, j \in [n]$ and more generally that the map V that maps $f \in \mathcal{H} \mapsto R^{-\top}(\langle \phi(x_1), f\rangle, \ldots, \langle \phi(x_n), f\rangle) \in \mathbb{R}^n$ is a partial isometry and

that $\Phi_i = V\phi(x_i)$. The map V will be crucial to characterize the properties of the finite dimensional version of the operator A_*

Lemma 2 (Characterizing Φ_j in terms of ϕ). *Let k be a kernel satisfying Assumption 2(a). There exists a linear operator $V : \mathcal{H} \to \mathbb{R}^n$ such that*

$$\Phi_i = V\phi(x_i), \qquad \forall i \in [n].$$

Moreover V is a partial isometry: VV^* is the identity on $\mathbb{R}^n$, $P = V^*V$ is a rank n projection operator satisfying $P\phi(x_i) = \phi(x_i), \forall i \in [n]$.

The proof of Proposition 2 is given in Appendix C.1 and is based on the fact that the kernel matrix K is positive definite and invertible when k is *universal* [17], property that is implied by Assumption 2(a), and that R is an invertible matrix that satisfies $K = R^\top R$.

5.1 Uniform inequality from scattered constraints

In this section we derive Theorem 4. Here we want to guarantee that a function g satisfies $g(x) \geq -\varepsilon$ on Ω, by imposing some constraints on $g(x_i)$ for $i \in [n]$. If we use the most natural discretization, that consists in the constraints $g(x_i) \geq 0$, by Lipschitzianity of g we can guarantee only $\varepsilon = |g|_{\Omega,1} h_{\widehat{X},\Omega}$ (recall the definition of $|\cdot|_{\Omega,m}$ for $m \in \mathbb{N}$ from Eq. (3.1)). In the case of *equality constraints*, instead, standard results for *functions with scattered zeros* [18] (recalled in Appendix B) guarantee for all $x \in \Omega$

$$|u(x)| \leq \varepsilon, \quad \varepsilon = C h^m_{\widehat{X},\Omega} |u|_{\Omega,m},$$

when u is m-times differentiable and satisfies $u(x_i) = 0$ for any $i \in [n]$ (see [18, 20] or Theorem 11 for more details). Thus, in this case the discretization leverages the degree of smoothness of u, requiring much less points to achieve a given ε than in the inequality case.

The goal here is to derive a guarantee for *inequality constraints* that is as strong as the one for the equality constraints. In particular, given a function g defined on Ω and that satisfies $g(x_i) - \Phi_i B \Phi_i = 0$ on $\widehat{X}$, with $B \succeq 0$, we first derive a function u defined on the whole Ω and matching $g(x_i) - \Phi_i B \Phi_i$ on $\widehat{X}$. This is possible since we know that $\Phi_i = V\phi(x_i)$, by Proposition 2, then $u(x) = g(x) - \langle \phi(x), V^*BV\,\phi(x) \rangle$ satisfies $u(x_i) = g(x_i) - \Phi_i B \Phi_i$ for any $i \in [n]$. Finally, we apply the results for functions with scattered zeros on u. The desired result is obtained by noting that, since $\langle \phi(x), V^*BV\,\phi(x) \rangle \geq 0$ for any $x \in \Omega$, by construction, then for all $x \in \Omega$

$$\begin{aligned} -g(x) &\leq -g(x) + \left\langle \phi(x), V^*BV\,\phi(x) \right\rangle \leq |g(x) - \left\langle \phi(x), V^*BV\,\phi(x) \right\rangle| \\ &= |u(x)| \leq \varepsilon, \end{aligned}$$

i.e., $g(x) \geq -\varepsilon$ for all $x \in \Omega$ with $\varepsilon = Ch^m_{\widehat{X},\Omega}|u|_{\Omega,m}$. In the following theorem we provide a slightly more general result, that allows for $|g(x_i) - \Phi_i B \Phi_i| \leq \tau$ with $\tau \geq 0$.

Theorem 4 (Uniform inequality from scattered constraints). *Let Ω satisfy Assumptions 1(a) for some $r > 0$. Let k be a kernel satisfying Assumptions 2(a) and 2(d) for some $m \in \mathbb{N}_+$. Let $\widehat{X} = \{x_1, \ldots, x_n\} \subset \Omega$ with $n \in \mathbb{N}_+$ such that $h_{\widehat{X},\Omega} \leq r \min(1, \frac{1}{18(m-1)^2})$. Let $g \in C^m(\Omega)$ and assume there exists $B \in \mathbb{S}_+(\mathbb{R}^n)$ and $\tau \geq 0$ such that*

$$|g(x_i) - \Phi_i^\top B \Phi_i| \leq \tau, \quad \forall i \in [n], \tag{5.2}$$

where the Φ_i's are defined in Sect. 2. The following statement holds:

$$g(x) \geq -(\varepsilon + 2\tau) \ \ \forall x \in \Omega, \quad \text{where} \quad \varepsilon = Ch^m_{\widehat{X},\Omega}, \tag{5.3}$$

and $C = C_0(|g|_{\Omega,m} + \mathsf{MD}_m Tr(B))$ with $C_0 = 3\frac{\max(\sqrt{d}, 3\sqrt{2d}(m-1))^{2m}}{m!}$. The constants $m, \mathsf{M}, \mathsf{D}_m$, defined in Assumptions 2(a) and 2(d), do not depend on $n, \widehat{X}, h_{\widehat{X},\Omega}, B$ or g.

Proof Let the partial isometry $V : \mathcal{H} \to \mathbb{R}^n$ and the projection operator $P = V^*V$ be defined as in Proposition 2. Given $B \in \mathbb{S}_+(\mathbb{R}^n)$ satisfying Eq. (5.2), define the operator $A \in \mathbb{S}_+(\mathcal{H})$ as $A = V^*BV$ and the functions $u, r_A : \Omega \to \mathbb{R}$ as follows

$$r_A(x) = \langle \phi(x), A\phi(x) \rangle, \quad u(x) = g(x) - r_A(x), \quad \forall x \in \Omega.$$

Since $\Phi_i = V\phi(x_i)$ for all $i \in [n]$, then for all $i \in [n]$:

$$r_A(x_i) = \left\langle \phi(x_i), V^*BV\phi(x_i) \right\rangle = (V\phi(x_i))^\top B(V\phi(x_i)) = \Phi_i^\top B \Phi_i,$$

and hence $u(x_i) = g(x_i) - \Phi_i^\top B \Phi_i$. Thus, $|u(x_i)| \leq \tau$ for any $i \in [n]$. This allows to apply one of the classical results on functions with scattered zeros [18, 20] to bound $\sup_{x\in\Omega} |u(x)|$, which we derived again in Theorem 11 to obtain explicit constants. Since we have assumed $h_{\widehat{X},\Omega} \leq r/\max(1, 18(m-1)^2)$, applying Theorem 11, the following holds

$$\sup_{x\in\Omega} |u(x)| \leq 2\tau + \varepsilon, \qquad \varepsilon = c\ R_m(u)\ h^m_{\widehat{X},\Omega},$$

where $c = 3\max(1, 18(m-1)^2)^m$ and $R_m(v) = \sum_{|\alpha|=m} \frac{1}{\alpha!} \sup_{x\in\Omega} |\partial^\alpha v(x)|$ for any $v \in C^m(\Omega)$ using the multi-index notation (recalled in Sect. 3.1). Since $r_A(x) = \langle \phi(x), A\phi(x) \rangle \geq 0$ for any $x \in \Omega$ as $A \in \mathbb{S}_+(\mathcal{H})$, it holds :

$$g(x) \geq g(x) - r_A(x) = u(x) \geq -|u(x)| \geq -(2\tau + \varepsilon), \qquad \forall x \in \Omega. \tag{5.4}$$

The last step is bounding $R_m(u)$. Recall the definition of $|\cdot|_{\Omega,m}$ from Eq. (3.1). First, note that $A = V^*BV$ is finite rank (hence trace-class). Applying the cyclicity of the trace and the fact that VV^* is the identity on $\mathbb{R}^n$, it holds

$$\mathrm{Tr}(A) = \mathrm{Tr}(V^*BV) = \mathrm{Tr}(BVV^*) = \mathrm{Tr}(B).$$

Since k satisfies Assumption 2(a), by Lemma 9, $r_A \in \mathcal{H}$ and $\|r_A\|_{\mathcal{H}} \leq \mathsf{M}\mathrm{Tr}(A) = \mathsf{M}\mathrm{Tr}(B)$ where M is fixed in Assumption 2(a). Moreover, since the kernel k satisfies Assumption 2(d) with m and D_m, then $|v|_{\Omega,m} \leq \mathsf{D}_m\|v\|_{\mathcal{H}}$, for any $v \in \mathcal{H}$ as recalled in Remark 2. In particular, this implies $|r_A|_{\Omega,m} \leq \mathsf{D}_m\|r_A\|_{\mathcal{H}} \leq \mathsf{D}_m\mathsf{M}\mathrm{Tr}(B)$. To conclude, note that, by the multinomial theorem,

$$R_m(u) = \sum_{|\alpha|=m} \frac{1}{\alpha!} \sup_{x\in\Omega} \left|\partial^\alpha u(x)\right| \leq \sum_{|\alpha|=m} \frac{1}{\alpha!} |u|_{\Omega,m} = \frac{d^m}{m!} |u|_{\Omega,m}.$$

Since $|u|_{\Omega,m} \leq |g|_{\Omega,m} + |r_A|_{\Omega,m}$, combining all the previous bounds, it holds

$$\varepsilon \leq C_0 \left(|g|_{\Omega,m} + \mathsf{D}_m\mathsf{M}\mathrm{Tr}(B)\right) h^m_{\widehat{X},\Omega}, \qquad C_0 = 3\frac{d^m \max(1, 18(m-1)^2)^m}{m!}.$$

The proof is concluded by bounding ε in Eq. (5.4) with the inequality above. □

In the theorem above we used a domain satisfying Assumption 1(a) and a version of a bound for functions with scattered zeros (that we derived in Theorem 11 following the analysis in [18]), to have explicit and relatively small constants. However, by using different bounds for functions with scattered zeros, we can obtain the same result as Theorem 4, but with different assumptions on Ω (and different constants). For example, we can use Corollary 6.4 in [20] to obtain a result that holds for $\Omega = [-1, 1]^d$ or Theorem 11.32 with $p = q = \infty$, $m = 0$ in [18] to obtain a result that holds for Ω with locally Lipschitz-continuous boundary.

5.2 Convergence properties of the finite-dimensional problem

Now we use Theorem 4 to bound the error of Eq. (2.4). First, to apply Theorem 4 we need to prove the existence of at least one finite-dimensional $B \succeq 0$ that satisfies the constraints of Eq. (2.4) and such that the trace of B is independent of n and $h_{\widehat{X},\Omega}$. This is possible since we proved in Theorem 3 that there exists at least one finite rank operator A that solves Eq. (2.2) and thus satisfies its constraints, of which the ones in Eq. (2.4) constitute a subset. In the next lemma we construct $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$, such that $\langle\phi(x_i), A\phi(x_i)\rangle = \Phi_i^\top \overline{B}\Phi_i$. In particular, $\overline{B} = VA_*V^* = R^{-\top}CR^{-1}$, with $C_{i,j} = \langle\phi(x_i), A_*\phi(x_j)\rangle$ for $i, j \in [n]$, where A_* is one solution of Eq. (2.2) with minimum trace-norm, since the bound in Theorem 4 depends on the trace of the resulting matrix.

Lemma 3 *Let Ω be an open set and $\{x_1, \ldots, x_n\} \subset \Omega$ with $n \in \mathbb{N}_+$. Let $g : \Omega \to \mathbb{R}$ and k be a kernel on Ω. Denote by $\mathcal{H}$ the associated RKHS and by ϕ the associated*

canonical feature map. Let $A \in \mathbb{S}_+(\mathcal{H})$ *satisfy* $Tr(A) < \infty$ *and* $\langle \phi(x), A\phi(x)\rangle = g(x),\ x \in \Omega$. *Then there exists* $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$ *such that* $Tr(\overline{B}) \leq Tr(A)$ *and* $g(x_i) = \Phi_i^\top \overline{B}\Phi_i,\ \forall i \in [n]$.

Proof Let $V : \mathcal{H} \to \mathbb{R}^n$ be the partial isometry defined in Proposition 2 and $P = V^*V$ be the associated projection operator. Define $\overline{B} \in \mathbb{R}^{n\times n}$ as $\overline{B} = VAV^*$. Since by Proposition 2, $\Phi_i = V\phi(x_i)$ and P satisfies $P\phi(x_i) = \phi(x_i)$ for $i \in [n]$,

$$\begin{aligned}\Phi_i^\top \overline{B}\Phi_i &= (V\phi(x_i))^\top (VAV^*)(V\phi(x_i)) = \left\langle V^*V\phi(x_i), AV^*V\phi(x_i)\right\rangle \\ &= \left\langle P\phi(x_i), AP\phi(x_i)\right\rangle = \langle \phi(x_i), A\phi(x_i)\rangle \quad \forall i \in [n].\end{aligned}$$

Note that $\overline{B}$ satisfies: (a) $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$, by construction; (b) the requirement $\Phi_i^\top \overline{B}\Phi_i = g(x_i)$, indeed $\Phi_i^\top \overline{B}\Phi_i = \langle \phi(x_i), A\phi(x_i)\rangle$ and $\langle \phi(x), A\phi(x)\rangle = g(x)$ for any $x \in \Omega$; (c) $\mathrm{Tr}(\overline{B}) \leq \mathrm{Tr}(A)$, indeed, by the cyclicity of the trace,

$$\mathrm{Tr}(\overline{B}) = \mathrm{Tr}(VAV^*) = \mathrm{Tr}(AV^*V) = \mathrm{Tr}(AP).$$

The proof is concluded by noting that, since $A \succeq 0$ and $\|P\|_{\mathrm{op}} \leq 1$ because P is a projection, then $\mathrm{Tr}(AP) \leq \|P\|_{\mathrm{op}}\mathrm{Tr}(|A|) = \|P\|_{\mathrm{op}}\mathrm{Tr}(A) \leq \mathrm{Tr}(A)$. □

We are now ready to prove the convergence rates of Eq. (2.4) to the global minimum. We will use the bound for the inequality on scattered data that we derived Theorem 4 and the fact that there exists $\overline{B} \succeq 0$ that satisfies the constraints of Eq. (2.4) with a trace bounded by $\mathrm{Tr}(A_*)$ as we proved in the lemma above (that is in turn bounded by the the trace of the operator explicitly constructed in Theorem 2). The proof is organized as follows. We will first show that Eq. (2.4) admits a minimizer, that we denote by $(\hat{c}, \hat{B})$. The existence of $\overline{B}$ allows to derive a lower-bound on $\hat{c} - f_*$. Using Theorem 4 on the constraints of Eq. (2.4) and evaluating the resulting inequality in one minimizer ζ of f allows to find an upper bound on $\hat{c} - f_*$ and an upper bound for $\mathrm{Tr}(\hat{B})$.

Theorem 5 (Convergence rates of Eq. (2.4) to the global minimum). *Let* Ω *be a set satisfying Assumption 1(a) for some* $r > 0$. *Let* $n \in \mathbb{N}_+$ *and* $\widehat{X} = \{x_1, \ldots, x_n\} \subset \Omega$ *with fill distance* $h_{\widehat{X},\Omega}$. *Let* k *be a kernel and* $\mathcal{H}$ *the associated RKHS satisfying Assumption 2 for some* $m \in \mathbb{N}_+$. *Let* f *be a function satisfying Assumption 1(b) and Assumption 3 for* $\mathcal{H}$. *The problem in Eq. (2.4) admits a solution. Let* $(\hat{c}, \hat{B})$ *be any solution of Eq. (2.4), for a given* $\lambda > 0$. *The following holds*

$$|\hat{c} - f_*| \ \leq\ 2\eta\,|f|_{\Omega,m} \ +\ \lambda\, Tr(A_*), \qquad \eta \ =\ C_0\, h_{\widehat{X},\Omega}^m, \tag{5.5}$$

when $h_{\widehat{X},\Omega} \leq r\min(1, \frac{1}{18(m-1)^2})$ *and* $\lambda \geq 2\mathsf{M}\mathsf{D}_m\eta$. *Here* $C_0 = 3\frac{\max(\sqrt{d}, 3\sqrt{2d}(m-1))^{2m}}{m!}$, D_m, M *are defined in Assumption 2 and* A_* *is given by Theorem 3. Moreover, under the same conditions*

$$Tr(\hat{B}) \ \leq\ 2\,Tr(A_*) \ +\ 2\tfrac{\eta}{\lambda}\,|f|_{\Omega,m}. \tag{5.6}$$

Proof We divide the proof in few steps.
Step 0. Problem Eq. (2.4) **admits always a solution.**
(a) On the one hand, *c cannot be larger than* $c_0 = \min_{i\in[n]} f(x_i)$, otherwise there would be a point x_j for which $f(x_j)-c < 0$ and so the constraint $\Phi_j^\top B \Phi_j = f(x_j)-c$ would be violated, since does not exist any positive semi-definite matrix for which $\Phi_j^\top B \Phi_j < 0$.
(b) On the other, *there exists an admissible point.* Indeed let (c_*, A_*) be the solution of Eq. (2.2) such that A_* has minimum trace norm. By Theorem 3, we know that this solution exists with $c_* = f_*$, under Assumptions 1 to 3. Then, by Lemma 3 applied to $g(x) = f(x) - c_*$ and $A = A_*$, given $\widehat{X} = \{x_1, \dots, x_n\}$ we know that there exists $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$ satisfying $\mathrm{Tr}(\overline{B}) \leq \mathrm{Tr}(A_*)$ such that the constraints of Eq. (2.4) are satisfied for $c = c_*$. Then $(c_*, \overline{B})$ is admissible for the problem in Eq. (2.4).
Thus, since there exists an admissible point for the constraints of Eq. (2.4) and its functional cannot be larger than c_0 without violating one constraint, the SDP problem in Eq. (2.4) admits a solution (see [21]).
Step 1. Consequences of existence of A_*. Let $(\hat{c}, \hat{B})$ be one minimizer of Eq. (2.4). The existence of the admissible point $(c_*, \overline{B})$ proven in the step above implies that

$$\hat{c} - \lambda \mathrm{Tr}(\hat{B}) \geq c_* - \lambda \mathrm{Tr}(\overline{B}) \geq f_* - \lambda \mathrm{Tr}(A_*),$$

from which we derive,

$$\lambda \mathrm{Tr}(\hat{B}) - \lambda \mathrm{Tr}(A_*) \leq \Delta, \quad \Delta := \hat{c} - f_*. \tag{5.7}$$

Step 2. $\boldsymbol{f|_\Omega \in C^{m+2}(\Omega)}$. Assumption 3 guarantees that $f|_\Omega \in C^2(\Omega)$ and that for all $i, j \in [d]$, $\frac{\partial}{\partial x_i \partial x_j} f|_\Omega \in \mathcal{H}$. Since under Assumption 2(d), $\mathcal{H} \subset C^m(\Omega)$ by Remark 2, we see that $\frac{\partial}{\partial x_i \partial x_j} f|_\Omega \in C^m(\Omega)$ for all $i, j \in [d]$ and hence $f|_\Omega \in C^{m+2}(\Omega)$.
Step 3. $\boldsymbol{L^\infty}$ **bound due to the scattered zeros.** Let $(\hat{c}, \hat{B})$ be one minimizer of Eq. (2.4) and define $\hat{g}(x) = f(x) - \hat{c}$ for all $x \in \Omega$. Note that $\hat{g}(x_i) = \Phi_i^\top \hat{B} \Phi_i$ for $i \in [n]$. Moreover, $\hat{g} \in C^m(\Omega)$ because $f \in C^m(\Omega)$ and $\hat{c}$ is a constant. Considering that $h_{\widehat{X},\Omega} \leq \frac{r}{\max(1, 18(m-1)^2)}$, by assumption, then all the conditions in Theorem 4 are satisfied for $g = \hat{g}$, $\tau = 0$ and $B = \hat{B}$. Applying Theorem 4, we obtain,

$$\forall x \in \Omega, \ f(x) - \hat{c} = \hat{g}(x) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \mathrm{Tr}(\hat{B})), \qquad \eta = C_0 h_{\widehat{X},\Omega}^m,$$

where C_0 is defined in Theorem 4. Since the inequality above holds for any $x \in \Omega$, by evaluating it in one global minimizer $\zeta \in \Omega$, we have $f(\zeta) = f_*$ and hence

$$-\Delta = f_* - \hat{c} = f(\zeta) - \hat{c} = \hat{g}(\zeta) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \mathrm{Tr}(\hat{B})).$$

Since $\hat{g} = f - \hat{c}\mathbf{1}_\Omega$, and since for any $m \in \mathbb{N}_+$, $|\mathbf{1}_\Omega|_{\Omega,m} = 0$, we have $|\hat{g}|_{\Omega,m} \leq |f|_{\Omega,m} + |\mathbf{1}_\Omega|_{\Omega,m} = |f|_{\Omega,m}$. Injecting this in the previous equation yields

$$\Delta \leq \eta |f|_{\Omega,m} + \eta \mathsf{MD}_m \mathrm{Tr}(\hat{B}). \tag{5.8}$$

Conclusion. Combining Eq. (5.8) with Eq. (5.7), and since $\lambda \geq 2\mathsf{MD}_m\eta$ by assumption,

$$\tfrac{\lambda}{2}\mathrm{Tr}(\hat{B}) \leq (\lambda - \mathsf{MD}_m\eta)\mathrm{Tr}(\hat{B}) \leq \eta|f|_{\Omega,m} + \lambda\mathrm{Tr}(A_*).$$

Note that Eq. (5.6) is obtained from the one above, by dividing by $\frac{\lambda}{2}$. Finally the inequality Eq. (5.5) is derived by bounding Δ from below as $\Delta \geq -\lambda\mathrm{Tr}(A_*)$ by Eq. (5.7), since $\mathrm{Tr}(\hat{B}) \geq 0$ by construction, and bounding it from above as

$$\Delta \leq 2\eta|f|_{\Omega,m} + \lambda\mathrm{Tr}(A_*),$$

obtained by combining Eq. (5.8) with Eq. (5.6) and with the assumption $\mathsf{MD}_m\eta \leq \frac{\lambda}{2}$. □

The result above holds for any kernel satisfying Assumption 2 and any function f, Ω satisfying the geometric conditions in Assumption 1 and with $f \in C^2(\Omega)$ and $\frac{\partial^2 f}{\partial x_i \partial x_j} \in \mathcal{H}$ for $i, j \in [d]$. The latter requirement is quite easy to verify for example when $\mathcal{H}$ contains $C^s(\Omega)$ and $f \in C^{s+2}(\Omega)$ for some $s > 0$ as in the case of $\mathcal{H}$ being a Sobolev space with $s > d/2$. Moreover the proposed result holds for any discretization $\widehat{X}$ (random, or deterministic). We would like to conclude with the following remark on the sufficiency of the assumptions on f.

Remark 3 (Sufficiency of Assumptions 1(b) and 3). Assumptions Assumptions 1(b) and 3 are sufficient for Theorems 3 and 5 to hold. However, by inspecting their proof it is clear that they hold by requiring only the existence of a trace-class operator $A_* \in \mathbb{S}_+(\mathcal{H})$ such that $f(x) - f_* = \langle \phi(x), A_*\phi(x)\rangle$ for any $x \in \Omega$, where $f_* = \inf_{x\in\Omega} f(x)$. Note that this is implied by Assumptions 1(b) and 3 via Corollary 1.

In the next subsection we are going to apply the theorem above to the specific setting of Algorithm 1.

5.3 Result for Sobolev kernels and discussion

In this we are going to apply Theorem 5 to Algorithm 1 which corresponds to $\mathcal{H}$ be the Sobolev space of smoothness s and the points $\widehat{X}$ selected independently and uniformly at random. First, in the next lemma we bound in high probability the fill distance $h_{\widehat{X},\Omega}$ with respect to the number of points n that we sample, i.e., the cardinality of $\widehat{X}$.

Lemma 4 (Random sets of points). *Let $\Omega \subset \mathbb{R}^d$ be a bounded set with diameter $2R$, for some $R > 0$, and satisfying Assumption 1(a) for a given $r > 0$. Let $\widehat{X} = \{x_1, \ldots, x_n\}$ independent points sampled from the uniform distribution on Ω. When $n \geq 2(\frac{6R}{r})^d \left(\log\frac{2}{\delta} + 2d\log\frac{4R}{r}\right)$, then the following holds with probability at least $1-\delta$:*

$$h_{\widehat{X},\Omega} \quad \leq \quad 11R\, n^{-\frac{1}{d}} \left(\log\tfrac{n}{\delta} + d\log\tfrac{2R}{r}\right)^{1/d}.$$

The proof of Theorem 4 is in Appendix E.1 and is a simpler version (with explicit constants) of more general results [22, Thm. 13.7]. In the next theorem we apply the bound in the lemma above with the explicit constants for Sobolev spaces derived in Proposition 1 to Theorem 5. The derivation of the theorem below is in Appendix E.2.

Theorem 6 (Convergence rates of Algorithm 1 to the global minimum). *Let $\Omega \subset \mathbb{R}^d$ be a bounded set with diameter $2R$, for some $R > 0$, and satisfying Assumption 1(a) for a given $r \in (0, R]$ (e.g. if Ω is a ball with radius R, then $r = R$). Let s satisfying $s > d/2$. Let k be Sobolev kernel of smoothness s (see Example 1). Assume that f satisfies Assumption 1(b) and that $f|_\Omega \in W_2^{s+2}(\Omega)$. Let $\hat{c}$ be the result of Algorithm 1 executed with $n \in \mathbb{N}_+$ points chosen uniformly at random in Ω and $\lambda > 0$. Let $\delta \in (0, 1]$. When $m \in \mathbb{N}_+$ satisfies $m < s - d/2$ and $n \geq \max(4, 15(m-1))^{2d} \left(\frac{R}{r}\right)^d \left(2 \log \frac{2}{\delta} + 4d \log \frac{20R\, m}{r}\right)$ choose any λ satisfying*

$$\lambda \quad \geq \quad n^{-\frac{m}{d}} \quad \left(\log \frac{2^d n}{\delta}\right)^{\frac{m}{d}} \quad R^m C_{m,s,d},$$

where $C_{m,s,d} = 11^m C_0 \max(1, \mathsf{MD}_m)$ with C_0 defined in Theorem 5 and MD_m defined in Proposition 1. Note that $C_{m,s,d}$ is explicitely bounded in the proof in terms of s, m, d. Then, with probability at least $1 - \delta$, the following holds

$$|\hat{c} - f_*| \quad \leq \quad 3\,\lambda \, \left(Tr(A_*) + |f|_{\Omega,m}\right).$$

A direct consequence of the theorem above, already stated in Remark 1, is the nearly-optimality of Algorithm 1 for the cases of Sobolev functions. Indeed by applying Theorem 6 with m equal to the largest integer strictly smaller than $s - d/2$ we have that $m \geq s - d/2 - 1$, and so Algorithm 1 achieves the global minimum with a rate that is $O(n^{-\frac{s}{d}+\frac{1}{2}+\frac{1}{d}})$. The lower bounds from information based complexity state that, by observing the functions in n points, it is not possible to find the minimum with error smaller than $n^{-\frac{s}{d}+\frac{1}{2}}$ for functions in $W_2^s(\Omega)$ (see, e.g., [1], Prop. 1.3.11, page 36). Since in Theorem 6 we assume f belongs to $W_2^{s+2}(\Omega)$, the optimal rate would be $n^{-\frac{s}{d}+\frac{1}{2}-\frac{2}{d}}$ so we are a factor $n^{3/d}$ slower than the optimal rate. Note that this factor is negligible if the function is very smooth, i.e., $s \gg d$, or d is very large. An interesting corollary that corresponds to Theorem 1, can be derived considering that $C^{s+2}(\Omega) \subseteq W_2^{s+2}(\Omega)$, since Ω is bounded.

6 Algorithm

We need to solve the following optimization problem:

$$\max_{B \succcurlyeq 0, c \in \mathbb{R}} c - \lambda \mathrm{Tr}(B) \quad \text{such that} \quad f(x_i) - c - \Phi_i^\top B \Phi_i = 0, \quad \forall i \in [n].$$

This is a semi-definite programming problem with n constraints and a semi-definite constraint of size n. It can thus be solved with precistion ε in time $O(n^{3.5} \log(1/\varepsilon))$ and

memory $O(n^2)$ by standard software packages [21]. However, to allow applications to $n = 1000$ or more, and on parallel architectures, we provide a simple Newton algorithm, which relies on penalization by a self-concordant barrier, that is, we aim to solve

$$\max_{B \succcurlyeq 0, c \in \mathbb{R}} c - \lambda \mathrm{Tr}(B) + \frac{\varepsilon}{n} \log\det(B) \text{ such that } f(x_i) - c - \Phi_i^\top B \Phi_i = 0, \ \forall i \in [n],$$

for which we know that at optimum, the deviation with the optimal value is at most ε [23, Sec. 4.4]. By standard Lagrangian duality, we get, with $\Phi \in \mathbb{R}^{n \times n}$ the matrix with rows $\Phi_1, \ldots, \Phi_n$, so that $\Phi\Phi^\top = K$:

$$\begin{aligned}
&\sup_{B \succcurlyeq 0, c} \inf_{\alpha \in \mathbb{R}^n} c + \sum_{i=1}^n \alpha_i \big(f(x_i) - c - \Phi_i^\top B \Phi_i\big) - \lambda \mathrm{Tr}(B) + \frac{\varepsilon}{n} \log\det(B) \\
&= \inf_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i f(x_i) - \frac{\varepsilon}{n} \log\det\big(\Phi^\top \mathrm{Diag}(\alpha) \Phi + \lambda I\big) + \frac{\varepsilon}{n} \log \frac{\varepsilon}{n} - \varepsilon \text{ s. t. } \alpha^\top 1_n = 1.
\end{aligned}$$

With the barrier term, this thus defines a dual function $H(\alpha)$, and we get the following gradient

$$H'(\alpha)_i = f_i - \frac{\varepsilon}{n} \Phi_i^\top \Big(\Phi^\top \mathrm{Diag}(\alpha) \Phi + \lambda I\Big)^{-1} \Phi_i = f_i - \frac{\varepsilon}{n\alpha_i} \Big[K(K + \lambda \mathrm{Diag}(\alpha)^{-1})^{-1}\Big]_{ii},$$

and Hessian

$$H''(\alpha)_{ij} = \frac{\varepsilon}{n} \Big[\Phi_i^\top \Big(\Phi^\top \mathrm{Diag}(\alpha) \Phi + \lambda I\Big)^{-1} \Phi_j\Big]^2,$$

which can be rewritten

$$H''(\alpha)_{ij} = \frac{\varepsilon}{n\alpha_j\alpha_i} \Big[K(K + \lambda \mathrm{Diag}(\alpha)^{-1})^{-1}\Big]_{ij} \Big[K(K + \lambda \mathrm{Diag}(\alpha)^{-1})^{-1}\Big]_{ji}.$$

We can then compute the step for the Damped Newton algorithm: $\alpha^+ = \alpha - \frac{1}{1 + \sqrt{\frac{n}{\varepsilon}} \lambda(\alpha)} \Delta$, where $\Delta = H''(\alpha)^{-1} H'(\alpha) - \frac{1_n^\top H''(\alpha)^{-1} H'(\alpha)}{1_n^\top H''(\alpha)^{-1} 1_n} H''(\alpha)^{-1} 1_n$ and $\lambda(\alpha)^2 = \Delta^\top H''(\alpha) \Delta$ is the Newton decrement (which can serve as a stopping criterion). Note that the algorithm is always feasible, without a need for any eigenvalue decomposition. The overall complexity is $O(n^3)$ per iteration due to matrix inversions and linear systems. Note that the conditioning of these linear systems is at least as bad as the conditioning of the kernel matrix K. Fortunately, for the s-th Sobolev kernels in dimension d, the m-th eigenvalue of the kernel matrix typically decay as $m^{-2s/d}$ [24, Sec. 2.3].

Retrieving c and B. From an optimal α, we can recover $B = \frac{\varepsilon}{n} \big(\Phi^\top \mathrm{Diag}(\alpha) \Phi + \lambda I\big)^{-1} = \frac{\varepsilon}{n\lambda} \big(I - \Phi^\top (\Phi\Phi^\top + \lambda \mathrm{Diag}(\alpha)^{-1})^{-1} \Phi\big)$ and $c = \frac{1}{n} H'(\alpha)^\top 1_n$ (since c is the

Lagrange multiplier for the constraint $\alpha^\top 1_n = 1$). Thus, computing the model for a test point, can be done as $\frac{\varepsilon}{n\lambda}\big(k(x,x) - q(x)^\top (K + \lambda \mathrm{Diag}(\alpha)^{-1})^{-1} q(x)\big)$, where $q(x)_i = k(x, x_i)$. Alternatively, when Φ is invertible, we can use $q(x)^\top \Phi^{-\top} B \Phi^{-1} q(x)$.

Retrieving a minimizer. Given the dual solution, based on our localizing arguments presented in Sect. 7, a good candidate solution will be

$$\hat{z} = \sum_{i=1}^{n} \alpha_i x_i \tag{6.1}$$

A more principled way to find a minimizer is provided in Sect. 7, of which the equation above corresponds to the limit solution of Eq. (7.4) for $\nu \to 0$ (see Sect. 7.1).

Number of iterations. In order to reach a Newton decrement $n^{1/2}\varepsilon^{-1/2}\lambda(\alpha) \le \kappa$, a number of steps equal to a universal constant times $\frac{n}{\varepsilon}[H(\alpha_0) - H(\alpha_*)] + \log\log\frac{1}{\kappa}$ is sufficient. [23].

When initializing with $\alpha_0 = \frac{1}{n}1_n$, we have $H(\alpha_0) = \frac{1}{n}\sum_{i=1}^n f_i - \frac{\varepsilon}{n}\log\det\big(K + n\lambda I\big) + \frac{\varepsilon}{n}\log\varepsilon - \varepsilon$, and $H(\alpha_*) \ge c_* - \lambda \mathrm{Tr}(A_*) - \varepsilon$. This leads to a number of Newton steps less than

$$\frac{n}{\varepsilon}\big[\langle f\rangle - \inf f\big] + \log\det\big(K + n\lambda I\big) + \frac{n}{\varepsilon}\lambda\mathrm{Tr}(A_*) + \log\varepsilon + \log\log\frac{1}{\kappa}.$$

In our experiments, we do not perform path following (that would lead the classical interior-point method) and instead fixed value $\varepsilon = 10^{-3}$, and a few hundred Newton steps.

Behavior for $\lambda = 0$. If the kernel matrix K is invertible (which is the case almost surely for Sobolev kernels and points sampled independently from a distribution with a density with respect to the Lebesgue measure), then we show that for $\lambda = 0$, the optimal value of of the finite-dimensional problem in Eq. (2.4) is equal to $\min_{i\in[n]} f(x_i)$. Since $f(x_i) \ge c + \Phi_i^\top B \Phi_i$ implies $f(x_i) \ge c$, the optimal value has to be less than $\min_{i\in[n]} f(x_i)$. We therefore just need to find a feasible B that achieves it. Since K is assumed invertible (and thus its Cholesky factor as well), we can simply take $B = R^{-\top}\mathrm{Diag}[(f(x_j) - \min_{i\in[n]} f(x_i))_j]R^{-1}$.

7 Finding the global minimizer

In this section we provide and study the problem in Eq. (7.4), that is a variation of the problem in Eq. (2.4), and allows to find also the minimizer of f as we prove in Theorem 8. As in Sect. 2 we start from a convex representation of the optimization problem and then we derive our sampled version, passing by an intermediate infinite-

dimensional problem that is useful to derive the theoretical properties of the method. While the problem in Eq. (2.1) can be seen as finding the largest constant c such that $f - c$ is still non-negative, in the problem below we find the parabola of the form $p_{z,\gamma}(x) = \frac{\nu}{2}\|x\|^2 - \nu x^\top z + c = \frac{\nu}{2}\|x - z\|^2 + c - \frac{\nu}{2}\|z\|^2$ with the highest vertex such that $f - p_{z,c}$ is still non-negative. Since the height of the vertex of $p_{z,c}$ corresponds to $c - \frac{\nu}{2}\|z\|^2$, the resulting optimization problem is the following,

$$\max_{c\in\mathbb{R}, z\in\mathbb{R}^d} c - \frac{\nu}{2}\|z\|^2 \quad \text{such that} \quad f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top z - c \geq 0 \quad \forall x \in \Omega. \tag{7.1}$$

It is easy to see that if $f \in C^2(\mathbb{R}^d)$ has a unique minimizer ζ that belongs to Ω and is locally strongly convex around ζ then there exists a $\nu > 0$ such that the problem above achieves an optimum (c_*, z_*) with $z_* = \zeta$ and $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$. In particular, to characterize ν explicitly we introduce the stronger assumption below.

Assumption 4 (Geometric assumption to find global minimizer). *The function $f : \mathbb{R}^d \to \mathbb{R}$ has a unique global minimizer in Ω.*

If f satisfies Assumptions 1(b) and 4, denote with ζ the unique minimizer of f in Ω and with $f_* = f(\zeta)$ the corresponding minimum.

Remark 4 Under Assumptions 1(b) and 4 f can be lower bounded by a parabola with value f_* at ζ, i.e., there exists $\beta > 0$ such that

$$\forall x \in \Omega,\ f(x) - f_* \geq \frac{\beta}{2}\|x - \zeta\|^2. \tag{7.2}$$

The remark above is derived in Appendix F.1. In what follows, whenever f satisfies Assumptions 1(b) and 4, then β will be assumed to be the supremum among the value satisfying Eq. (7.2). Now we are ready to summarize the reasoning above on the fact that Eq. (7.1) achieves the minimizer of f.

Lemma 5 *Suppose f satisfies Assumptions 1 and 4. Let ζ be the unique minimizer of f in Ω and $f_* = f(\zeta)$ be the corresponding minimum. Let $\beta > 0$ such that Eq. (7.2) holds. If $\nu < \beta$ then the problem in Eq. (7.1) has a unique solution (c_*, z_*) such that $z_* = \zeta$ and $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$.*

The lemma above guarantees that the problem in Eq. (7.1) achieves the global minimum and the global minimizer of f, when f satisfies the geometric conditions Assumptions 1 and 4. Now, as we did for Eq. (2.1), we consider the following problem of which Eq. (7.1) is a tight relaxation.

$$\begin{aligned} \max_{c\in\mathbb{R}, z\in\mathbb{R}^d, A\in\mathbb{S}_+(\mathcal{H})} & \ c - \frac{\nu}{2}\|z\|^2 \\ \text{such that} & \ f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top z - c \ = \langle \phi(x), A\phi(x)\rangle \quad \forall x \in \Omega. \end{aligned} \tag{7.3}$$

Indeed, since $\langle \phi(x), A\phi(x)\rangle \geq 0$ for any $x \in \Omega$ and $A \in \mathbb{S}_+(\mathcal{H})$, for any triplet (c, z, A) satisfying the constraints in the problem above, the couple (c, z) satisfies the

constraints in Eq. (7.1). The contrary may be not true in general. In the next theorem we prove that when $\mathcal{H}$ satisfies Assumption 2 and Ω, f satisfy Assumptions 1, 3 and 4, then the relaxation is tight and, in particular, when $\nu < \beta$, there exists a finite rank operator A_* such that the triplet $(f_* + \frac{\nu}{2}\|\zeta\|^2, \zeta, A_*)$ is optimal.

Theorem 7 *Let $\Omega \subset \mathbb{R}^d$ be an open set, k be a kernel, $\mathcal{H}$ the associated RKHS, and $f : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumptions 1 to 3, and Assumption 4. Let β satisfying Eq. (7.2). For any $\nu < \beta$, the problem in Eq. (7.3) admits an optimal solution (c_*, z_*, A_*) with $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$, $z_* = \zeta$, and A_* a positive semi-definite operator on $\mathcal{H}$ with rank at most $d + 1$.*

The proof of the theorem above is essentially the same of Theorem 3 and is reported for completeness in Appendix F.2. In particular, to prove the existence of A_* we applied Corollary 1 to the function $f(x) - \frac{\nu}{2}\|x - \zeta\|^2$ that still satisfies Assumptions 1 and 3 when f does and $\nu < \beta$. Now we are ready to consider the finite-dimensional version of Eq. (7.3). Given a set of points $\widehat{X} = \{x_1, \ldots, x_n\}$ with $n \in \mathbb{N}_+$,

$$\begin{aligned} \max_{c\in\mathbb{R}, z\in\mathbb{R}^d, B\in\mathbb{S}_+(\mathbb{R}^n)} \; & c - \tfrac{\nu}{2}\|z\|^2 - \lambda \mathrm{Tr}(B) \\ \text{such that} \quad & \forall i \in [n],\ f(x_i) - \tfrac{\nu}{2}\|x_i\|^2 + \nu x_i^\top z - c = \Phi_i^\top B \Phi_i. \end{aligned} \tag{7.4}$$

For the problem above we can derive similar convergence guarantees as for Eq. (2.4) and also a convergence of the estimated minimizer z to ζ, as reported in the following theorem.

Theorem 8 (Convergence rates of Eq. (7.4) to the global minimizer). *Let Ω be a set satisfying Assumption 1(a) for some $r > 0$. Let $\widehat{X} = \{x_1, \ldots, x_n\} \subset \Omega$ with fill distance $h_{\widehat{X},\Omega}$. Let k be a kernel satisfying Assumption 2 for some $m \geq 2$ and f satisfying Assumptions 1, 3 and 4. The problem in Eq. (7.4) admits a solution. Denote by $(\hat{c}, \hat{z}, \hat{B})$ any solution of Eq. (7.4), for a given $\lambda > 0$. Then*

$$\tfrac{\nu}{2}\|\hat{z} - \zeta\|^2 \leq 3\eta(|f|_{\Omega,m} + \nu) + 2\lambda Tr(A_*), \qquad \eta = C\, h_{\widehat{X},\Omega}^m, \tag{7.5}$$

when $h_{\widehat{X},\Omega} \leq \frac{r}{18(m-1)^2}$ and $\lambda \geq 2\mathsf{M}\mathsf{D}_m\eta$. Here $C = 3\frac{(3\sqrt{2d}(m-1))^{2m}}{m!}$ and D_m, M are defined in Assumption 2. A_* is from Theorem 7. Moreover under the same conditions

$$|\hat{c} - \tfrac{\nu}{2}\|\hat{z}\|^2 - f_*| \leq 2\eta\,|f|_{\Omega,m} + \lambda\, Tr(A_*) + 2\eta\nu, \tag{7.6}$$

$$Tr(\hat{B}) \leq 2\, Tr(A_*) + 2\tfrac{\eta}{\lambda}\,|f|_{\Omega,m} + 2\nu\tfrac{\eta}{\lambda}. \tag{7.7}$$

The proof of the theorem above is similar to the one of Theorem 5 and it is stated for completeness in Appendix F.3. The same comments to Theorem 5 that we reported in the related section and the rates for Sobolev functions, apply also in this case. In the next section we describe the algorithm to solve the problem in Eq. (7.4).

7.1 Algorithm

We can use the same dual technique as presented in Sect. 6, and obtain a dual problem to Eq. (7.4) with the additional penalty $\frac{\varepsilon}{n} \log\det B$. The dual problem can readily be obtained as (up to constants)

$$\inf_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i f(x_i) - \frac{\varepsilon}{n} \log\det\left(\Phi^\top \mathrm{Diag}(\alpha)\Phi + \lambda I\right) + \frac{\nu}{2}\left(-\sum_{i=1}^n \alpha_i \|x_i\|_2^2 + \left\|\sum_{i=1}^n \alpha_i x_i\right\|_2^2\right),$$

such that $\alpha^\top 1_n = 1$, with the optimal z that can be recovered as $z = \sum_{i=1}^n \alpha_i x_i$. We note that when ν tends to zero, we recover the dual problem from Sect. 6, and we keep the candidate above in Ω even when $\nu = 0$.

7.2 Warm restart scheme for linear rates

It is worth noting that Theorem 8 provides strong guarantees on the distance $\|\hat{z} - \zeta\|$ where $\hat{z}$ is the solution of the problem Eq. (7.4) and ζ the global optimum of f. This suggests that we can implement a warm restart scheme that leverage the additional knowledge of the position of ζ. Assume indeed that Ω is a ball of radius R centered in z_0. For $t = 1, \dots, T$ with $T = \lceil \log \frac{1}{\varepsilon} \rceil$, we apply Eq. (7.4) to a set $\widehat{X}_t$ that contains enough points sampled uniformly at random in the ball $B_{r_{t-1}}(z_{t-1})$ such that Theorem 8 guarantees that $\|z_t - \zeta\| \le r_{t-1}/e$ where z_t is the solution of Eq. (7.4). The cycle is repeated with $r_t = r_{t-1}/e$ and the new center be z_t. By plugging the estimate of Theorem 4 for $h_{\widehat{X}_t, B_{r_{t-1}}(z_{t-1})}$ in Theorem 8 for each step t, we obtain a total number of points n to achieve $\|z_T - \zeta\| \le \varepsilon$ with probability $1 - T\delta$, that is

$$n = O\left(C_{d,m}^{d/m} \left(\frac{\mathcal{F}}{\nu}\right)^{d/m} R^d \log\frac{1}{\varepsilon}\right)$$

modulo logarithmic terms in n and δ, where $C_{d,m} = 3^m C \mathsf{MD}_m$ with C defined in Theorem 8 and $\mathcal{F} = |f|_{\Omega,m} + \nu + \mathrm{Tr}(A_*)$. This means that under the additional assumption of a unique minimizer in Ω, we achieve a convergence rate that is only logarithmic in ε, moreover when $m \gg d$ also the dependence with respect to $C_{d,m}$ (which is exponential in m and d in the case of the Sobolev kernel) and $\mathcal{F}$ improves, since d/m tends to 0.

8 Extensions

In this section we deal with two aspects: (a) the effect of solving approximately the problem in Eq. (2.4), and (b) how can we certify explicitly (no dependence on quantities of theoretical interest as $\mathrm{Tr}(A_*)$) how close is a given (approximate) solution to the optimum;

8.1 Approximate solutions

In this section we extend Theorem 5 to consider the case when we solve Eq. (2.4) in an approximate way. In particular, let $\lambda > 0$, $n \in \mathbb{N}_+$ and $\widehat{X} = \{x_1, \ldots, x_n\}$. Denote by $p_{\lambda,n}$ the optimal value achieved by Eq. (2.4) for such λ, n. We say that $(\tilde{c}, \tilde{B})$ is an *approximate solution* of Eq. (2.4) with parameters $\theta_1, \theta_2, \tau_1, \tau_2 \geq 0$ if it satisfies the following inequalities

$$p_{\lambda,n} \; - \; \tilde{c} + \lambda \mathrm{Tr}(\tilde{B}) \leq \theta_1 + \theta_2 \mathrm{Tr}(\tilde{B}), \tag{8.1}$$

$$|f(x_i) - \tilde{c} - \Phi_i^\top \tilde{B} \Phi_i| \leq \tau_1 + \tau_2 \mathrm{Tr}(\tilde{B}), \quad \forall i \in [n]. \tag{8.2}$$

Theorem 9 (Error of approximate solutions of Eq. (2.4)). *Let $(\tilde{c}, \tilde{B})$ be an approximate solution of Eq. (2.4) for a given $n \in \mathbb{N}_+, \lambda > 0$ as defined in Eqs. (8.1) and (8.2) w.r.t. $\tau_1, \tau_2, \theta_1, \theta_2 \geq 0$. Under the same assumptions and notation of Theorem 5 and Remark 3, when $\tau_2, \theta_2 \leq \frac{\lambda}{8}$*

$$|\tilde{c} - f_*| \; \leq \; 7(2\tau_1 + \eta |f|_{\Omega,m}) + 6(\theta_1 + \lambda Tr(A_*)), \tag{8.3}$$

$$Tr(\tilde{B}) \; \leq \; 8\, Tr(A_*) \; + \; 8\tfrac{\eta}{\lambda}\, |f|_{\Omega,m} \; + \; 8\tfrac{\theta_1 + 2\tau_1}{\lambda}. \tag{8.4}$$

The proof of the theorem above is reported for completeness in Appendix G.1, and is a variation of the one of Theorem 5 where we used Theorem 4 with $\tau = \tau_1 + \tau_2 \mathrm{Tr}(\tilde{B})$ and we further bound $p_{\lambda,n}$ via Eq. (8.1). From a practical side, the theorem above allows to use a wide range of methods and techniques to approximate the solution of Eq. (2.4). In particular, it is possible to use lower dimensional approximations of $\Phi_1, \ldots, \Phi_n$ and algorithms based on early stopping as described in Sect. 11, since $\tau_1, \tau_2, \theta_1, \theta_2$ will take into account the error incurred in the approximations. An interesting application of the theorem above, from a theoretical side is that it allows also to deal with situations where f does not have a representer A_* in $\mathbb{S}_+(\mathcal{H})$ as we are going to discuss in the next section.

8.2 Certificate of optimality

While in Theorem 5 we provide a bound on the convergence of Eq. (2.4) *a priori*, i.e., only depending on properties of $f, \Omega, \mathcal{H}$, in this section we provide a bound *a posteriori*, that is a *certificate of optimality*. Indeed, the next theorem quantifies $f(z) - f^*$ for a candidate minimizer z, in terms of only $(\hat{c}, \hat{B})$, an (approximate) solution of Eq. (2.4) and $|f|_{\Omega,m}$. A candidate minimizer based on Eq. (2.4) is provided in Eq. (6.1). In Sect. 7 we study a different algorithm Eq. (7.4) that explicitly provides a minimizer and whose certificate is studied in Appendix G.3.

Theorem 10 (Certificate of optimality a minimizer from Eq. (2.4)). *Let Ω satisfy Assumption 1(a) for some $r > 0$. Let k be a kernel satisfying Assumptions 2(a) and 2(d) for some $m \in \mathbb{N}_+$. Let $\widehat{X} = \{x_1, \ldots, x_n\} \subset \Omega$ with $n \in \mathbb{N}_+$ such that $h_{\widehat{X},\Omega} \leq$*

$\frac{r}{18(m-1)^2}$. *Let* $f \in C^m(\Omega)$ *and let* $\hat{c} \in \mathbb{R}$, $\hat{B} \in \mathbb{S}_+(\mathbb{R}^n)$ *and* $\tau \geq 0$ *satisfying*

$$|f(x_i) - \hat{c} - \Phi_i^\top \hat{B} \Phi_i| \leq \tau, \quad i \in [n], \tag{8.5}$$

where the Φ_i*'s are defined in Sect. 2. Let* $f_* = \min_{x \in \Omega} f(x)$. *Then the following holds*

$$|f(z) - f_*| \leq f(z) - \hat{c} + \varepsilon + 2\tau, \quad \forall z \in \Omega, \quad \text{where} \quad \varepsilon = C h^m_{\widehat{X},\Omega}, \tag{8.6}$$

and $C = C_0(|f|_{\Omega,m} + \mathsf{M}\mathsf{D}_m Tr(\hat{B}))$. *The constants* C_0, *defined in Theorem 4, and* $m, \mathsf{M}, \mathsf{D}_m$, *defined in Assumptions 2(a) and 2(d), do not depend on* $n, \widehat{X}, h_{\widehat{X},\Omega}, \hat{c}, \hat{B}$ *or* f.

Proof By applying Theorem 4 with $g(x) = f(x) - \hat{c}$, we have $f(x) - \hat{c} \geq -\varepsilon - 2\tau$ for any $x \in \Omega$. In particular this implies that $f(\zeta) - \hat{c} \geq -\varepsilon - \tau$. The proof is concluded by noting that $f(z) \geq f_*$ by definition of f_*. □

9 Relationship with polynomial hierarchies

The formulation as an infinite-dimensional sum-of-squares bears some strong similarities with polynomial hierarchies. There are several such hierarchies allowing to solve any polynomial optimization problem [6, 25, 26], but one has a clear relationship to ours. The goal of the following discussion is to shed light on the benefits in terms of condition number and dimensionality of the problem, deriving by using an infinite dimensional feature map in the finite dimensional problem, instead of an explicit finite-dimensional polynomial map as in the case considered by the papers cited above.

Adding small perturbations. We start this discussion from the following result from Lasserre [25], that is, for any multivariate non-negative polynomial f on $\mathbb{R}^d$, and for any $\eta > 0$, there exists a degree $r(f, \eta)$ such that the function

$$f_\eta(x) = f(x) + \eta \sum_{k=0}^{r(f,\varepsilon)} \frac{1}{k!} \sum_{j=1}^{d} x_j^{2k}$$

is a sum of squares, and such that the ℓ_1-norm between the coefficients of f and f_η tends to zero (here this ℓ_1-norm is equal to $\eta d \sum_{k=0}^{r(f,\varepsilon)} \frac{1}{k!} \leq \eta d e$).

This implies that for the kernel $k_r(x, y) = \sum_{k=0}^{r} \frac{(x^\top y)^k}{k!}$, with feature map $\phi_r(x)$ composed of all weighted monomials of degree less than r, the function

$$f(x) + \eta \|\phi_r(x)\|_2^2 = f(x) + \eta k_r(x, x)$$

is a sum of squares, for any $r \geq r(f, \eta)$, with η arbitrarily close to zero (this can be obtained by adding the required squares to go from $\sum_{j=1}^{d} x_j^{2k}$ to $\|x\|^{2k} =$

$(\sum_{j=1}^{d} x_j^2)^k)$. This result implies that minimizing f arbitrarily precisely over any compact set K (such that $\sup_{x\in K} k_r(x,x)$ is finite), can be done by minimizing $f(x) + \eta k(x,x)$, with sum-of-squares polynomials of sufficiently large degree. We already showed that in this paper that if f satisfies the geometric condition in Assumption 1(b), our framework is able to find the global minimum by the finite dimensional problem in Eq. (2.4), which, in turn, is based on a kernel associated to an infinite dimensional space (as the Sobolev kernel, see Example 1). We now show how our framework can provide approximation guarantees and potentially efficient algorithms for the problem above even when Assumption 1(b) may not hold and we use a polynomial kernel of degree r (with r that may not be large enough). However, in this case the resulting problem would suffer of a possibly infinite condition number and a larger dimensionality than the one achievable with an infinite dimensional feature map.

Modified optimization problem. Given the representation of $x \mapsto f(x) - f_* + \eta\|\phi_r(x)\|_2^2$ as a sum-of-squares, we can explicitly model the function as

$$f(x) - c + \eta\|\phi_r(x)\|_2^2 = \langle \phi_r(x), A\phi_r(x)\rangle$$

with A positive definite and $\eta \geq 0$. Note that if r is greater than twice the degree of f this problem is always feasible by taking η sufficiently large. Moreover, for feasible (c, η, A), we have for any $x \in \Omega$,

$$f(x) \geq c - \eta\|\phi_r(x)\|^2 \geq c - \eta \sup_{y\in\Omega} \|\phi_r(y)\|_2^2.$$

Thus, a relaxation of the optimization problem is

$$\begin{aligned}\sup_{c\in\mathbb{R}, A\succcurlyeq 0, \eta\geq 0} \quad & c - \eta \sup_{y\in\Omega} \|\phi_r(y)\|_2^2 \quad \text{s. t.} \quad \forall x \in \Omega, \ f(x) \\ & = c + \phi_r(x)^\top A\phi_r(x) - \eta\|\phi(x)\|_2^2.\end{aligned}$$

Moreover, if we choose r larger than $r(f - f_*, \eta)$, we know that there exists a feasible A which is positive semi-definite, with $c = f_* - \eta \sup_{y\in\Omega} \|\phi_r(y)\|_2^2$, and thus the objective value is greater than $f_* - \eta \sup_{y\in\Omega} \|\phi_r(y)\|_2^2$. Thus, the objective value of the problem above converges to f_*, when η go to zero (and thus $r(f - f_*, \eta)$ goes to infinity), while always providing a lower bound. Note that if $f - f_*$ is a sum of squares, then the optimal value η can be taken to be zero, and we recover the initial problem.

Subsampling and regularization. At this point, since r is finite, subsampling $\binom{d}{2r}$ points leads to an equivalent finite-dimensional problem. We can also add some regularization to sub-sample the problem and avoiding such a large number of points. Note here that the kernel matrix will probably be ill-conditioned, and the problem computationally harder to solve and difficult ro regularize.

Infinite-degree polynomials. In the approach outlined above, we need to let r increase to converge to the optimal value. We can directly take $r = \infty$, since $k_r(x, y) = \sum_{k=0}^{r} \frac{(x^\top y)^k}{k!}$ tends to the kernel $\exp(x^\top y)$, and here use subsampling. Again, it may lead to numerical difficulties. However, we can use Sobolev kernels (with guarantees on performance and controlled conditioning of kernel matrices), on the function $f(x) + \eta e^{\|x\|_2^2}$ for which we now there exists a sum of squares representation as soon as f is a polynomial.

10 Experiments

In this section, we illustrate our results with experiments on synthetic data.

Finding hyperparameters. Given a function to minimize and a chosen kernel, there are three types of hyperparameters: (a) the number n of sample points, (b) the regularization parameter λ, and (c) the kernel parameters. Since n drives the running time complexity of the method, we will always set it manually, while we will estimate the other parameters (regularization and kernel), by "cross-validation" (i.e., selecting the parameters of the algorithm that lead to the minimum value of f at the candidate optimum, among a logarithmic range of parameters). This adds a few function evaluations, but allows to choose good parameters.

Functions to minimize. We consider first a simple functions defined in $\mathbb{R}^2$ with their global minimimizer on $[-1, 1]^d$, which is minus the sum of Gaussian bumps (see Fig. 1). To go to higher even dimensions with the possibility of computing the global minimum with high precision by grid search, we consider functions of the form $f(x) = f(x_1, x_2) + f(x_3, x_4) + \cdots + f(x_{d-1}, x_d)$. We also consider adding a high-frequency cosine on the coordinate directions representing a more general scenario for a non-convex function. Note that in this second setting the gradient based methods cannot work properly (while ours can) as we are going to see in the simulations.

All results are reported by normalizing function values so that the range of values is 1, that is, $\max_{x\in[-1,1]^d} f(x) = 1$ and $\min_{x\in[-1,1]^d} f(x) = 0$.

Baseline algorithms. We compare our algorithm with the exponential kernel and points sampled from a quasi-random sequence in $[-1, 1]^d$, such as the Halton sequence [27], to:

- Random search: select a quasi-random sequence in $[-1, 1]^d$ and take the point with minimal function value.
- Random search with gradient descent: starting gradient descent for a certain number of iterations from quasi-random points, with a number of initialization divided

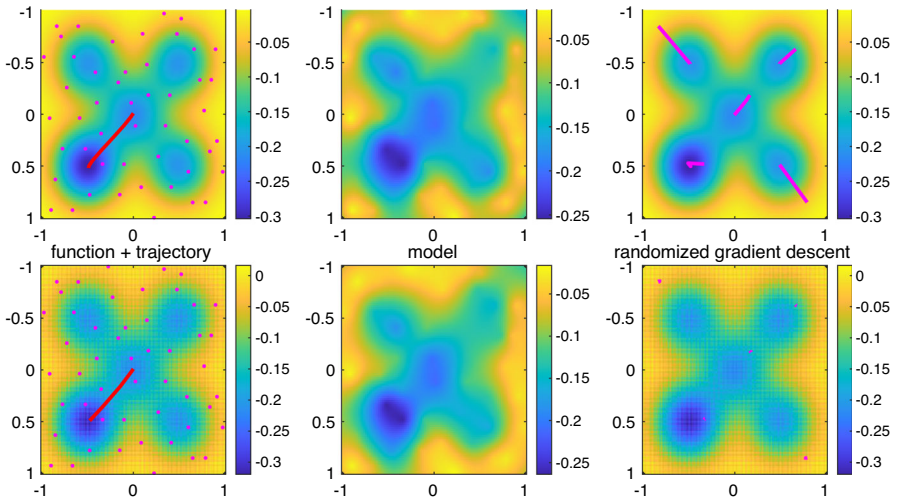

Fig. 1 Top: 2D function without small-amplitude high-frequency components. Bottom: 2D function with small-amplitude high-frequency components. Left: sampled points and the trajectory of the proposed algorithm. Center: model reconstructed by the algorithm (see Eq. (10.1)). Right: the trajectory of gradient descent starting from random points. As it is possible to see, even a small local non-convexity prevents the random+GD algorithms to converge properly, while the proposed method is quite robust to it

by $d + 1$ and the number of gradient steps, to account for gradient evaluations based on $d + 1$ function evaluations (by finite-difference). The step-size for gradient descent is taken constant, but its values is optimized for smallest final value while providing a descent algorithm.

Illustration in two dimensions. We show in Fig. 1 a function in two dimensions, with sampled point in purple, the trajectory of the candidate optimum along Newton iterations in red, and the final model of the function. We also compare to gradient descent with random starting points. We consider two functions below, one without extra high-frequency component (top), and one with (bottom). We can make the following observations:

- Our algorithm outperforms random search, that is, it improves on the function values of the sampled points.
- For the smoother function, gradient descent performs quite well, but is not robust when high-frequency components are added.

Note that the proposed algorithm provides also a model of the function reconstructed starting from its evaluation on the sampled points. In particular, if $(\hat{c}, \hat{B})$ is a solution of the algorithm, the approximate function $\hat{g} \approx f - f^*$ corresponds to

$$\hat{g}(x) = \langle \phi(x), V^* B V \phi(x) \rangle = v(x)^\top R^{-1} \hat{B} R^{-\top} v(x), \quad \forall x \in \Omega \tag{10.1}$$

with $v(x) = (k(x_1, x), \ldots, k(x_n, x))$ for $x \in \Omega$ and where $V : \mathcal{H} \to \mathbb{R}^n$ is in Sect. 5.

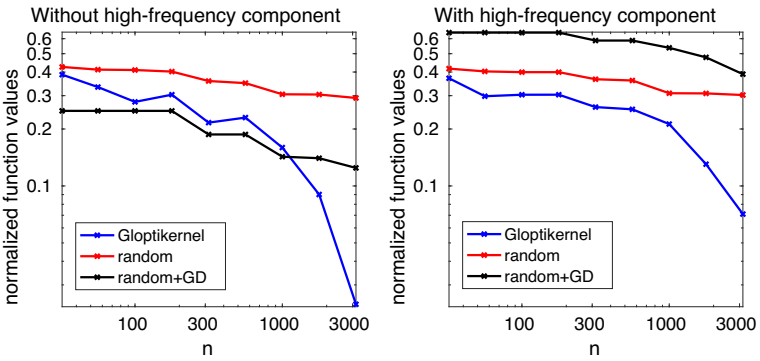

Fig. 2 Multivariate case $d = 8$. Minimization error of our algorithm (gloptikernel) compared with random evaluations or random evaluations + GD. The function considered is built as described at the beginning of this section with domain $[-1, 1]^d$ and shifted and rescaled to have minimum in 0 and output in $[0, 1]$. Left: function without small-amplitude high-frequency components. Right: function with small-amplitude high-frequency components

Higher dimensions. We compare the algorithms on a problem in dimension $d = 8$, as n increases, in order to assess how we approach the global optimum. We perform 4 replications with different random seeds for the sampling of points in $[-1, 1]^d$. The function to be minimized is built as described at the beginning of this section and is shifted and rescaled to have output in $[0, 1]$ and the minimum in 0. We can see that as n gets large, the performance of the proposed algorithm improves, and that with high frequency components, gradient descent with random restarts has worse performance and seem to show a slower rate overall, even in the case of the function without high-frequency components (Fig. 2).

10.1 Experiments on benchmarks for global optimization

In this section, we perform experiments using the algorithm described in the section above on the more than 200 global optimization problems in multiple dimensions constituting the well-known benchmark "Global Optimization Benchmarks" [28–30] http://infinity77.net/global_optimization/index.html. The functions to optimize come with their minimizer and their minimum to be used as a ground truth and with a region of interest where to look for the minimizers.

In this section, we present only the results for dimensions 4 and higher, as our method seems particularly interesting for these dimensions. The results for dimensions 2 and 3 can be found in Table 4, in the "Appendix". In Table 1, we report the results obtained by our algorithm. The algorithm we implemented is warm-restart scheme described in Sect. 7.2. The implementation details are reported in Appendix H. The algorithm was performed with $N = 200$ restart iterations, and was repeated 5 times (we select the best estimator out of the 5 restarts, to take into account of the high-probability factors). In Table 1, we report the following : (a) the problem name; (b) its

Table 1 Results of our algorithm for functions in dimension greater than four

	d	iters thresh	Final absolute error	fevs/iter
Colville	4	32	1.87E−03	31
Corana	4	1	0.00E+00	31
Shekel07	4	20	9.54E−07	31
PowerSum	4	2	3.26E−04	31
Ratkowsky01	4	90	3.69E+02	31
MieleCantrell	4	3	9.03E−13	31
Powell	4	6	2.85E−07	31
Shekel10	4	18	0.00E+00	31
Shekel05	4	18	1.91E−06	31
BiggsExp04	4	12	7.88E−05	31
Gear	4	2	1.18E−09	31
Kowalik	4	15	4.87E−05	31
DeVilliersGlasser01	4	4	1.06E+03	31
DeVilliersGlasser02	5	NaN	2.28E+03	36
Dolan	5	2	3.78E−13	36
BiggsExp05	5	3	2.64E−03	36
Trid	6	10	0.00E+00	41
Watson	6	11	1.09E−03	41
Hartmann6	6	8	0.00E+00	41
LennardJones	6	2	0.00E+00	41
Thurber	7	125	9.70E+03	46
Xor	9	NaN	6.99E−03	56
Paviani	10	23	1.03E−04	61
Cola	17	68	3.35E−01	96

dimension; (c) the number of iterations needed to achieve a threshold of 0.01 relative error; we define the relative error as $r(x) = \frac{f(x)-f(x_*)}{f(x_1)-f(x_*)}$; (d) the final absolute error $f(\widehat{x}) - f(x_*)$; (e) the number m of new function evaluations at each step (without counting those in order to select λ).

Note that the dimension of the optimization problem is $n = 3m$ and that the SDP constraint is also of size $n \times n$. Moreover, the choice we make to evaluate the relative error is in order to avoid very high values of the function f; comparing to $f(x_1)$ somewhat shows the importance of the iterative scheme.

Discussion, interpretation The set of functions on which we have tested our algorithm originally is a challenge allowing a maximum of 2000 function evaluations to reach a precision in absolute error of order 10^{-6}. We do not try to compete in this specific challenge, which models the fact that the number of function evaluations in certain real-life problems is very costly. In order to tackle this challenge, we would

Table 2 Polynomials used in the experiments

Poly1	$4x_1^2 + x_1x_2 - 4x_2^2 - 2.1x_1^4 + 4x_2^4 + x_1^6/3$
Poly2	$x_1^2x_4x_6x_7 + 4x_1x_2^2x_6x_8 + x_1x_2x_3x_4x_6 - x_2^4x_7 + 3x_2x_4^3x_7 + 3x_3x_4x_5x_6x_8$ $+x_3x_5x_7^2x_8 + \sum_{i=1}^{8} x_i^6$
Poly3	$-9x_2^2 + 8x_3x_7 + 2x_1x_4x_5 + 3x_3x_5x_6 + x_1^4 + x_2^4 + x_3^4 + x_4^4 + x_5^4$ $+x_6^4 + x_7^4 + x_1^6 + x_2x_3^5$
Poly4	$-15x_6 - 2x_1x_7^2 - 3x_2^2x_4 - x_3^2x_4 + x_1^4 + x_2^4 + x_3^4 + x_4^4 + x_5^4 + x_6^4 + x_7^4$
Poly5	$2x_5x_8 + 4x_1x_8x_9 + 4x_4x_6x_9 + x_1^4 + x_2^4 + x_3^4 + x_4^4 + x_5^4$ $+x_6^4 + x_7^4 + x_8^4 + x_9^4 + x_{10}^4$
Poly6	$-9x_2x_7x_{10} - 2x_3x_{11}x_{13} + 5x_5x_7x_{15} - 3x_9x_{11}x_{15} + \sum_{i=1}^{15} x_i^4$
Poly7	$8x_2x_8x_{11} + 3x_2x_{14}x_{15} - 5x_4x_7x_{13} - 13x_{12}^2x_{17} + \sum_{i=1}^{17} x_i^4$
Poly8	$-11x_2x_6x_{11} - 4x_3x_4x_{11} + 3x_4x_{10}x_{11} - x_5x_8x_{10} + \sum_{i=1}^{12} x_i^4$
Poly9	$12x_2x_4x_5x_8 + 5x_1x_2x_4x_5x_7 + 5x_2x_3x_4^2x_7 + x_1^6 + x_2^6 + x_3^6 + x_4^6 + x_5^6$ $+x_6^6 + x_7^6 + x_8^6 + x_9^6$

need to reduce the cost in function evaluations of certain steps such as that of the selection of λ (which we believe can be done without much difficulty).

Note that the fact that we achieve a relative error of 0.01 in almost all cases shows that the iterative scheme is indeed effective.

The performance on certain problems is bad, but this seems to be linked to the fact that the functions at hand have very high oscillations (hence high derivatives).

Remark 5 (NaN values). NaN values simply mean that we never reach a relative precision of 0.01.

10.2 Comparison with SOS polynomials

In this section, we present a second set of experiments with the same setting as before but optimizing polynomial functions.

One of the reference algorithms in order to optimize polynomials (on semi-algebraic sets) is the Lasserre Hierarchy, implemented in the toolbox `gloptipoly 3` [31]. Applying this toolbox on a minimization problem constrained on a hyper-rectangle will yield either a lower bound (if the hierarchy does not converge) or the exact minimum as well as a minimizer if the hierarchy does converge.

The idea of this section is not to compete with the Lasserre hierarchies, which are tailored for polynomials. Rather, we wish to compare both methods, and show that they can complement each other by providing an approximation of the minimizer with a certificate (i.e. with an upper bound on its distance to the optimum). This is particularly interesting in high dimensions, or with polynomials with high degree: in that case, the size of the polynomial problem becomes intractable, while our algorithm still runs and returns a solution.

In this experiment, we consider polynomials whose expression can be found below, and wish to find a minimizer for these polynomials in the hypercube $[-2, 2]^d$. Note that this domain is chosen such that we can easily sample from it (while Lasserre hierarchies can adapt to much more flexible sets of constraints). In particular, when using the `gloptipoly 3` we imposed the additional constraint $\|x\|^2 \leq 4$, that, while redundant, improves the convergence behavior of the algorithm, as suggested by an anonymous reviewer.

Selection of polynomials Almost all of the polynomials considered in these experiments are of the form

$$P(x) = \sum_{i=1}^{d} x_i^{2k} + Q(x), \deg(Q) \leq 2k - 1.$$

We randomly select a few non-zero indices for the Q as well as a random integer. The exact expressions of all the polynomials used can be found below in Table 2.

Results and Discussion. We report the results of these experiments in Table 3. The following columns are reported: (a) the name of the polynomial function; (b) the dimension d of the underlying domain; (c) the degree deg of the polynomial function; (d) wether or not gloptipoly3 has converged (cv column); (e) the relaxation order we have tested for before computational issues (relax column); (f) the dimensions of the PSD constraints for the Lasserre hierarchy (PSD moment matrix + the ones due to the constraints); (g) the gap between our solution and the Lasserre lower bound; (h) the dimension of the PSD constraint in our method.

Our method is statistical and therefore does not enjoy the same precision that `gloptipoly3` achieves in the case when it converges. However, our method is clearly more scalable in the sense that it returns an approximate solution for polynomials of high degree and dimension, for any chosen dimension of the PSD matrix, and very small matrices allow to achieve already interesting precision, as it is possible to see in Table 3.

11 Discussion

In this section, we discuss our results and propose a series of extensions.

Main technical contribution and extensions. We see that from Eq. (2.1), the problem of minimization can be easily written in terms of an infinite set of inequality constraints on $u(x) = f(x) - c$ that must hold for every $x \in \Omega$. While it is well known how to approximate efficiently an infinite set of equality constraints via a finite subset (e.g. via *bounds on functions with scattered zeros* [18] from the field of approximation theory), leading to optimal rates for the approximation problem, the situation is more difficult in the case of an infinite set of inequality constraints. The main technical contribution

Table 3 Results of the experiments when using both `gloptipoly3` and our method

	d	deg	cv	relax	PSD dim glopti	gap	PSD dim (ours)
Poly1	2	6	True	3	$10 \times 10 + 4 \times (6 \times 6)$	0.0000E+00	63×63
Poly3	7	6	True	4	$330 \times 330 + 14 \times (120 \times 120)$	5.0819E−02	138×138
Poly4	7	4	True	4	$330 \times 330 + 14 \times (120 \times 120)$	1.9073E−05	138×138
Poly2	8	6	True	3	$165 \times 165 + 16 \times (45 \times 45)$	1.7166E−04	153×153
Poly9	9	6	True	3	$455 \times 455 + 20 \times (91 \times 91)$	5.1392E−02	168×168
Poly5	10	4	True	3	$286 \times 286 + 20 \times (66 \times 66)$	2.4796E−05	183×183
Poly8	12	4	True	2	$91 \times 91 + 24 \times (13 \times 13)$	3.6388E−02	213×213
Poly6	15	4	True	2	$136 \times 136 + 30 \times (16 \times 16)$	2.2190E−02	258×258
Poly7	17	4	True	2	$171 \times 171 + 34 \times (18 \times 18)$	1.6233E+00	288×288

of this paper, on which the whole result of the paper is based, is Theorem 4, which allows to deal with an infinite set of inequality constraints as efficiently as in the equality case as discussed in Sect. 5.1. In particular, we rewrite the infinite set of inequalities $g(x) \geq 0,\ \forall x \in \Omega$ in terms of a very sparse set of constraints of the form $g(x_i) = \Phi_i B \Phi_i$, for some points $x_1, \ldots, x_n \in \Omega$ and a matrix $B \in \mathbb{S}_+(\mathbb{R}^n)$, with n in the same order of the one required by the equality case. Assume for simplicity that Ω is contained in the unit ball and the points are uniformly distributed in Ω. From Theorem 4 we derive that if B exists,

$$g(x) \ \geq -C\, n^{-m/d}\ (|g|_{\Omega,m} + \mathrm{Tr}(B)),$$

modulo logarithmic factors, where m is the order of smoothness of g. This result is particularly useful for two reasons. First, it recovers the same dependence on m, the smoothness of g, and n the number of sample points, as in the case of equality constraints. This is particularly convenient when $m \gg d$, e.g. with $m \geq d$ the rate becomes $O(n^{-1})$, that is independent from d in the exponent (the dependence of d is still present in the hidden constants and it is exponential in the worst case). Second, if used in an optimization problem, the matrix B can be found via a convex formulation, by requiring $u(x_i) = \Phi_i^\top B \Phi_i$ for $i \in [n]$ and penalizing $\mathrm{Tr}(B)$ in the functional. This technique allows, for example, to deal with more general optimization problems with infinite constraints than the one considered in this paper, as

$$\min_{\theta \in \Theta} F(\theta) \quad \text{such that} \quad g(\theta, x) \geq 0,\ \forall x \in \Omega,$$

by translating it as follows

$$\min_{\theta \in \Theta,\, B \succeq 0} F(\theta) + \lambda \mathrm{Tr}(B) \quad \text{such that} \quad g(\theta, x_i) = \Phi_i B \Phi_i\ \forall i \in [n].$$

If F and u are convex in θ and Θ a convex set, then the second is a convex problem that has the potential to approximate very efficiently the first, due to Theorem 4. From this viewpoint this paper is an application of this principle to Eq. (2.1).

Duality. Beyond using duality in Sect. 6 for algorithmic purposes, there is also a dual for the infinite-dimensional problem, which can be written as,

$$\inf_{p:\Omega \to \mathbb{R}} \int_\Omega p(x) f(x) dx \text{ such that } \int_\Omega p(x) dx = 1 \text{ and } \int_\Omega p(x) \phi(x) \otimes \phi(x) dx \succcurlyeq 0.$$

Replacing the constraint $\int_\Omega p(x)\phi(x)\otimes\phi(x)dx \succcurlyeq 0$ by $\forall x \in \Omega,\ p(x) \geq 0$ leads to the usual relaxation of optimization with probability measures. Thus, like for polynomial optimization [32], our formulation corresponds also to a relaxation in the dual formulation to signed measures.

Comparison with algorithms based on SOS polynomials. Our approach is related to the field of optimization via polynomial sum-of-squares [6, 33]. Indeed we also transform the problem of non-convex optimization to an SDP problem and we use a sum-of-squares representation. Our analysis however takes a different path, indeed (1) we select a given function space and define the infinite-dimensional cone of sum-of-squares of functions belonging to it (2) then we derive sufficient conditions to guarantee that a non-negative function belongs to such cone (Theorem 2) (3) at this point the quantitative approximation results are built naturally in a modular way on top of the approximation results of the original function class and can leverage the ample literature of approximation theory, see Theorem 4 (here, in particular we use the results related to scattered data approximation [18], but we could have used other approximation results, e.g. the one based on wavelet or Fourier series approximation), from which we quantify explicitly the error of the optimization algorithm in Theorem 5.

According to recent results on SOS polynomials (see [34] and references therein) which apply to polynomial relaxations as described in Sect. 9, when f is a polynomial, such algorithms can achieve the global mininum with a rate $O(1/r^2)$ via an SDP problem based on the representation of SOS polynomials of degree r in terms of positive definite matrices. Since the dimension of the corresponding matrix is $n = \binom{d+r}{r}$ corresponding to $n = O(r^d)$, by expressing the rate with respect to the dimensionality of the matrix, such methods achieve the global minimum with an error that is in the order of $O(n^{-2/d})$. This can be compared with the approach proposed in this paper as Algorithm 1. By sampling n points from the domain of interest, we cast an SDP problem in terms of a n-dimensional positive definite matrix, achieving a rate that is $C_{s,d} n^{-s/d+1/2}$ (see Theorem 6) modulo logaritmic factors, by using a Sobolev kernel k_{s+3} with $s > d/2$ (see Example 1). Since the polynomials are arbitrarily differentiable, we can choose s arbitrarily large at the cost of a larger constant $C_{s,d}$ completely characterized in Theorem 6. For example, by choosing $s = 5d/2$ we achieve the global minimum with a rate $O(n^{-2})$ that does not suffer of the curse of dimensionality except in the constants, and that is faster than the one obtained by SOS polynomial methods especially when $d \gg 1$. It must be noted that our result holds under the sufficient assumption Assumption 1(b) that can be relaxed according to Remark 3, but that it is not required by SOS polynomial methods. It would be of interest to know if such methods can achieve our rates under the same assumption. The difference [33]

Comparison with simpler algorithms. Similar reasoning can be done with respect to simple algorithms for global optimization. We consider here the algorithm that consists in sampling n points at random in Ω and taking the one with minimum value. A simple analysis, that we report below shows that this method achieves a rate of $O(n^{-2/d})$. So our method is strictly better than taking the minimum $f(x_i)$ for $i \in [n]$ when f is at least 3-times differentiable. Note that even in the case when the function f is infinitely differentiable, the algorithm that consists in sampling n points at random in Ω and taking the one with minimum value cannot go faster that $O(n^{-2/d})$. To see this, consider $\Omega = [0, 1]^d$ and the points x_i to be chosen as a grid of step τ. This means that $n = O(\tau^{-d})$. Now let $f(x) = \|x - y\|^2$ for some $y \in [0, 1]^d$. This function is

infintely differentiable. Nevertheless, in general the best approximation of y on the grid will be $\tilde{y} = \tau \lfloor y/\tau + 1/2 \rfloor$ (componentwise). Since, for any τ, there exists always an $y \in [0, 1]^d$ such that $\lfloor y/\tau + 1/2 \rfloor - y/\tau = 1/2$, we have that in the worst case

$$f(\tilde{y}) - f(y) = \|\tilde{y} - y\|^2 = \tau^2 \|\lfloor y/\tau + 1/2 \rfloor - y/\tau\|^2 = \tau^2/4.$$

Now if we express τ w.r.t., n, i.e., $\tau = n^{-1/d}$, we see that we obtain an error that is in the order of $n^{-2/d}$. So this simple algorithm cannot be better than $n^{-2/d}$ even if the function is infinitely differentiable. A similar argument can be obtained when the points are a generic covering of Ω.

Obtaining optimal rates. Our current analysis, even for functions f in Sobolev spaces, does not lead to the optimal rate of convergence (we obtain an extra term of $2/d$ in the exponents). We conjecture, that this could be removed by a more refined analysis (in particular in the construction of the operator A_*).

Modelling gradients. Our current framework only used function values. If gradients are observed, it could be possible to use them to reduce the number of sampled points, using tools from [35].

The choice of Ω. Since we assume that f has at least one global minimum, then there exists always an open set Ω that contains it and that satisfies the required properties. In this work, we don't discuss how to find Ω. While, in general, this could be not an easy problem. In practice, many non-convex optimization problems come already with a region of interest where to look for the global minimum. Such a region is typically obtained by considering some basic properties of the function of interest. For example, if are minimizing a polynomial of the form $f(x) = B(x) + p(x)$, where $B(x) = x_1^{2r} + \cdots + x_d^{2r}$ for some $r \in \mathbb{N}$ and $p(x)$ is a polynomial of degree $q \leq 2r - 1$. Note that by construction f admits a global minimum, since it goes to $+\infty$ at infinity and has $p(0) < \infty$ (while any polynomial without this structure does not have a minimizer). Now it is possible to easily derive a hypercube that contains the global minimum. Indeed by construction $f(x_*) \leq f(0) = p(0)$. Denote by L the sum of the absolute values of the coefficients of p. Now take the smallest $R \geq 1$ such that $R^{2r} - LR^q \geq p(0) + \varepsilon$ for an $\varepsilon > 0$. For any $x \notin (-R, R)^d$, we have

$$f(x) = B(x) + p(x) \geq R^{2r} - LR^q \quad > \quad p(0) \geq f(x_*).$$

Then the region $(-R, R)^d$ contains all the global minimizers.

Efficient kernel approximations. The current algorithm has a complexity of $O(n^3)$ for n sampled points, partly due to the need to compute inverse of kernel matrices. There is a large literature within machine learning aiming at providing low-rank

approximations, either from approximations of K from a subset of its columns (see, e.g., [36, 37] and references therein) or using random feature vectors (see, e.g., [38, 39] and references therein). This requires to relax the equality constraint on the subset $\widehat{X}$ to an mean square deviations, as allowed by Sect. 8.

Constrained optimization. Following [6], we can apply the same algorithmic technique to constrained optimization, by formulating the problem of minimizing $f(x)$ such that $g(x) \geq 0$ as maximizing c such that $f(x) = c + p(x) + g(x)q(x)$, and p, q non-negative functions. We can then replace the non-negative constraints by $p(x) = \langle \phi(x), A\phi(x) \rangle$ and $q(x) = \langle \phi(x), B\phi(x) \rangle$ for positive operators A and B. We can then subsample and penalize the traces of A and B to obtain an algorithm. A detailed study of the approximation properties of this algorithm remains to be done.

Acknowledgements We would like to thank Jean-Bernard Lasserre and Edouard Pauwels for their feedback on an earlier version of the manuscript. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We acknowledge support from the European Research Council (grant SEQUOIA 724063). We also acknowledge support from the European Research Council (grant REAL 947908).

A Additional notation and definitions

We provide here some basic notation that will be used in the rest of the appendices.

Multi-index notation. Let $\alpha \in \mathbb{N}^d$, $x \in \mathbb{R}^d$ and f be an infinitely differentiable function on $\mathbb{R}^d$, we introduce the following notation

$$|\alpha| = \sum_{j \in [d]} \alpha_i, \quad \alpha! = \prod_{j \in [d]} \alpha_j !, \quad x^\alpha = \prod_{j \in [d]} x_j^{\alpha_j}, \quad \partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Some useful space of functions. Let Ω be an open set. In this paper we will denote by $C^s(\Omega)$, $s \in \mathbb{N}$, the set of s-times differentiable functions on Ω and by $C_0^s(\Omega)$ the set of functions that are differentiable at least s times and that are supported on a compact in Ω. Denote by $L^p(\Omega)$ the *Lebesgue space* of p-integrable functions with respect to the Lebesgue measure and denote by $\| \cdot \|_{L^p(\Omega)}$ the associated norm [11].

A.1 Fourier Transform

Given two functions $f, g : \Omega \to \mathbb{R}$ on some set Ω, we denote by $f \cdot g$ the function corresponding to *pointwise product* of f, g, i.e.,

$$(f \cdot g)(x) = f(x)g(x), \quad \forall x \in \Omega.$$

Let $f, g \in L^1(\mathbb{R}^d)$ we denote the *convolution* by $f \star g$

$$(f \star g)(x) = \int_{\mathbb{R}^d} f(y) g(x - y) dy.$$

Let $f \in L^1(\mathbb{R}^d)$. The Fourier transform of f is denoted by $\tilde{f}$ and is defined as

$$\tilde{f}(\omega) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-i\,\omega^\top x}\, f(x)\, dx,$$

We now recall some basic properties, that will be used in the rest of the appendix.

Proposition 2 (Basic properties of the Fourier transform [18], Chapter 5.2.).

(a) *Let $f \in L^1(\mathbb{R}^d)$ and let $r > 0$. Denote by $\tilde{f}$ its Fourier transform and by f_r the function $f_r(x) = f(x/r)$ for all $x \in \mathbb{R}^d$, then*

$$\tilde{f}_r(\omega) = r^d \tilde{f}(r\omega).$$

(b) *Let $f, g \in L^1(\mathbb{R}^d)$, then*

$$\widetilde{f \cdot g} = (2\pi)^{d/2} \tilde{f} \star \tilde{g}.$$

(c) *Let $\alpha \in \mathbb{N}_0^d$, $f : \mathbb{R}^d \to \mathbb{R}$ and $f, \partial^\alpha f \in L^1(\mathbb{R}^d)$, then*

$$\widetilde{\partial^\alpha f}\,(\omega) = i^{|\alpha|} \omega^\alpha \tilde{f}(\omega), \quad \forall \omega \in \mathbb{R}^d.$$

(d) *Let $f \in L^1(\mathbb{R}^d)$, then*

$$\|\tilde{f}\|_{L^\infty(\mathbb{R}^d)} \leq (2\pi)^{-d/2} \|f\|_{L^1(\mathbb{R}^d)}.$$

(e) *Let $f \in L^1(\mathbb{R}^d)$ and assume that $\tilde{f} \in L^1(\mathbb{R}^d)$, then*

$$f(x) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{i\,\omega^\top x}\, \tilde{f}(\omega)\, dx, \quad \text{and} \quad \|f\|_{L^\infty(\mathbb{R}^d)} \leq (2\pi)^{-d/2} \|\tilde{f}\|_{L^1(\mathbb{R}^d)}.$$

(f) *There exists a linear isometry $\mathcal{F} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ satisfying*

$$\mathcal{F} f = \tilde{f}, \quad f \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d).$$

The isometry is uniquely determined by the property in the equation above. For any $f \in L^2(\mathbb{R}^d)$ we denote by $\tilde{f}$ the function $\tilde{f} = \mathcal{F} f$.

A.2 Sobolev Spaces

For this section we refer to [11]. For any $\alpha \in \mathbb{N}_0^d$ we say that $v_\alpha \in L^1_{loc}(\mathbb{R}^d)$ is the *α-weak derivative* of $u \in L^1_{loc}(\mathbb{R}^d)$ if, for all compactly supported smooth functions $\tau \in C_0^\infty(\mathbb{R}^d)$, we have

$$\int_{\mathbb{R}^d} v_\alpha(x)\tau(x)dx = (-1)^{|\alpha|} \int_{\mathbb{R}^d} u(x)(\partial^\alpha \tau)(x)dx,$$

and we denote v_α by $D^\alpha u$. Let $\Omega \subseteq \mathbb{R}^d$ be an open set. For $s \in \mathbb{N}$, $p \in [1, \infty]$ the Sobolev spaces $W_p^s(\Omega)$ are defined as

$$W_p^s(\Omega) = \{f \in L^p(\Omega) \mid \|f\|_{W_p^s(\Omega)} < \infty\}, \quad \|f\|_{W_p^s(\Omega)} = \sum_{|\alpha| \le s} \|D^\alpha f\|_{L^p(\Omega)}.$$

We now recall some basic results about Sobolev spaces that are useful for the proofs in this paper. First we start by recalling the restriction properties of Sobolev spaces. Let $\Omega \subseteq \Omega' \subseteq \mathbb{R}^d$ be two open sets. Let $\beta \in \mathbb{N}$ and $p \in [1, \infty]$. By definition of the Sobolev norm above we have

$$\|g|_\Omega\|_{W_p^s(\Omega)} \le \|g\|_{W_p^s(\Omega')},$$

and so $g|_\Omega \in W_p^s(\Omega)$ for any $g \in W_p^s(\Omega')$. Now we recall the extension properties of Sobolev spaces.

Proposition 3 (Extension operator, 5.24 in [11]). *Let Ω be a bounded open subset of $\mathbb{R}^d$ with locally Lipschitz boundary* [11]. *Let $\beta \in \mathbb{N}$ and $p \in [1, \infty]$. There exists a bounded operator $E : W_p^\beta(\Omega) \to W_p^\beta(\mathbb{R}^d)$ and a constants C_3 depending only on β, p, Ω such that for any $h \in W_p^\beta(\Omega)$ the following holds (a) $h = (Eh)|_\Omega$ (b) $\|Eh\|_{W_p^\beta(\mathbb{R}^d)} \le C_3 \|h\|_{W_p^\beta(\Omega)}$ with $C_3 = \|E\|_{op}$.*

A.3 Reproducing Kernel Hilbert spaces

For this section we refer to [15–17]. Let S be a set and $k : S \times S \to \mathbb{R}$ be a p.d. kernel. We denote by $\mathcal{H}_k(S)$ the reproducing kernel Hilbert space (RKHS) associated to the kernel k, and by $\langle \cdot, \cdot \rangle_k$ the associated inner product. In particular, we will omit the dependence in k from $\mathcal{H}$ and $\langle \cdot, \cdot \rangle$ when the used kernel is clear from the context. We will omit also the dependence on S when $S = \Omega$, the region we are using in this paper. In particular we will use the following shortcuts $\mathcal{H} = \mathcal{H}_k(\Omega)$ and $\mathcal{H}(\mathbb{R}^d) = \mathcal{H}_k(\mathbb{R}^d)$.

Concrete constructions and useful characterizations. In the rest of the section we provide other methods to build RKHS and some interesting characterizations of $\mathcal{H}_k(S)$ and $\langle \cdot, \cdot \rangle_k$ that will be useful int the rest of the appendix.

Proposition 4 (Construction of RKHS given S, ϕ, Thm. 4.21 of *[17]*). *Let $\phi : S \to V$ be a continuous map, where V is separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$. Let $k(x, x') = \langle \phi(x), \phi(x') \rangle_V$ for any $x, x' \in S$. Then k is a p.d. kernel and the associated RKHS is characterized as follows*

$$\mathcal{H}_k(S) = \{\langle w, \phi(\cdot) \rangle_V \mid w \in V\}, \quad \|f\|_{\mathcal{H}_k(S)} = \inf_{u \in V} \|u\|_V \ \ s.t. \ \ f = \langle u, \phi(\cdot) \rangle_V .$$

Proposition 5 (Restriction of a RKHS $\mathcal{H}_{k_1}(S_1)$ on a subset $S_0 \subset S_1$ [15, 16]). *Let k_0 be the restriction on S_0 of the kernel k_1 defined on S_1. Then the following holds*

(a) *k_0 is a p.d. kernel,*
(b) *the RKHS $\mathcal{H}_{k_0}(S_0)$ is characterized as $\mathcal{H}_{k_0}(S_0) = \{f|_{S_0} \mid f \in \mathcal{H}_{k_1}(S_1)\}$,*
(c) *the norm $\|\cdot\|_{\mathcal{H}_{k_0}(S_0)}$ is characterized by*

$$\|f\|_{\mathcal{H}_{k_0}(S_0)} = \inf_{g \in \mathcal{H}_{k_1}(S_1)} \|g\|_{\mathcal{H}_{k_1}(S_1)}, \quad s.t. \quad f(x) = g(x) \ \forall x \in S_0,$$

(d) *there exist a linear bounded* extension *operator $E : \mathcal{H}_{k_0}(S_0) \to \mathcal{H}_{k_1}(S_1)$ such that $(Ef)(x) = f(x)$ for any $x \in S_0$ and $f \in \mathcal{H}_{k_0}(S_0)$ and such that*

$$\|f\|_{\mathcal{H}_{k_0}(S_0)} = \|Ef\|_{\mathcal{H}_{k_1}(S_1)}, \quad \forall f \in \mathcal{H}_{k_0}(S_0),$$

(e) *there exist a linear bounded* restriction *operator $R : \mathcal{H}_{k_1}(S_1) \to \mathcal{H}_{k_0}(S_0)$ such that $(Rf)(x) = f(x)$ for any $x \in S_0$ and $f \in \mathcal{H}_{k_1}(S_1)$,*
(f) *R and E are partial isometries. In particular $E = R^*$ and RE is the identity on $\mathcal{H}_{k_0}(S_0)$, while ER is a projection operator on $\mathcal{H}_{k_1}(S_1)$.*

Proposition 6 (Translation invariant kernels on $\mathbb{R}^d$). *Let $v : \mathbb{R}^d \to \mathbb{R}$ such that its Fourier transform $\tilde{v}$ is integrable and satisfies $\tilde{v} \geq 0$ on $\mathbb{R}^d$. Then*

(a) *The function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as $k(x, x') = v(x - x')$ for any $x, x' \in \mathbb{R}^d$ is a kernel and is called* translation invariant kernel.
(b) *The RKHS $\mathcal{H}_k(\mathbb{R}^d)$ and the norm $\|\cdot\|_{\mathcal{H}_k(\mathbb{R}^d)}$ are characterized by*

$$\mathcal{H}_k(\mathbb{R}^d) = \{f \in L^2(\mathbb{R}^d) \mid \|f\|_{\mathcal{H}_k(\mathbb{R}^d)} < \infty\}, \quad \|f\|^2_{\mathcal{H}_k(\mathbb{R}^d)} = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \frac{|(\mathcal{F}f)(\omega)|^2}{\tilde{v}(\omega)} d\omega,$$

where $\mathcal{F}f$ is the Fourier transform of f (see Proposition 2 for more details on $\mathcal{F}$).
(c) *The inner product $\langle \cdot, \cdot \rangle_k$ is characterized by*

$$\langle f, g \rangle_k = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \frac{(\mathcal{F}f)(\omega)\overline{(\mathcal{F}g)(\omega)}}{\tilde{v}(\omega)} d\omega.$$

A.4 Auxiliary results on C^∞ functions

Proposition 7 *Let U be an open set of $\mathbb{R}^d$ and $K \subset U$ be a compact set. Let $u \in C^\infty(U)$, then there exists $v \in C_0^\infty(\mathbb{R}^d)$ (with compact support), such that $v(x) = u(x)$ for all $x \in K$.*

Proof By Thm. 1.4.1, pag. 25 of [40] there exists $z_{K,U} \in C_0^\infty(U)$, i.e., a smooth function with compact support, such that $z_{K,U}(x) \in [0, 1]$ for any $x \in U$ and $z(x) = 1$ for any $x \in K$. Consider now the function $v_{K,U}$ defined as $v_{K,U}(x) = z_{K,U}(x)u(x)$ for all $x \in U$. The function $v_{K,U}$ is in $C_0^\infty(U)$, since it is the product of a $C_0^\infty(U)$ and a $C^\infty(U)$ function, moreover $v_{K,U}(x) = u(x)$ for all $x \in K$. The theorem is concluded by defining v as the extension of $v_{K,U}$ to $\mathbb{R}^d$, i.e., the function $v_K(x) = z_{K,U}(x)$ for any $x \in U$ and $v_K(x) = 0$ for any $x \in \mathbb{R}^d \setminus U$. This is always possible since $v_{K,U}$ is supported on a compact set K' which is contained in the open set U, so $v_{K,U}$ is already identically zero in the open set $U \setminus K'$. □

Lemma 6 *Given $\zeta \in \mathbb{R}^d$ and $r > 0$, there exists $u \in C_0^\infty(\mathbb{R}^d)$ such that for any $x \in \mathbb{R}^d$, it holds*

(i) $u(x) \in [0, 1]$;
(ii) $\|x\| \geq r \implies u(x) = 0$;
(iii) $\|x\| \leq r/2 \implies u(x) = 1$.

Proof Assume without loss of generality that $\zeta = 0$ and $r = 1$. Consider the following functions :

$$u_1(x) = \begin{cases} \exp\left(-\frac{1}{1-\|x\|^2}\right) & \text{if } \|x\| < 1 \\ 0 & \text{otherwise} \end{cases}, \qquad u_2(x) = \begin{cases} \exp\left(-\frac{1}{\|x\|^2-1/4}\right) & \text{if } \|x\| > 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

Both u_1 and u_2 belong to $C^\infty(\mathbb{R}^d)$ with values in $[0, 1]$. Moreover, $u_1 > \alpha_1$ on $B_{3/4}(0)$ and $u_2 \geq \alpha_2$ for some $\alpha_1, \alpha_2 > 0$ on $\mathbb{R}^d \setminus B_{3/4}(0)$, which implies that $u_1 + u_2 \in I$ on $\mathbb{R}^d$, where $I = [\min(\alpha_1, \alpha_2), 2]$. Since $(\cdot)^{-1}$ is infinitely differentiable on $(0, \infty)$ we see that $1/(u_1 + u_2)$ is well defined on all $\mathbb{R}^d$ and belongs to $C^\infty(\mathbb{R}^d)$, since $I \subset\subset (0, \infty)$. Consider the function

$$u_0 = \frac{u_1}{u_1 + u_2}.$$

It is non-negative, bounded by 1, and infinitely differentiable as a product. Moreover :

$$\forall x \in B_{1/2}(0),\ u_2(x) = 0 \implies u_0(x) = 1,$$
$$\forall x \in \mathbb{R}^d,\ u_1(x) = 0 \Leftrightarrow u_0(x) = 0 \Leftrightarrow x \in \mathbb{R}^d \setminus B_1(0).$$

To conclude the proof, given $r > 0$ and $\zeta \in \mathbb{R}^d$ we will take $u(x) = u_0((x - \zeta)/r)$. □

Lemma 7 *Let $N \in \mathbb{N}_+$, $\zeta_1, ..., \zeta_N \in \mathbb{R}^d$ and $r_1, ..., r_N > 0$. For $n \in \{1, \dots, N\}$, let $B_n = B_{r_n}(\zeta_n)$ be the open ball centered in ζ_n of radius r_n and $B'_n = B_{r_n/2}(\zeta_n) \subset B_n$ be the open ball centered in ζ_n of radius $r_n/2$. Then there exists functions $v_0, v_1, ..., v_N \in C^\infty(\mathbb{R}^d)$ such that*

(i) $v_0 = v_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n=1}^N B'_n}$
(ii) $v_n = v_n \cdot \mathbf{1}_{B_n}, \ \forall n \in \{1, \dots, N\}$
(iii) $\sum_{n=0}^N v_n^2 = 1$.

Proof For all $n \in [N]$, take u_n as in Lemma 6 with $r = r_n, \zeta = \zeta_n$ and define $u_0 = \prod_{n=1}^N (1 - u_n)$. Since $\forall n \in [N], \ u_n \in [0, 1]$, we also have $u_0 \in [0, 1]$. Moreover, let $R = \max_{n \in [N]} \|\zeta_n\| + r_n$, then

$$\forall \|x\| \geq R, \ \forall 1 \leq n \leq N, \ u_n(x) = 0 \text{ and } u_0(x) = 1.$$

Step 1. $u_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n \in [N]} B'_n} = u_0$ *and for all* $n \in [N]$, $u_n \cdot \mathbf{1}_{B_n} = u_n$.

By point (iii) of Lemma 6, $u_n = 1$ on B'_n for all $n \in [N]$, which shows that $u_0 = 0$ on $\bigcup_{n=1}^N B'_n$ and hence $u_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n \in [N]} B'_n} = u_0$. On the other hand, for all $n \in [N]$, point (ii) of Lemma 6 directly implies $u_n \cdot \mathbf{1}_{B_n} = u_n$.

Step 2. *The function* $\frac{1}{\sqrt{\sum_{n=0}^N u_i^2}}$ *is well defined and in* $C^\infty(\mathbb{R}^d)$.

By definition of u_0, if $u_0(x) = 0$, then there exists $n \in [N]$ such that $u_n(x) = 1$. Since all the u_n are non-negative, this shows that $s := \sum_{n=0}^N u_n^2 > 0$. Moreover, consider the closed ball $\bar{B}$ of radius R and centered in 0. Since $\bar{B}$ is compact, s is continuous and $s(x) > 0$ for any $x \in \bar{B}$, then there exists $0 < m_R \leq M_R < \infty$ such that $s(x) \in [m_R, M_R]$ for any $x \in \bar{B}$. Moreover, since for any $\|x\| \geq R, \ u_0(x) = 1$ and $\forall n \in [N], \ u_n(x) = 0$, we see that

$$\forall x \in \mathbb{R}^d \setminus B_R(0), \ \sum_{n=0}^N u_n^2(x) = 1.$$

Then $s \in [m, M]$ for any $x \in \mathbb{R}^d$, where $m = \min(m_R, 1)$ and $M = \max(M_R, 1)$.

Since the interval $I = [m, M]$ is a compact set included in the open set $(0, \infty)$ and $1/\sqrt{\cdot}$ is infinitely differentiable on $(0, \infty)$ then by Proposition 7 there exists $q_I \in C_0^\infty(\mathbb{R})$ such that $q_I(x) = 1/\sqrt{x}$ for any $x \in I$. Since $s(x) \in I$ for any $x \in \mathbb{R}^d$ we have

$$\frac{1}{\sqrt{\sum_{n=0}^N u_i^2}} = q_I \circ s.$$

Finally $q_I \circ s \in C^\infty(\mathbb{R}^d)$ since it is the composition of $q_I \in C_0^\infty(\mathbb{R})$ and $s = \sum_{n=0}^N u_n^2 \in C^\infty(\mathbb{R}^d)$ (since all the u_n are in $C^\infty(\mathbb{R}^d)$) and $s \in [m, M]$.

Step 3.

Finally, defining $v_n = \frac{u_n}{\sqrt{\sum_{n=0}^N u_n^2}}$ for all $0 \leq n \leq N$, $v_n \in C^\infty(\mathbb{R}^d)$ since it is the product of two infinitely differentiable functions. Moreover, $\sum_{n=0}^N v_i^2 = 1$ by

construction and $v_0 = v_0 \cdot \mathbf{1}_{\mathbb{R}^d \setminus \bigcup_{n=1}^N B'_n}$ since u_0 satisfies the same equality and v_0 is the product of u_0 by the strictly positive function $1/\sqrt{s}$. Analogously $v_n = v_n \cdot \mathbf{1}_{B_n}$, $\forall n \in \{1, \ldots, N\}$, since u_n satisfy the same equality and v_n is the product of u_n by the strictly positive function $1/\sqrt{s}$. □

B Fundamental results on scattered data approximation

We recall here some fundamental results about local polynomial approximation. In particular, we report here the proofs to track explicitly the constants. The proof techniques are essentially from [18, 20]. Denote by $\pi_k(\mathbb{R}^d)$ the set of multivariate polynomials of degree at most k, with $k \in \mathbb{N}$. In this section $B_r(x) \subset \mathbb{R}^d$ denotes the open ball of radius r and centered in x.

Proposition 8 ([18], Corollary 3.11. Local polynomial reproduction on a ball). *Let $k \in \mathbb{N}$, $d, m \in \mathbb{N}_+$ and $\delta > 0$. Let B_δ be an open ball of radius $\delta > 0$ in $\mathbb{R}^d$. Let $\widehat{Y} = \{y_1, \ldots, y_m\} \subset B_\delta$ be a non empty finite subset of B_δ. If either $k = 0$ or $h_{\widehat{Y}, B_\delta} \leq \frac{\delta}{9k^2}$, there exist $u_j : B_\delta \to \mathbb{R}$ with $j \in [m]$ such that*

(a) $\sum_{j\in[m]} p(y_j)u_j(x) = p(x), \quad \forall x \in B_\delta, p \in \pi_k(\mathbb{R}^d)$
(b) $\sum_{j\in[m]} |u_j(x)| \leq 2, \quad \forall x \in B_\delta.$

Lemma 8 (Bounds on functions with scattered zeros on a small ball [18, 20]). *Let $k \in \mathbb{N}$, $d, m \in \mathbb{N}_+$ and $\delta > 0$. Let $B_\delta \subset \mathbb{R}^d$ be a ball of radius δ in $\mathbb{R}^d$. Let $f \in C^{k+1}(B_\delta)$. Let $\widehat{Y} = \{y_1, \ldots, y_m\} \subset B_\delta$ be a non empty finite subset of B_δ. If either $k = 0$ or $h_{\widehat{Y}, B_\delta} \leq \frac{\delta}{9k^2}$, it holds:*

$$\sup_{x\in B_\delta} |f(x)| \leq 3C\delta^{k+1} + 2 \max_{i\in[m]} |f(y_i)|, \qquad C := \sum_{|\alpha|=k+1} \frac{1}{\alpha!} \|\partial^\alpha f\|_{L^\infty(B_\delta)}.$$

Proof Note that since either $k = 0$ or $h_{\widehat{Y}, B_\delta} \leq \frac{\delta}{9k^2}$, then we can apply Proposition 8 obtaining u_j with $j \in [m]$ with the local polynomial reproduction property. Define the function $s_{f,\widehat{Y}} = \sum_{j\in[m]} f(y_j)u_j$ and let $\tau = \max_{i\in[m]} |f(y_i)|$. Now, by using both Propositions 8(a) and 8(b), we have that for any $p \in \pi_k(\mathbb{R}^d)$ and any $x \in B_\delta$,

$$\begin{aligned}
|f(x)| &\leq |f(x) - p(x)| + |p(x) - s_{f,\widehat{Y}}(x)| + |s_{f,\widehat{Y}}(x)| \\
&\leq |f(x) - p(x)| + \sum_{j\in[m]} |p(y_j) - f(y_j)||u_j(x)| + \max_{j\in[m]} |f(y_j)| \sum_{j\in[m]} |u_j(x)| \\
&\leq \|f - p\|_{L^\infty(B_\delta)} \left(1 + \sum_{j\in[m]} |u_j(x)|\right) + \tau \sum_{j\in[m]} |u_j(x)| \\
&\leq 3\|f - p\|_{L^\infty(B_\delta)} + 2\tau.
\end{aligned}$$

In particular, consider the Taylor expansion of f at the center x_0 of B_δ up to order k (e.g. [41] Eq. 4.2.5 pag 95). For any $x \in B_\delta$, it holds

$$f(x) = \sum_{|\alpha|\le k} \frac{1}{\alpha!}\partial^\alpha f(x_0)(x-x_0)^\alpha + \sum_{|\alpha|=k+1} \frac{k+1}{\alpha!}(x-x_0)^\alpha \int_0^1 (1-t)^k \partial^\alpha f((1-t)x_0 + tx)dt.$$

By choosing $p(x) = \sum_{|\alpha|\le k} \frac{1}{\alpha!}\partial^\alpha f(x_0)(x-x_0)^\alpha \ \in \pi_k(\mathbb{R}^d)$ it holds:

$$\|f - p\|_{L^\infty(B_\delta)} \le \sum_{|\alpha|=k+1} \frac{\delta^{k+1}}{\alpha!}\|\partial^\alpha f\|_{L^\infty(B_\delta)} = C\delta^{k+1},$$

where $C = \sum_{|\alpha|=k+1} \frac{1}{\alpha!}\|\partial^\alpha f\|_{L^\infty(B_\delta)}$ is defined in the lemma. Gathering the previous equations,

$$\sup_{x\in B_\delta} |f(x)| \le 2\tau + 3C\delta^{k+1}.$$

□

Theorem 11 (Bounds on functions with scattered zeros [18, 20]). *Let $k, m \in \mathbb{N}$ s.t. $k \le m$ and $n, d \in \mathbb{N}_+$. Let $r > 0$ and Ω an open set of $\mathbb{R}^d$ of the form $\Omega = \bigcup_{x\in S} B_r(x)$ for some subset S of $\mathbb{R}^d$. Let $\widehat{X} = \{x_1, \dots, x_n\}$ be a non-empty finite subset of Ω. Let $f \in C^{m+1}(\Omega)$. If $h_{\widehat{X},\Omega} \le r \max(1, \frac{1}{18k^2})$, then*

$$\sup_{x\in\Omega} |f(x)| \ \le \ CC_f h_{\widehat{X},\Omega}^{k+1} \ + \ 2\max_{i\in[n]} |f(x_i)|,$$

where $C = 3\max(1, 18\,k^2)^{k+1}$ and $C_f = \sum_{|\alpha|=k+1} \frac{1}{\alpha!}\|\partial^\alpha f\|_{L^\infty(\Omega)}$.

Proof First, note that the condition that there exists a set S such that $\Omega = \bigcup_{x\in S} B_r(x)$ implies

$$\forall \delta \le r, \ \Omega = \bigcup_{x_0\in S_\delta} B_\delta(x_0), \qquad S_\delta = \{x' \in \Omega \ : \ \exists x \in S, \ \|x - x'\| \le r - \delta\}.$$

We will now prove the theorem for $k \ge 1$ and then the easier case $k = 0$, where we will use essentially only the Lipschitzianity of f.

Proof of the case $k \ge 1$. The idea of the proof is to apply Lemma 8 to a collection of balls of radius δ for a well chosen $\delta \le r$ and centered in $x_0 \in S_\delta$ defined above. Given $\widehat{X}$, to apply Lemma 8 on a ball of radius δ we have to restrict the points in $\widehat{X}$ to the subset belonging to that ball, i.e., $\widehat{Y}_{x_0,\delta} = \widehat{X} \cap B_\delta(x_0)$, $x_0 \in S_\delta$ and $\delta > 0$. The set $\widehat{Y}_{x_0,\delta}$ will have a fill distance $h_{x_0,\delta} = h_{\widehat{Y}_{x_0,\delta}, B_\delta(x_0)}$. First we are going to show that

$\widehat{Y}_{x_0,\delta}$ is not empty, when $r > \delta > h_{\widehat{X},\Omega}$. To obtain this result we need to study also the ball $B_{\delta'}(x_0)$ with $\delta' = \delta - h_{\widehat{X},\Omega}$.

Step 1. Showing that $\widehat{Y}_{x_0,\delta}$ is not empty and for any $y \in B_{\delta'}(x_0)$ there exists $z \in \widehat{Y}_{x_0,\delta}$ satisfying $\|y - z\| \leq h_{\widehat{X},\Omega}$. Let $x_0 \in S_\delta$ and $\delta \leq r$. This implies that $B_\delta(x_0) \subseteq \Omega$ by the characterization of Ω in terms of S_δ we gave above. Define now $\delta' = \delta - h_{\widehat{X},\Omega}$ and note that $B_{\delta'}(x_0)$ is non empty, since $\delta' > 0$, and that $B_{\delta'}(x_0) \subset B_\delta(x_0) \subseteq \Omega$. Now note that by definition of fill distance, for any $y \in B_{\delta'}(x_0)$ there exists a $z \in \widehat{X}$ such that $\|z - y\| \leq h_{\widehat{X},\Omega}$. Moreover note that $z \in B_\delta(x_0)$, since $\|x_0 - z\| \leq \|x_0 - y\| + \|y - z\| < \delta - h_{\widehat{X},\Omega} + h_{\widehat{X},\Omega} = \delta$. Since $z \in \widehat{X}$ and also in $B_\delta(x_0)$, then $z \in \widehat{Y}_{x_0,\delta}$ by definition of $\widehat{Y}_{x_0,\delta}$.

Step 2. Showing that $h_{x_0,\delta} \leq 2h_{\widehat{X},\Omega}$. Let $x \in B_\delta(x_0)$. We have seen in the previous step that the ball $B_{\delta'}(x_0)$ is well defined and non empty, with $\delta' = \delta - h_{\widehat{X},\Omega}$. Now note that also $B_{h_{\widehat{X},\Omega}}(x) \cap B_{\delta'}(x_0)$ is not empty, indeed the distance between the centers x, x_0 is strictly smaller than the sum of the two radii, indeed $\|x - x_0\| < \delta = \delta' + h_{\widehat{X},\Omega}$, since $x \in B_\delta(x_0)$. Take $w \in B_{h_{\widehat{X},\Omega}}(x) \cap B_{\delta'}(x_0)$. Since $w \in B_{\delta'}(x_0)$ by Step 1 we know that there exists $z \in \widehat{Y}_{x_0,\delta}$ with $\|w - z\| \leq h_{\widehat{X},\Omega}$. Since $w \in B_{h_{\widehat{X},\Omega}}(x)$, then we know that $\|x - w\| < h_{\widehat{X},\Omega}$. So $\|x - z\| \leq \|x - w\| + \|w - z\| < 2h_{\widehat{X},\Omega}$.

Step 3. Applying Lemma 8. Since, by assumption $h_{\widehat{X},\Omega} \leq r/(18k^2)$ and $k \geq 1$, then the choice $\delta = 18k^2 h_{\widehat{X},\Omega}$ implies $r \geq \delta > h_{\widehat{X},\Omega}$. So we can use the characterization of Ω in terms of S_δ and the results in the previous two steps, obtaining that for any $x_0 \in S_\delta$ the set $B_\delta(x_0) \subseteq \Omega$ and moreover the set $\widehat{Y}_{x_0,\delta}$ is not empty and covers $B_\delta(x_0)$ with a fill distance $h_{x_0,\delta} \leq 2h_{\widehat{X},\Omega}$. Since, $h_{x_0,\delta} \leq 2h_{\widehat{X},\Omega} \leq \delta/(9k^2)$ then we can apply Lemma 8 to each ball $B_\delta(x_0)$ obtaining

$$\sup_{x \in B_\delta(x_0)} |f(x)| \leq 3C_{\delta,x_0}\delta^{k+1} + 2 \max_{z \in \widehat{Y}_{x_0,\delta}} |f(z)|,$$

$$C_{\delta,x_0} := \sum_{|\alpha|=k+1} \frac{1}{\alpha!} \|\partial^\alpha f\|_{L^\infty(B_\delta(x_0))}.$$

The proof is concluded by noting that $\Omega = \bigcup_{x_0 \in S_\delta} B_\delta(x_0)$ and that for any $x_0 \in S_\delta$ we have $C_{\delta,x_0} \leq C_f$, $\delta^{k+1} \leq (18k^2)^{k+1} h_{\widehat{X},\Omega}^{k+1}$ and moreover that $\max_{z \in \widehat{Y}_{x_0,\delta}} |f(z)| \leq \max_{i \in [n]} |f(x_i)|$, since $\widehat{Y}_{x_0,\delta} \subseteq \widehat{X}$ by construction.

Proof of the case $k = 0$ Since $h_{\widehat{X},\Omega} \leq r$, by assumption, then $\delta = h_{\widehat{X},\Omega}$ implies that Ω admits a characterization as $\Omega = \bigcup_{x_0 \in S_\delta} B_\delta(x_0)$. Now let $x \in \Omega$ and choose $x_0 \in S_\delta$ such that $x \in B_\delta(x_0)$. One the one hand, since the segment $[x_0, x]$ is included in Ω, by Taylor inequality, $|f(x) - f(x_0)| \leq C_f \|x - x_0\| \leq C_f h_{\widehat{X},\Omega}$ and $C_f = \sum_{|\alpha|=1} \frac{1}{\alpha!} \|\partial^\alpha f\|_{L^\infty(\Omega)}$. One the other hand, by definition of $h_{\widehat{X},\Omega}$, there exists $z \in \widehat{X} \subset \Omega$ such that $\|z - x_0\| \leq h_{\widehat{X},\Omega} = \delta$. Since both the open segment $[x_0, z) \subset B_\delta(x_0) \subset \Omega$ and $z \in \Omega$, then the whole segment $[x_0, z] \subset \Omega$ and hence we can apply Taylor inequality to show $\|f(x_0) - f(z)\| \leq C_f \|z - x_0\| \leq C_f h_{\widehat{X},\Omega}$. Then we have

$$|f(x)| \leq |f(x) - f(x_0)| + |f(x) - f(z)| + |f(z)| \leq 2C_f h_{\widehat{X},\Omega} + \max_{i \in [n]} |f(x_i)|.$$

The proof of the step $k = 0$ is concluded by noting that the previous inequality holds for every $x \in \Omega$. □

C Auxiliary results on RKHS

We recall that the *nuclear norm* of a compact linear operator A is defined as $\|A\|_\star = \mathrm{Tr}(\sqrt{A^*A})$ or equivalently $\|A\|_\star = \sum_{j\in\mathbb{N}} \sigma_j$, where $(\sigma_j)_{j\in\mathbb{N}}$ are the singular values of A (Chapter 7 of [42] or [43] for the finite dimensional analogue).

Lemma 9 *Let Ω be a set, k be a kernel and $\mathcal{H}$ the associated RKHS. Let $A : \mathcal{H} \to \mathcal{H}$ be a trace class operator. If $\mathcal{H}$ satisfies Assumption 2(a), then*

$$\|r_A\|_{\mathcal{H}} \leq \mathsf{M}\|A\|_\star, \quad \textit{where} \quad r_A(x) := \langle \phi(x), A\phi(x)\rangle\,, \quad \forall x \in \Omega,$$

and $\|A\|_\star$ is the nuclear norm *of A. We recall that if $A \in \mathbb{S}_+(\mathcal{H})$ then $\|A\|_\star = Tr(A)$.*

Proof Since A is compact, it admits a singular value decomposition $A = \sum_{i\in\mathbb{N}} \sigma_i u_i \otimes v_i$. Here, $(\sigma_j)_{j\in\mathbb{N}}$ is a non-increasing sequence of non-negative eigenvalues converging to zero, and $(u_j)_{j\in\mathbb{N}}$ and $(v_j)_{j\in\mathbb{N}}$ are two orthonormal families of corresponding eigenvectors, (a family (e_j) is said to be orthonormal if for $i, j \in \mathbb{N}$, $\langle e_i, e_j\rangle = 1$ if $i = j$ and $\langle e_i, e_j\rangle = 0$ otherwise) [42]. Note that we can write r_A using this decomposition as $r_A(x) = \sum_{i\in\mathbb{N}} \sigma_i u_i(x) v_i(x) = \sum_{i\in\mathbb{N}} \sigma_i\, (u_i \cdot v_i)(x)$, for all $x \in \Omega$, where we denote by $\cdot$ the pointwise multiplication between two functions (this equality is justified by the following absolute convergence bound). By Assumption 2(a), the fact that A is trace-class (i.e., $\|A\|_\star < \infty$) and the fact that u_j, v_j satisfy $\|u_j\|_{\mathcal{H}} = \|v_j\|_{\mathcal{H}} = 1$, $j \in \mathbb{N}$, the following holds

$$\begin{aligned}\|r_A\|_{\mathcal{H}} &= \left\| \sum_{j\in\mathbb{N}} \sigma_j (u_j \cdot v_j) \right\|_{\mathcal{H}} \leq \sum_{j\in\mathbb{N}} \sigma_j \|u_j \cdot v_j\|_{\mathcal{H}} \\ &\leq \mathsf{M} \sum_{j\in\mathbb{N}} \sigma_j \|u_j\|_{\mathcal{H}} \|v_j\|_{\mathcal{H}} \leq \mathsf{M} \sum_{j\in\mathbb{N}} \sigma_j = \mathsf{M}\|A\|_\star.\end{aligned}$$

In the case where $A \in \mathbb{S}_+(\mathcal{H})$, we have $\|A\|_\star = \mathrm{Tr}(\sqrt{A^*A}) = \mathrm{Tr}(A)$. □

C.1 Proof of Proposition 2

Given the kernel k, the associated RKHS $\mathcal{H}$ and the canonical feature map $\phi : \Omega \to \mathcal{H}$ and a set of distinct points $\widehat{X} = \{x_1, \ldots, x_n\}$ define the *kernel matrix* $K \in \mathbb{R}^{n\times n}$ as $K_{i,j} = \langle \phi(x_i), \phi(x_j)\rangle = k(x_i, x_j)$ for all $i, j \in [n]$. Note that, since k is a p.d. kernel, then K is positive semidefinite, moreover when k is universal, then $\phi(x_1), \ldots, \phi(x_n)$ are linearly independent, so K is full rank and hence invertible. Universality of k is guaranteed since $\mathcal{H}$ contains the $C_0^\infty(\Omega)$ functions, by Assumption 1(a), and so can approximate continuous functions over compacts in Ω [17]. Denote by R the upper

triangular matrix corresponding to the Cholesky decomposition of K, i.e., R satisfies $K = R^\top R$. We are ready to start the proof of Proposition 2.

Proof Denote by $\widehat{S} : \mathcal{H} \to \mathbb{R}^n$ the linear operator that acts as follows

$$\widehat{S}g \;=\; (\,\langle \phi(x_1), g\rangle \;,\; \ldots,\; \langle \phi(x_n), g\rangle\,) \in \mathbb{R}^n, \qquad \forall g \in \mathcal{H}.$$

Define $\widehat{S}^* : \mathbb{R}^n \to \mathcal{H}$, i.e., the adjoint of $\widehat{S}$, as $\widehat{S}^*\beta = \sum_{i=1}^n \beta_i \phi(x_i)$ for $\beta \in \mathbb{R}^n$. Note, in particular, that $K = \widehat{S}\widehat{S}^*$ and that $\widehat{S}^* e_j = \phi(x_i)$, where e_j is the j-th element of the canonical basis of $\mathbb{R}^n$. We define the operator $V = R^{-\top}\widehat{S}$ and its adjoint $V^* = \widehat{S}^* R^{-1}$. By using the definition of V, the fact that $K = R^\top R$ by construction of R, and the fact that $K = \widehat{S}\widehat{S}^*$, we derive two facts.
On the one hand,

$$VV^* = R^{-\top}\widehat{S}\widehat{S}^* R^{-1} = R^{-\top} K R^{-1} = R^{-\top} R^\top R R^{-1} = I.$$

On the other hand, P is a projection operator, i.e., $P^2 = P$, P is positive definite and its range is $\mathrm{ran} P = \mathrm{span}\{\phi(x_i) \mid i \in [n]\}$, implying $P\phi(x_i) = \phi(x_i)$ for all $i \in [n]$. Indeed, using the equation above, $P^2 = V^*VV^*V = V^*(VV^*)V = V^*V = P$, and the positive-semi-definiteness of P is given by construction since it is the product of an operator and its adjoint. Moreover, the range of P is the same as that of V^* which in turn is the same as that of S^*, since R is invertible : $\mathrm{ran} P = \mathrm{span}\{\phi(x_i) \mid i \in [n]\}$. Finally, note that since $k(x, x') = \langle \phi(x), \phi(x')\rangle$, for any $x, x' \in \Omega$, then for any $j \in [n]$, Φ_j is characterized by

$$\begin{aligned}\Phi_j &= R^{-\top}(k(x_1, x_j), \ldots, k(x_n, x_j)) \\ &= R^{-\top}(\langle \phi(x_1), \phi(x_j)\rangle, \ldots, \langle \phi(x_n), \phi(x_j)\rangle) = R^{-\top}\widehat{S}\phi(x_j) = V\phi(x_j).\end{aligned}$$

□

D The constants of translation invariant and Sobolev kernels

D.1 Results for translation invariant and Sobolev kernels

Lemma 10 *Let Ω be a set and let $k(x, x') = v(x - x')$ for all $x, x' \in \Omega$, be a translation invariant kernel for some function $v : \mathbb{R}^d \to \mathbb{R}$. Denote by $\tilde{v}$ the Fourier transform of v. Let $\mathcal{H}$ be the associated RKHS. For any $f, g \in \mathcal{H}$ we have*

$$\begin{aligned}\|f \cdot g\|_{\mathcal{H}} &\le C \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}, \\ C &= (2\pi)^{d/4} \left\| \frac{\tilde{v} \star \tilde{v}}{\tilde{v}} \right\|^{1/2}_{L^\infty(\mathbb{R}^d)}.\end{aligned}$$

In particular, if there exists a non-increasing $g : [0, \infty] \to (0, \infty]$ *s.t.* $\tilde{v}(\omega) \leq g(\|\omega\|)$, *then*

$$C \leq \sqrt{2}(2\pi)^{d/2} v(0)^{1/2} \sup_{\omega \in \mathbb{R}^d} \sqrt{\frac{g(\frac{1}{2}\|\omega\|)}{\tilde{v}(\omega)}}.$$

Proof First note that by as recalled in Example 5, there exists an extension operator, i.e., a partial isometry $E : \mathcal{H} \to \mathcal{H}(\mathbb{R}^d)$ such that $r = Eu$ satisfies $r(x) = u(x)$ for all $x \in \Omega$ and $\|u\|_{\mathcal{H}} = \|r\|_{\mathcal{H}}$, for any $u \in \mathcal{H}$. Moreover there exists a restriction operator $R : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}$, as recalled in Example 5, such that $RE : \mathcal{H} \to \mathcal{H}$ is the identity operator and $ER : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}(\mathbb{R}^d)$ is a projection operator whose range is $\mathcal{H}$. Moreover, note that $f \cdot g = R(Ef \cdot Eg)$ since for any $x \in \Omega$, $(R(Ef \cdot Eg))(x) = (Ef)(x)(Eg)(x) = f(x)g(x) = (f \cdot g)(x)$. Since ER is a projection operator, then $\|ER\|_{\text{op}} \leq 1$, hence

$$\begin{aligned}\|f \cdot g\|_{\mathcal{H}} &= \|R(Ef \cdot Eg)\|_{\mathcal{H}} = \|ER(Ef \cdot Eg)\|_{\mathcal{H}(\mathbb{R}^d)} \\ &\leq \|ER\|_{\text{op}} \|Ef \cdot Eg\|_{\mathcal{H}(\mathbb{R}^d)} \leq \|Ef \cdot Eg\|_{\mathcal{H}(\mathbb{R}^d)}.\end{aligned}$$

Let $a = Ef$ and $b = Eg$. Denote by $\tilde{a}, \tilde{b}$ their Fourier transform and by $\widetilde{a \cdot b}$ the Fourier transform of $a \cdot b$ (see Proposition 2 for more details). By expanding the definition of the Hilbert norm of translation invariant kernel

$$\|Ef \cdot Eg\|^2_{\mathcal{H}(\mathbb{R}^d)} = \|a \cdot b\|^2_{\mathcal{H}(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\widetilde{a \cdot b}(\omega)|^2}{\tilde{v}(\omega)} d\omega.$$

Now we bound $\widetilde{a \cdot b}$. Since $\widetilde{a \cdot b} = (2\pi)^{d/2} \tilde{a} \star \tilde{b}$ (see Proposition 2) where $\star$ corresponds to the convolution, by expanding it and by applying Cauchy-Schwarz we obtain

$$\begin{aligned}(2\pi)^{-d/2} |\widetilde{a \cdot b}(\omega)|^2 = |(\tilde{a} \star \tilde{b})(\omega)|^2 &= \left(\int_{\mathbb{R}^d} \tilde{a}(\sigma) \tilde{b}(\omega - \sigma) d\sigma \right)^2 \\ &= \left(\int_{\mathbb{R}^d} \frac{\tilde{a}(\sigma)}{\sqrt{\tilde{v}(\sigma)}} \frac{\tilde{b}(\omega - \sigma)}{\sqrt{\tilde{v}(\omega - \sigma)}} \sqrt{\tilde{v}}(\sigma) \sqrt{\tilde{v}}(\omega - \sigma) d\sigma \right)^2 \\ &\leq \int_{\mathbb{R}^d} \frac{\tilde{a}^2}{\tilde{v}}(\sigma) \frac{\tilde{b}^2}{\tilde{v}}(\omega - \sigma) d\sigma \int_{\mathbb{R}^d} \tilde{v}(\sigma) \tilde{v}(\omega - \sigma) d\sigma \\ &= \left(\frac{\tilde{a}^2}{\tilde{v}} \star \frac{\tilde{b}^2}{\tilde{v}} \right)(\omega) \, (\tilde{v} \star \tilde{v})(\omega).\end{aligned}$$

By using the bound above together with Hölder inequality and Young inequality for convolutions, we have

$$(2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\widetilde{a \cdot b}(\omega)|^2}{\tilde{v}(\omega)} d\omega \leq \int_{\mathbb{R}^d} \left(\frac{\tilde{a}^2}{\tilde{v}} \star \frac{\tilde{b}^2}{\tilde{v}} \right)(\omega) \frac{(\tilde{v} \star \tilde{v})(\omega)}{\tilde{v}(\omega)} d\omega$$

$$
\begin{aligned}
&\leq \left\| \frac{\tilde{a}^2}{\tilde{v}} \star \frac{\tilde{b}^2}{\tilde{v}} \right\|_{L^1(\mathbb{R}^d)} \left\| \frac{\tilde{v} \star \tilde{v}}{\tilde{v}} \right\|_{L^\infty(\mathbb{R}^d)} \\
&\leq \left\| \frac{\tilde{a}^2}{\tilde{v}} \right\|_{L^1(\mathbb{R}^d)} \left\| \frac{\tilde{b}^2}{\tilde{v}} \right\|_{L^1(\mathbb{R}^d)} \left\| \frac{\tilde{v} \star \tilde{v}}{\tilde{v}} \right\|_{L^\infty(\mathbb{R}^d)} \\
&= (2\pi)^{d/2} \left\| \frac{\tilde{v} \star \tilde{v}}{\tilde{v}} \right\|_{L^\infty(\mathbb{R}^d)} \|a\|^2_{\mathcal{H}(\mathbb{R}^d)} \|b\|^2_{\mathcal{H}(\mathbb{R}^d)} = C^2,
\end{aligned}
$$

where in the last step we used the definitions of inner products for translation invariant kernels. The proof is concluded by noting that $\|a\|_{\mathcal{H}(\mathbb{R}^d)} = \|Ef\|_{\mathcal{H}(\mathbb{R}^d)} = \|f\|_{\mathcal{H}}$ and the same holds for b, i.e., $\|b\|_{\mathcal{H}(\mathbb{R}^d)} = \|g\|_{\mathcal{H}}$. A final consideration is that C can be further bounded by applying Proposition 9 and noting that $v(0) = (2\pi)^{-d/2} \int \tilde{v}(\omega) d\omega = (2\pi)^{-d/2} \|\tilde{v}\|_{L^1(\mathbb{R}^d)}$, via the characterization of v in terms of $\tilde{v}$ in Proposition 2(e), since $\tilde{v}(\omega) \geq 0$ and integrable. □

Proposition 9 *Let $u \in L^1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ be $u(x) \geq 0$ for $x \in \mathbb{R}^d$ and such that there exists a non-increasing function $g : [0, \infty) \to (0, \infty)$ satisfying $u(x) \leq g(\|x\|)$ for all $x \in \mathbb{R}^d$. Then it holds :*

$$
\forall x \in \mathbb{R}^d,\ 0 \leq (u \star u)(x) \leq 2\|u\|_{L^1(\mathbb{R}^d)} g\left(\tfrac{1}{2}\|x\|\right).
$$

In particular, if $u > 0$, it holds

$$
\left\| \frac{u \star u}{u} \right\|_{L^\infty(\mathbb{R}^d)} \leq 2\|u\|_{L^1(\mathbb{R}^d)} \sup_{x \in \mathbb{R}^d} \frac{g\left(\frac{1}{2}\|x\|\right)}{u(x)}.
$$

Proof For any $x \in \mathbb{R}^d$,

$$
(u \star u)(x) = \sup_{x \in \mathbb{R}^d} \int_{\mathbb{R}^d} u(y) u(x - y) dy.
$$

Let $S_x = \{y \mid \|x - y\| \leq \frac{1}{2}\|x\|\}$. Note that, when $y \in \mathbb{R}^d \setminus S_x$, then $\|x - y\| > \frac{1}{2}\|x\|$. Instead, when $y \in S_x$, then

$$
\tfrac{1}{2}\|x\| \leq \|x\| - \|x - y\| \leq \|y\|.
$$

Since g is non-increasing, for any $x \in \mathbb{R}^d$ we have

$$
\begin{aligned}
\int_{\mathbb{R}^d} u(y) u(x - y) dy &= \int_{S_x} u(y) u(x - y) dy \ + \ \int_{\mathbb{R}^d \setminus S_x} u(y) u(x - y) dy \\
&\leq \int_{S_x} g(\|y\|) u(x - y) dy \ + \ \int_{\mathbb{R}^d \setminus S_x} u(y) g(\|x - y\|) dy \\
&\leq \int_{S_x} g\left(\tfrac{1}{2}\|x\|\right) u(x - y) dy \ + \ \int_{\mathbb{R}^d \setminus S_x} u(y) g\left(\tfrac{1}{2}\|x\|\right) dy
\end{aligned}
$$

$$
\begin{aligned}
&\leq \int_{\mathbb{R}^d} g\left(\tfrac{1}{2}\|x\|\right) u(x-y)dy \;+\; \int_{\mathbb{R}^d} u(y) g\left(\tfrac{1}{2}\|x\|\right) dy \\
&= \int_{\mathbb{R}^d} g\left(\tfrac{1}{2}\|x\|\right) u(y)dy \;+\; \int_{\mathbb{R}^d} u(y) g\left(\tfrac{1}{2}\|x\|\right) dy \\
&= 2\, g\left(\tfrac{1}{2}\|x\|\right) \int_{\mathbb{R}^d} u(y)dy,
\end{aligned}
$$

where: in the first inequality we bounded $u(y)$ with $g(\|y\|)$ and $u(x-y)$ with $g(\|x-y\|)$, in the first and the second integral, respectively; in the second inequality we bounded $g(\|y\|)$ with $g(\frac{1}{2}\|x\|)$, since $\|y\| \geq \frac{1}{2}\|x\|$ when $y \in S_x$ and we bounded $g(\|x-y\|)$ with $g(\frac{1}{2}\|x\|)$, since $\|x-y\| \geq \frac{1}{2}\|x\|$ when $y \in \mathbb{R}^d \setminus S_x$; in the third we extended the integration domains to $\mathbb{R}^d$. □

D.2 Proof of Proposition 1

Proof We prove here that the Sobolev kernel satisfies Assumption 2. Let $k = k_s$ from Eq. (3.2). As we have seen in Example 1 $\mathcal{H} = W_2^s(\Omega)$ and $\|\cdot\|_{W_2^s(\Omega)}$ is equivalent to $\|\cdot\|_{\mathcal{H}}$, when $s > d/2$ and Ω satisfies Assumption 1(a) since this assumption implies that Ω satisfies the *cone condition* [18].
Recall that k is translation invariant, i.e., $k(x, x') = v(x-x')$ for any $x, x' \in \mathbb{R}^d$, with v defined in Example 1. The Fourier transform of v is $\tilde{v}(\omega) = C_0(1+\|\omega\|^2)^{-s}$ with $C_0 = \frac{2^{d/2}\Gamma(s)}{\Gamma(s-d/2)}$ [18]. In the rest of the proof, C_0 will always refer to this constant.
We are going to divide the proof in one step per point of Assumption 2.
Proof of Assumption 2(d) for the Sobolev kernel. Let $\alpha \in \mathbb{N}^d$, $m = |\alpha|$. Assume $m < s - d/2$, i.e., $m \in \{1, \ldots, \lfloor s - (d+1)/2 \rfloor\}$. Since k is translation invariant, then $\partial_x^\alpha \partial_y^\alpha k(x, y) = (-1)^m\, v_{2\alpha}(x-y)$ with $v_{2\alpha}(z) = \partial_z^{2\alpha} v(z)$ for all $z \in \mathbb{R}^d$. So

$$
\begin{aligned}
&\sup_{x,y\in\Omega} |\partial_x^\alpha \partial_y^\alpha k(x, y)| = \sup_{x,y\in\Omega} |\partial_x^\alpha \partial_y^\alpha v(x-y)| \leq \sup_{z\in\mathbb{R}^d} |\partial_z^{2\alpha} v(z)| \\
&\quad \leq (2\pi)^{-d/2} \|\omega^{2\alpha}\tilde{v}(z)\|_{L^1(\mathbb{R}^d)},
\end{aligned}
$$

where in the last step we used elementary properties of the Fourier transform (in particular the ones recalled in Proposition 2(c) and 2(e)). Let $S_{d-1} = 2\frac{\pi^{d/2}}{\Gamma(d/2)}$ be the area of the $d-1$ dimensional sphere. Since $m < s - d/2$ and $\tilde{v} \geq 0$,

$$
\begin{aligned}
\|\omega^{2\alpha}\tilde{v}(z)\|_{L^1(\mathbb{R}^d)} &\leq \int_{\mathbb{R}^d} \|\omega\|^{2m} \tilde{v}(\omega)d\omega = C_0 S_{d-1} \int_0^\infty \frac{r^{2m+d-1}}{(1+r^2)^s} dr \\
&= C_0 S_{d-1} \int_0^\infty \frac{t^{m+d/2-1}}{2(1+t)^s} dt = C_0 S_{d-1} \frac{\Gamma(m+d/2)\Gamma(s-d/2-m)}{2\Gamma(s)},
\end{aligned}
$$

where we performed a change of variable $r = \sqrt{t}$ and $dr = \frac{dt}{2\sqrt{t}}$ and applied Eq. 5.12.3 pag. 142 of [44] to the resulting integral. Thus, Assumption 2(d) holds with

$$\mathsf{D}_m^2 = C_0 \frac{\pi^{d/2}\Gamma(m+d/2)\Gamma(s-m-d/2)}{\Gamma(d/2)\Gamma(s)} = \frac{(2\pi)^{d/2}\Gamma(m+d/2)\Gamma(s-d/2-m)}{\Gamma(s-d/2)\Gamma(d/2)}.$$

Proof of Assumption 2(a) for the Sobolev kernel. First, note that $C^\infty(\mathbb{R}^d)|_\Omega \subset W_\infty^s(\Omega) \subset W_2^s(\Omega)$. Indeed, since Ω is bounded, for any $f \in C^\infty(\mathbb{R}^d)$, $\|\partial^\alpha f|_\Omega\|_{L^\infty(\Omega)} < \infty$ for any $\alpha \in \mathbb{N}^d$. This shows that $f|_\Omega \in W_\infty^s(\Omega)$. Moreover $W_\infty^s(\Omega) \subset W_2^s(\Omega)$ since $\|\cdot\|_{L^2(\Omega)} \leq \mathrm{vol}(\Omega)^{1/2}\|\cdot\|_{L^\infty(\Omega)}$ because Ω is bounded. Second, since $\tilde{v}(\omega) = g_s(\|\omega\|)$ with $g_s(t) = C_0(1+t^2)^{-s}$, positive and non-increasing, we can apply Lemma 10. Therefore, for $C = \sqrt{2}(2\pi)^{d/2}v(0)^{1/2}\sup_{t\geq 0}\left(\frac{g_s(t/2)}{g_s(t)}\right)^{1/2}$ it holds $\|f\cdot g\|_{\mathcal{H}} \leq C\|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}$. In particular we have $\sup_{t\geq 0}\left(\frac{g_s(t/2)}{g_s(t)}\right)^{1/2} \leq 2^s$ and $v(0) = 1$, since $\lim_{t\to 0} t^{s-d/2}\mathcal{K}_{s-d/2}(t) = \Gamma(s-d/2)/2^{1+d/2-s} = 1/C_0$ ([44] Eq. 10.30.2 pag. 252) and $v(x) = C_0 t^{s-d/2}\mathcal{K}_{s-d/2}(t)$, $t = \|x\|$. Thus, Assumption 2(a) holds with constant

$$\mathsf{M} = \pi^{d/2}2^{(2s+d+1)/2}.$$

Proof of Assumption 2(b) for the Sobolev kernel. First we recall from [11] that for any $s > d/2$, there exists a constant C_s such that

$$\forall h \in W_2^s(\mathbb{R}^d),\ \|h\|_{L^\infty(\mathbb{R}^d)} \leq C_s\|h\|_{W_2^s(\mathbb{R}^d)}.$$

In particular, this shows that $W_2^s(\mathbb{R}^d) \subset L^\infty(\mathbb{R}^d)$. Fix such a constant C_s in the rest of the proof.

Let $p \in \mathbb{N}$ and $g \in C^\infty(\mathbb{R}^p)$ with $g(0,0,\ldots,0) = 0$. From (i) of Thm. 11 in [45], there exists a constant c_g depending only on g, p, s such that for any $h_1,\ldots,h_p \in W_2^s(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, it holds

$$\|g(h_1,\ldots,h_p)\|_{W_2^s(\mathbb{R}^d)} \leq c_g \sup_{i\in[p]}\|h_i\|_{W_2^s(\mathbb{R}^d)}\left(1+\|h_i\|_{L^\infty(\mathbb{R}^d)}^{\max(0,s-1)}\right).$$

Since $s > d/2$, the bound above shows, in particular, that for any $h_1,\ldots,h_p \in W_2^s(\mathbb{R}^d)$, it holds

$$\begin{aligned}\|g(h_1,\ldots,h_p)\|_{W_2^s(\mathbb{R}^d)} &\leq c_g' \sup_{i\in[p]}\left(\|h_i\| + \|h_i\|_{W_2^s(\mathbb{R}^d)}^{\max(1,s)}\right),\\ c_g' &= c_g \max\left(1, C_s^{\max(0,s-1)}\right).\end{aligned}$$

Since $W_2^s(\mathbb{R}^d) = \mathcal{H}(\mathbb{R}^d)$ and $\|\cdot\|_{W_2^s(\mathbb{R}^d)}$ and $\|\cdot\|_{\mathcal{H}(\mathbb{R}^d)}$ are equivalent (see [11]), the previous inequality holds for $\|\cdot\|_{\mathcal{H}(\mathbb{R}^d)}$ with a certain constant c_g' depending only on g, p, s, d. In particular, this implies that $g(h_1,\ldots,h_p) \in \mathcal{H}(\mathbb{R}^d)$ for any $h_1,\ldots,h_p \in \mathcal{H}(\mathbb{R}^d)$. Now we are going to prove the same implication for the restriction on Ω.
First note that any function in $a \in C^\infty(\mathbb{R}^p)$ can be written as $a(z) = q\,1(z) + g(z)$, $z \in \mathbb{R}^p$ where $q = a(0,0,\cdots,0) \in \mathbb{R}$, $g \in C^\infty(\mathbb{R}^p)$ with $g(0,0,\cdots,0) = 0$

and $1(z) = 1$ for all $z \in \mathbb{R}^p$. Recall the definition and basic results on the extension operator $E : \mathcal{H} \to \mathcal{H}(\mathbb{R}^d)$ from Example 5. For any $f_1, \ldots, f_p \in \mathcal{H}$, note that $g((Ef_1)(x), \ldots, (Ef_p)(x)) = g(f_1(x), \ldots, f_p(x))$ for all $x \in \Omega$. We can now apply the results of Example 5 to show that $g(f_1, \ldots, f_p) \in \mathcal{H}$:

$$\begin{aligned}\|g(f_1, \ldots, f_p)\|_{\mathcal{H}} &= \inf_u \|u\|_{\mathcal{H}(\mathbb{R}^d)} \ \ s.t. \ \ u(x) = g(f_1(x), \ldots, f_p(x)) \ \forall x \in \Omega \\ &\leq \|g(Ef_1, \ldots, Ef_p)\|_{\mathcal{H}(\mathbb{R}^d)} \\ &\leq c_g' \sup_{j \in [p]} \|Ef_j\|_{\mathcal{H}(\mathbb{R}^d)} + \|Ef_j\|_{\mathcal{H}(\mathbb{R}^d)}^{\max(1,s)} \\ &= c_g' \sup_{j \in [p]} \|f_j\|_{\mathcal{H}} + \|f_j\|_{\mathcal{H}}^{\max(1,s)} < \infty,\end{aligned}$$

where in the last step we used the fact that $\|\cdot\|_{\mathcal{H}} = \|E \cdot\|_{\mathcal{H}(\mathbb{R}^d)}$. The proof of this point is concluded by noting that, $a(f_1, \ldots, f_p) \in \mathcal{H}$, since $1 \in \mathcal{H}$, due to the Point (a) above, and

$$\|a(f_1, \ldots, f_p)\|_{\mathcal{H}} \leq q\|1\|_{\mathcal{H}} + \|g(f_1, \ldots, f_p)\|_{\mathcal{H}} < \infty.$$

Proof of Assumption 2(c) for the Sobolev kernel. This proof is done in Lemma 11, right below. □

Before stating Lemma 11 we are going to recall some properties. First, recall the Young inequality :

$$\forall f \in L^2(\mathbb{R}^d), \ \forall g \in L^1(\mathbb{R}^d), \ \|f \star g\|_{L^2(\mathbb{R}^d)} \leq \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^1(\mathbb{R}^d)}.$$

Moreover, by definition of the Sobolev kernel, it is a translation-invariant kernel with v defined in Example 1, with Fourier transform $\tilde{v}(\omega) = C_0(1 + \|\omega\|^2)^{-s}$. Let $\mathcal{H}(\mathbb{R}^d)$ be the reproducing kernel Hilbert space on $\mathbb{R}^d$ associated to the Sobolev kernel k_s. As recalled in Example 6, the $\mathcal{H}(\mathbb{R}^d)$-norm is characterized by

$$\forall f \in \mathcal{H}(\mathbb{R}^d), \ \|f\|_{\mathcal{H}(\mathbb{R}^d)} = (2\pi)^{-d/4} \|\tilde{f}/\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)}, \tag{D.1}$$

where $\tilde{f} = \mathcal{F}(f)$ is the Fourier transform of f (see [11]). Then we recall that $\tilde{v} \in L^1(\mathbb{R}^d)$, since $s > d/2$, so for any $f \in \mathcal{H}(\mathbb{R}^d)$

$$\|\tilde{f}\|_{L^1(\mathbb{R}^d)} = \|\sqrt{\tilde{v}}\tilde{f}/\sqrt{\tilde{v}}\|_{L^1(\mathbb{R}^d)} \leq \|\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)} \|\tilde{f}/\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)} = C_1 \|f\|_{\mathcal{H}(\mathbb{R}^d)}. \tag{D.2}$$

where $C_1 = (2\pi)^{d/4} \|\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)}$. A useful consequence of the inequality above is obtained by considering that $\|f\|_{L^\infty(\mathbb{R}^d)}$ is bounded by the L^1 norm of $\tilde{f}$ (see Proposition 2(e)), then

$$\|f\|_{L^\infty} \leq (2\pi)^{-d/2} \|\tilde{f}\|_{L^1(\mathbb{R}^d)} \leq C_2 \|f\|_{\mathcal{H}(\mathbb{R}^d)}, \tag{D.3}$$

where $C_2 = (2\pi)^{-d/4} \|\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)}$.

Lemma 11 (Assumption 2(c) for Sobolev Kernels). *Let $\mathcal{H}$ be the RKHS associated to the translation invariant Sobolev Kernel defined in Example 1, with $s > d/2$. Then Assumption 2(c) is satisfied.*

Proof For the rest of the proof we fix $u : \Omega \to \mathbb{R}$ with $u \in \mathcal{H}$, $r > 0$ and $z \in \mathbb{R}^d$ such that $B_r(z) \subset \Omega$. Let $E_\Omega : \mathcal{H} \to \mathcal{H}(\mathbb{R}^d)$ be the extension operator from Ω to $\mathbb{R}^d$ (its properties are recalled in Example 5).Let $\chi \in C_0^\infty(\mathbb{R}^d)$ be given by Lemma 6 such that $\chi = 1$ on $B_r(z)$, $\chi = 0$ on $\mathbb{R}^d \setminus B_{2r}(z)$ and $\chi \in [0, 1]$. Define for any $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$

$$h_t(x) = \chi(x) w_t(x), \qquad w_t(x) = w((1-t)z + tx), \qquad w = E_\Omega u.$$

In particular we recall that, since E_Ω is a partial isometry (see Example 5) then $\|w\|_{\mathcal{H}(\mathbb{R}^d)} = \|u\|_{\mathcal{H}}$.

Step 1. Fourier transform of w_t. Denote with $\widetilde{w}$ the Fourier transform of w which is well defined since $w \in \mathcal{H}(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$ (see [11]), with $\tilde{\chi}$ the Fourier transform of χ. Since For any $t \neq 0$, denote with $\widetilde{w}_t$ the Fourier transform of w_t which is well defined using the results of Proposition 2, and which satisfies

$$\forall t \neq 0,\ \forall \omega \in \mathbb{R}^d,\ \widetilde{w}_t(\omega) = |t|^{-d} e^{i \frac{1-t}{t} z^\top \omega} \tilde{w}(\omega/t).$$

Step 2. Separating low and high order derivatives of h_t, and bounding the low order terms. For $t \neq 0$, denote with $\widetilde{h_t}$ the Fourier transform of h_t which is well defined since χ is bounded and $w_t \in L^2(\mathbb{R}^d)$. We will now bound $\|h_t\|_{\mathcal{H}(\mathbb{R}^d)}$ for all $t \neq 0$, by using the characterization in Eq. (D.1). Since $(x+y)^s \leq 2^{\max(s-1,0)}(x^s + y^s)$ for any $x, y \geq 0$, $s \geq 0$, then $(1 + \|\omega\|^2)^{s/2} \leq c_1(1 + \|\omega\|^s)$ for any $\omega \in \mathbb{R}^d$, with $c_1 = 2^{\max(s/2-1,0)}$ so using Eq. (D.1), we have

$$\begin{aligned}\sqrt{C_0}(2\pi)^{d/4}\|h_t\|_{\mathcal{H}(\mathbb{R}^d)} &= \|(1 + \|\cdot\|^2)^{s/2}\widetilde{h_t}\|_{L^2(\mathbb{R}^d)} \\ &\leq c_1 \|\widetilde{h_t}\|_{L^2(\mathbb{R}^d)} + c_1 \| \, |\cdot|^s_{\mathbb{R}^d} \widetilde{h_t}\|_{L^2(\mathbb{R}^d)}.\end{aligned}$$

The first term on the right hand side can easily be bounded using the fact that the Fourier transform is an isometry of $L^2(\mathbb{R}^d)$ (see Proposition 2 for more details), indeed

$$\|\widetilde{h_t}\|_{L^2(\mathbb{R}^d)} = \|h_t\|_{L^2(\mathbb{R}^d)} = \|\chi \cdot w_t\|_{L^2(\mathbb{R}^d)} \leq \|w_t\|_{L^\infty(\mathbb{R}^d)} \|\chi\|_{L^2(\mathbb{R}^d)} < \infty.$$

since $\chi \in C_0^\infty(\mathbb{R}^d)$ by definition, so it it bounded and has compact support, implying that $\|\chi\|_{L^2(\mathbb{R}^d)} < \infty$, moreover $\|w_t\|_{L^\infty(\mathbb{R}^d)} = \|w\|_{L^\infty(\mathbb{R}^d)}$ and $\|w\|_{L^\infty(\mathbb{R}^d)} \leq C_2 \|w\|_{\mathcal{H}(\mathbb{R}^d)}$ as recalled in Eq. (D.3) (the constant C_2 is defined in the same equation).

Step 3. Decomposing the high order derivatives of h_t. Note that since $\widetilde{h_t} = \widetilde{\chi \cdot w_t}$, by property of the Fourier transform (see Proposition 2(b)), $\widetilde{\chi \cdot w_t} = (2\pi)^{d/2} \widetilde{\chi} \star \widetilde{w_t}$.

Moreover, since $\|\omega\|^s \leq (\|\omega-\eta\|+\|\eta\|)^s \leq c_s(\|\omega-\eta\|^s+\|\eta\|^s)$ for any $\omega, \eta \in \mathbb{R}^d$, with $c = 2^{\max(s-1,0)}$, then, for all $\omega \in \mathbb{R}^d$ we have

$$\begin{aligned}
\|\omega\|^s|\widetilde{h}_t(\omega)| &= \|\omega\|^s|\widetilde{\chi \cdot w_t}(\omega)| = \|\omega\|^s(2\pi)^{\frac{d}{2}}|(\tilde{\chi}\star\tilde{w}_t)(\omega)| \\
&= (2\pi)^{\frac{d}{2}}|\int_{\mathbb{R}^d}\|\omega\|^s\widetilde{\chi}(\eta)\widetilde{w}_t(\omega-\eta)d\eta| \\
&\leq (2\pi)^{\frac{d}{2}}c\int_{\mathbb{R}^d}(|\widetilde{\chi}(\eta)|\,\|\eta\|^s)\,|\widetilde{w}_t(\omega-\eta)|\,d\eta \\
&\quad + (2\pi)^{\frac{d}{2}}c\int_{\mathbb{R}^d}|\widetilde{\chi}(\eta)|\,(|\widetilde{w}_t(\omega-\eta)|\,\|\omega-\eta\|^s)\,d\eta \\
&= c\,((J_s|\widetilde{\chi}|)\star|\widetilde{w}_t|)(\omega) \;+\; c\,(|\tilde{\chi}|\star(J_s|\widetilde{w}_t|))(\omega),
\end{aligned}$$

where we denoted by J_s the function $J_s(\omega) = \|\omega\|^s$ for any $\omega \in \mathbb{R}^d$. Applying Young's inequality, it holds :

$$\begin{aligned}
\|J_s\widetilde{h}_t\|_{L^2(\mathbb{R}^d)} &\leq c\,\|(J_s|\widetilde{\chi}|)\star|\widetilde{w}_t|\|_{L^2(\mathbb{R}^d)} \;+\; c\||\tilde{\chi}|\star(J_s|\widetilde{w}_t|)\|_{L^2(\mathbb{R}^d)} \\
&\leq c\|J_s\widetilde{\chi}\|_{L^2(\mathbb{R}^d)}\|\widetilde{w}_t\|_{L^1(\mathbb{R}^d)} + c\|J_s\widetilde{w}_t\|_{L^2(\mathbb{R}^d)}\,\|\widetilde{\chi}\|_{L^1(\mathbb{R}^d)}.
\end{aligned}$$

Step 4. Bounding the elements of the decomposition. Now we are ready to bound the four terms of the decomposition of $\|J_s\widetilde{h}_t\|_{L^2(\mathbb{R}^d)}$. First term, since $\chi \in C_0^\infty(\mathbb{R}^d) \subset \mathcal{H}(\mathbb{R}^d)$, and $J_s(\omega) \leq \sqrt{C_0/\tilde{v}(\omega)}$ for any $\omega \in \mathbb{R}^d$, then $\|J_s\widetilde{\chi}\|_{L^2(\mathbb{R}^d)} \leq \sqrt{C_0}\|\widetilde{\chi}/\sqrt{\tilde{v}}\|_{L^2(\mathbb{R}^d)} = (2\pi)^{d/4}\sqrt{C_0}\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}$, where we used Eq. (D.1). Second term, $\|\widetilde{\chi}\|_{L^1(\mathbb{R}^d)} < \infty$, since $\|\widetilde{\chi}\|_{L^1(\mathbb{R}^d)} \leq C_1\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}$, via Eq. (D.2) (the constant C_1 is defined in the same equation) and we have seen already that $\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}$ is bounded. Third term, by a change of variable $\tau = \omega/t$,

$$\begin{aligned}
\|\widetilde{w}_t\|_{L^1(\mathbb{R}^d)} &= \int_{\mathbb{R}^d}|\widetilde{w}_t(\omega)|d\omega = \int_{\mathbb{R}^d}|t|^{-d}|\tilde{w}(\omega/t)|d\omega \\
&= \int_{\mathbb{R}^d}|\tilde{w}(\tau)|d\tau = \|\tilde{w}\|_{L^1(\mathbb{R}^d)},
\end{aligned}$$

moreover $\|\tilde{w}\|_{L^1(\mathbb{R}^d)} \leq C_1\|w\|_{\mathcal{H}(\mathbb{R}^d)} = C_1\|u\|_{\mathcal{H}}$ via Eq. (D.2) and the fact that $\|w\|_{\mathcal{H}(\mathbb{R}^d)} = \|u\|_{\mathcal{H}}$ as recalled at the beginning of the proof. Finally, fourth term, for $t \in \mathbb{R}\setminus\{0\}$,

$$\begin{aligned}
\|J_s\widetilde{w}_t\|^2_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d}\|\omega\|^{2s}|\widetilde{w}_t(\omega)|^2\,d\omega = t^{-2d}\int_{\mathbb{R}^d}\|\omega\|^{2s}|\widetilde{w}(\omega/t)|^2d\omega \\
&= t^{2s-d}\int_{\mathbb{R}^d}\|\tau\|^{2s}|\widetilde{w}(\tau)|^2d\tau \leq t^{2s-d}\int_{\mathbb{R}^d}(1+\|\tau\|^2)^s|\widetilde{w}(\tau)|^2d\tau \\
&= t^{2s-d}(2\pi)^{d/2}C_0\|w\|^2_{\mathcal{H}(\mathbb{R}^d)}.
\end{aligned}$$

where we performed a change of variable $\omega = t\,\tau$, $t^d d\tau = d\omega$ and used the definition in Eq. (D.1) and the fact that $\|\tau\|^{2s} \leq (1+\|\tau\|^2)^s$ for any $\tau \in \mathbb{R}^d$. The proof of

the bound of the fourth term is concluded by recalling that $\|w\|_{\mathcal{H}(\mathbb{R}^d)} = \|u\|_{\mathcal{H}}$ as discussed in the proof of the bound for the previous term.

Conclusion. Putting all our bounds together, we get :

$$\forall t \in \mathbb{R} \setminus \{0\},\ \|h_t\|_{\mathcal{H}(\mathbb{R}^d)} \leq (A + B\, t^{s-d/2})\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}},$$

where $A = c_1C_2 + cc_1C_1(2\pi)^{d/4}\sqrt{C_0}$ and $B = cc_1C_1(2\pi)^{d/4}\sqrt{C_0}$, where $c = 2^{\max(s-1,0)}$, $c_1 = 2^{\max(s/2-1,0)}$, while C_1 is defined in Eq. (D.2), C_2 in Eq. (D.3). Now define

$$\forall x \in \mathbb{R}^d,\ \overline{g}_{z,r}(x) = \int_0^1 (1-t)h_t(x)dt,$$

and note that, by construction $\overline{g}_{z,r}(x) = \int_0^1 (1-t)u(tz + (1-t)x)dt$ for any $x \in B$ since u and χw coincide on B.

Note that the map $t \in (0, 1) \mapsto (1-t)\|h_t\|_{\mathcal{H}(\mathbb{R}^d)}$ is measurable, using the expression in Eq. (D.1).

Moreover, since for all $t \in (0, 1)$, it holds $\|h_t\|_{\mathcal{H}(\mathbb{R}^d)} \leq (A+Bt^{s-d/2})\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}}$ $\leq (A+B)\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}}$ since $s > d/2$, the map $t \mapsto (1-t)h_t$ is in integrable, and thus

$$\begin{aligned}\|\overline{g}_{z,r}\|_{\mathcal{H}(\mathbb{R}^d)} = \Big\|\int_0^1 (1-t)h_t dt\Big\|_{\mathcal{H}(\mathbb{R}^d)} &\leq \int_0^1 |1-t|\|h_t\|_{\mathcal{H}(\mathbb{R}^d)}dt \\ &\leq (A+B)\|\chi\|_{\mathcal{H}(\mathbb{R}^d)}\|u\|_{\mathcal{H}} < \infty,\end{aligned}$$

which implies that the function $\overline{g}_{z,r}$ belongs to $\mathcal{H}(\mathbb{R}^d)$. Finally, denote by $R_\Omega : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}$ the restriction operator (see Example 5 for more details). By construction $(R_\Omega g)(x) = g(x)$ for any $g \in \mathcal{H}(\mathbb{R}^d)$ and $x \in \Omega$, defining $g_{z,r} = R_\Omega \overline{g}_{z,r}$ the lemma is proven. □

E Proofs for Algorithm 1

We start with two technical lemmas that will be used by the proofs in this section.

Lemma 12 (Technical result). *Let $\alpha \geq 1$, $\beta \geq 2$ and $n \in \mathbb{N}$. If $n \geq 2\alpha\log(2\beta\alpha)$, then it holds*

$$\frac{\alpha\log(\beta n)}{n} \leq 1.$$

Proof Note that the function $x \mapsto \frac{\log(\beta x)}{x}$ is strictly decreasing on $[\exp(1)/\beta, +\infty]$.

Moreover, $2\alpha\log(2\beta\alpha) \geq 2\log 4 \geq \exp(1)/2 \geq \exp(1)/\beta$ since $\beta \geq 2$ and $\alpha \geq 1$.

Now assume $n \geq c\alpha$ with $c = 2\log(2\beta\alpha)$. It holds:

$$\frac{\alpha\log(\beta n)}{n} \leq \frac{\log(\beta c\alpha)}{c} \leq \frac{\log(\frac{c}{2}) + \log(2\alpha\beta)}{c} \leq \frac{1}{2} + \frac{1}{2}\,\frac{2\log(2\beta\alpha)}{c} \leq 1,$$

where we used the definition of c and the fact that $\log(c/2) \leq c/2 - 1 \leq c/2$. □

Lemma 13 *Let $\overrightarrow{u} \in S_{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$, $\alpha \in [0, \pi/2]$, $x_0 \in \mathbb{R}^d$ and $t > 0$. Define the cone centered at x_0, directed by $\overrightarrow{u}$ of radius t with aperture α:*

$$C^{\alpha}_{x_0, \overrightarrow{u}, t} = \left\{ x \in B_t(x_0) \mid \tfrac{x - x_0}{\|x - x_0\|} \cdot \overrightarrow{u} \leq \cos(\alpha),\ x \neq x_0 \right\},$$

where we denoted by $\cdot$ the scalar product among vectors. Then the volume of this cone is lower bounded as

$$vol(C^{\alpha}_{x_0, \overrightarrow{u}, t}) \geq \frac{(\sqrt{\pi} \sin(\alpha))^{d-1} (t\ \cos \alpha)^d}{d \Gamma((d+1)/2)}.$$

Moreover, let $x_0 \in \mathbb{R}^d$ and $r > 0$. Let $x \in B_r(x_0)$ and $0 < t \leq r$. The intersection $B_t(x) \cap B_r(x_0)$ contains the cone $C^{\pi/3}_{x, \overrightarrow{u}, t}$, where $\overrightarrow{u} = \frac{x - x_0}{\|x - x_0\|}$ if $x \neq x_0$ and any unit vector otherwise.

Proof **1. Bound on the volume of the cone**. Without loss of generality, assume $x_0 = 0$ and $\overrightarrow{u} = e_1$ since the Lebesgue measure is invariant by translations and rotations. A simple change of variable also shows that $\text{vol}(C^{\alpha}_{0, \overrightarrow{u}, t}) = t^d \text{vol}(C^{\alpha}_{0, \overrightarrow{u}, 1})$. Now note the following inclusion (the proof is trivial):

$$\widetilde{C} := \left\{ x = (x_1, z) \in \mathbb{R}^d = \mathbb{R} \times \mathbb{R}^{d-1} \ :\ z \leq \cos(\alpha),\ \|z\|_{\mathbb{R}^{d-1}} \leq x_1 \sin(\alpha) \right\} \subset C^{\alpha}_{0, e_1, 1}.$$

It is possible to compute the volume of the left hand term explicitly :

$$\begin{aligned} \text{vol}(\widetilde{C}) &= \int_{\mathbb{R}} \mathbf{1}_{x_1 \leq \cos(\alpha)} \left(\int_{\mathbb{R}^{d-1}} \mathbf{1}_{\|z\| \leq x_1 \sin(\alpha)} dz \right) dx_1 \\ &= \int_0^{\cos(\alpha)} V_{d-1} (\sin \alpha x_1)^{d-1}\ dx \\ &= V_{d-1} \frac{\sin^{d-1}(\alpha) \cos^d(\alpha)}{d}, \end{aligned}$$

where $V_{d-1} = \pi^{(d-1)/2} / \Gamma((d-1)/2+1)$ denotes the volume of the $d-1$ dimensional ball.

2. Proof of the second point The case where $x = x_0$ is trivial since $t \leq r$. Assume therefore $x \neq x_0$ and note that by definition, $C^{\pi/3}_{x, \overrightarrow{u}, t} \subset B_t(x)$. We will now show that $C^{\pi/3}_{x, \overrightarrow{u}, t} \subset B_r(x_0)$. Let $y \in C^{\pi/3}_{x, \overrightarrow{u}, t}$ and assume $y \neq x$ (if $y = x$ then $y \in B_r(x_0)$). Expanding the dot product

$$\begin{aligned} \|y - x_0\|^2 &= \|y - x\|^2 + 2(y - x) \cdot (x - x_0) + \|x - x_0\|^2 \\ &= \|y - x\|^2 - 2\|y - x\|\ \|x_0 - x\|\ \tfrac{y - x}{\|y - x\|} \cdot \overrightarrow{u} + \|x - x_0\|^2 \end{aligned}$$

$$\leq \|x-y\|^2 - \|x-y\|\,\|x-x_0\| + \|x-x_0\|^2,$$

where the last inequality comes from the definition of the cone and $\cos \pi/3 = \frac{1}{2}$. Let us distinguish two cases:

- if $t > \|x_0 - x\|$, we have $-\|x-y\|\|x_0-x\| \leq -t^2$ and hence $\|y-x_0\|^2 \leq t^2 \leq r^2$;
- otherwise $\|x-y\| \leq t \leq \|x_0-x\|$ and thus $\|y-x_0\|^2 \leq \|x-x_0\|^2 \leq r^2$.

In any case, $y \in B_r(x_0)$, which concludes the proof. □

E.1 Proof of Theorem 4

Proof of Theorem 4 Fix Ω as in Theorem 4. Let U be the uniform probability over Ω, i.e., $U(A) = \frac{\mathrm{vol}(A\cap\Omega)}{\mathrm{vol}(\Omega)}$ for any Borel-measurable set A. Let $\mathbb{P} = U^{\otimes n}$ over Ω^n. Throughout this proof, we will use the notation V_d to denote the volume of the d-dimensional unit ball (recall that $V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$).

Step 1. Covering Ω. Let $t > 0$. We say that a subset $\overline{X}$ of Ω is a t (interior) covering of Ω if $\Omega \subset \bigcup_{x\in\overline{X}} B_t(x)$. Denote with N_t the minimal cardinal $|\overline{X}|$ of a t interior covering of Ω and fix $\overline{X}_t$ a t interior covering of Ω whose cardinal is minimum, i.e., $|\overline{X}_t| = N_t$. Since the diameter of Ω is bounded by $2R$, it is known that $N_t \leq (1+2R/t)^d$
To prove this fact , one defines a maximal $t/2$-packing of Ω as a maximal set $\overline{Y}_{t/2} \subset \Omega$ such that the balls $B_{t/2}(\overline{y})$ are disjoint. It is then easy to check that if $\overline{Y}_{t/2}$ is a maximal $t/2$-packing, then it is also a t-covering and hence $N_t \leq |\overline{Y}_{t/2}|$. Finally, since Ω is included in a ball of radius $B_{2R}(x_0)$ for some $x_0 \in \mathbb{R}^d$ and since $\overline{Y}_{t/2} \subset \Omega$, it holds $\bigcup_{\overline{y}\in\overline{Y}_{t'}} B_t(\overline{y}) \subset B_{R+t/2}(x_0)$. Since the $B_t(\overline{y})$ are two by two disjoint, the result follows from the following equation:

$$|\overline{Y}|_{t/2}(t/2)^d V_d = \mathrm{vol}\left(\cup_{\overline{y}\in\overline{Y}_{t'}} B_t(\overline{y})\right) \leq \mathrm{vol}(B_{R+t/2}(x_0)) = (R+t/2)^d V_d.$$

Step 2. Probabilistic analysis. Note that for any $(x_1, ..., x_n) \in \Omega^n$, writing $\widehat{X} = \{x_1, .., x_n\}$, it holds:

$$\begin{aligned} h_{\widehat{X},\Omega} &= \max_{x\in\Omega} \min_{i\in[n]} \|x - x_i\| = \max_{\overline{x}\in\overline{X}_t} \max_{x\in B_t(\overline{x})\cap\Omega} \min_{i\in[n]} \|x - x_i\| \\ &\leq t + \max_{\overline{x}\in\overline{X}_t} \min_{i\in[n]} \|\overline{x} - x_i\|. \end{aligned}$$

Define E to be the following event :

$$E = \{(x_1, \ldots, x_n) \in \Omega^n \mid \max_{j\in[m]} \min_{i\in[n]} \|\overline{x}_j - x_i\| < t\}.$$

The n tuple $(x_1, .., x_n)$ belongs to E if for each $\overline{x} \in \overline{X}_t$ there exists at least one $i \in [n]$ for which $\|\overline{x} - x_i\| < t$. E can therefore be rewritten as follows :

$$E = \bigcap_{\overline{x} \in \overline{X}_t} \bigcup_{i \in [n]} \{(x_1, \ldots, x_n) \in \Omega^n \mid \|\overline{x} - x_i\| < t\}.$$

In particular, note that

$$E^c = \Omega^n \setminus E = \bigcup_{\overline{x} \in \overline{X}_t} \bigcap_{i \in [n]} \{(x_1, \ldots, x_n) \in \Omega^n \mid \|\overline{x} - x_i\| \geq t\} = \bigcup_{\overline{x} \in \overline{X}_t} (\Omega \setminus B_t(\overline{x}))^n.$$

Applying a union bound, we get

$$\begin{aligned} \mathbb{P}(E^c) &= \mathbb{P}\left(\bigcup_{\overline{x} \in \overline{X}_t} (\Omega \setminus B_t(\overline{x}))^n \right) \\ &\leq \sum_{\overline{x} \in \overline{X}_t} \mathbb{P}\big((\Omega \setminus B_t(\overline{x}))^n\big) = \sum_{j \in [m]} U(\Omega \setminus B_t(\overline{x})))^n, \end{aligned}$$

where the last step is due to the fact that $\mathbb{P}$ is a product measure and so $\mathbb{P}(A^n) = U^{\otimes n}(A^n) = U(A)^n$. Now we need to evaluate $U(\Omega \setminus B_t(\overline{x})) = 1 - U(B_t(\overline{x}))$ for $\overline{x} \in \overline{X_t}$. Since $\overline{X}_t \subset \Omega$, it holds

$$\forall \overline{x} \in \overline{X}_t,\ U(B_t(\overline{x})) = \tfrac{\mathrm{vol}(B_t(\overline{x}) \cap \Omega)}{\mathrm{vol}(\Omega)} \geq \tfrac{\min_{x \in \Omega} \mathrm{vol}(B_t(x) \cap \Omega)}{\mathrm{vol}(\Omega)}.$$

Step 3. Bounding $\mathrm{vol}(B_t(x) \cap \Omega)$ **when** $t \leq r$. Let us now find a lower bound for $\min_{x \in \Omega} \mathrm{vol}(B_t(x) \cap \Omega)$. Recall that since Ω satisfies Assumption 1(a), Ω can be written $\Omega = \cup_{z \in S} B_r(z)$. .Let $t \leq r$, $x \in \Omega$. By the previous point, there exists $z \in S$ such that $x \in B_r(z) \subset \Omega$ and hence $B_t(x) \cap B_r(z) \subset B_t(x) \cap \Omega$. Let $C_{x,z,t}$ denote the cone centered in x and directed to z with aperture $\pi/3$. It is easy to see geometrically that $B_r(z) \cap B_t(x)$ contains the cone $C_{x,z,t}$ (this fact is proved in Lemma 13). Moreover, using the lower bound for the volume of this cone provided in Lemma 13, it holds:

$$\begin{aligned} \mathrm{vol}(\Omega \cap B_t(x)) &\geq \mathrm{vol}(B_r(z) \cap B_t(x)) \geq \mathrm{vol}(C_{x,z,t}) \\ &\geq \frac{2V_{d-1}}{\sqrt{3}d} \left(\frac{\sqrt{3}}{4}\right)^d t^d. \end{aligned}$$

Step 4. Expressing t **with respect to** n **and** δ **and guaranteeing that** $t \leq r$. To conclude, let $C = \frac{V_{d-1}}{2d\mathrm{vol}(\Omega)} \left(\frac{\sqrt{3}}{4}\right)^{d-1}$. Since $N_t \leq (1 + 2R/t)^d$, and $(1 - c)^x \leq e^{-cx}$ for any $x \geq 0$ and $c \in [0, 1]$, then

$$\mathbb{P}(E) \geq 1 - N_t\big(1 - Ct^d\big)^n \geq 1 - e^{-Ct^d n + d\log(1+2R/t)} \geq 1 - \delta,$$

where the last step is obtained by setting

$$t = (Cn)^{-1/d} \left(\log \frac{(1 + 2R(Cn)^{1/d})^d}{\delta} \right)^{1/d}.$$

Then $h_{\widehat{X},\Omega} \leq 2t$ with probability at least $1 - \delta$, when $t \leq r$. The desired result is obtained by further bounding C and t as follows.

Bounding C. It holds $\frac{2V_{d-1}}{\sqrt{3}dV_d} = \left(\frac{4}{3d^2\pi}\right)^{1/2} \frac{\Gamma(d/2+1)}{\Gamma(d/2+1/2)}$. Using Gautschi's inequality and the fact that $d \geq 1$,

$$\left(\tfrac{2}{3d\pi}\right)^{1/2} \leq \frac{2V_{d-1}}{\sqrt{3}dV_d} \leq \left(\tfrac{2(d+2)}{3d^2\pi}\right)^{1/2} \leq 1.$$

Since $\left(\frac{3d\pi}{2}\right)^{1/2d} \frac{4}{\sqrt{3}} \leq 2\sqrt{2\pi}$ for all $d \geq 1$, and since $V_d r^d \leq \mathrm{vol}(\Omega) \leq V_d R^d$, it holds

$$(2\sqrt{2\pi}R)^{-d} \leq C \leq (4r/\sqrt{3})^{-d} \implies \frac{n^{1/d}}{2\sqrt{2\pi}R} \leq (Cn)^{1/d} \leq \frac{\sqrt{3}n^{1/d}}{4r} \leq \frac{n^{1/d}}{2r}.$$

Bounding t. Since, $(1 + x)^d \leq (2x)^d$ for any $x \geq 1$ and $2R(Cn)^{1/d} \leq \frac{R}{r} n^{1/d}$, and $\frac{R}{r} n^{1/d} \geq 1$, it holds

$$t \leq 2\sqrt{2\pi}Rn^{-1/d}(\log \tfrac{n}{\delta} + d \log \tfrac{2R}{r})^{1/d}.$$

Guaranteeing $t \leq r$. Applying Lemma 12 to $\alpha = (2\pi)^{d/2}(2R/r)^d$ and $\beta = (2R/r)^d/\delta$, it holds that if

$$n \geq 2\alpha \log(2\alpha\beta) = 2\,(2\pi)^{d/2}(2R/r)^d \left(\log \frac{2}{\delta} + d/2 \log(2\pi) + 2d \log(2R/r) \right),$$

then $\alpha/n \log(\beta n) \leq 1$, so

$$t \leq 2\sqrt{2\pi}Rn^{-1/d}(\log \tfrac{n}{\delta} + d \log \tfrac{2R}{r})^{1/d} \leq r(\alpha/n \log(\beta n)^{1/d} \leq r.$$

□

E.2 Proof of Theorem 6

Proof Recall that $s > d/2$ and $m < s - \frac{d}{2}$ is a positive integer. Assume that Ω satisfies Assumption 1(a) for a certain r and that the diameter of Ω is bounded by $2R$. In particular, if Ω is a ball of radius R, then Ω satisfies Assumption 1(a) with $r = R$. In the first step of the proof we guarantee that n is large enough to apply Theorem 4 and that $h_{\widehat{X},\Omega}$, controlled by Theorem 4, satisfies the assumptions of Theorem 5. Then we apply Theorem 5.

Step 1. Guaranteeing n large enough and $h_{\widehat{X},\Omega} \leq r/(18(m-1)^2)$. Applying Lemma 12 to $\alpha = \left(\frac{2R}{r}\right)^d \max(3, 10(m-1))^{2d}$ and $\beta = \frac{(2R)^d}{r^d\,\delta}$, it holds that if

$$n \geq 2\alpha \log(2\alpha\beta) = \left(\frac{2R}{r}\right)^d \max(3, 10(m-1))^{2d} \left(2\log\frac{2}{\delta} + 4d\log\left(\tfrac{R}{r}\max(6, 20(m-1))\right)\right),$$

then $\alpha/n \log(\beta n) \leq 1$, which implies

$$n^{-1/d}(\log\frac{n}{\delta} + d\log\beta)^{1/d} \leq \frac{r}{2R\max(3, 10(m-1))^2}.$$

In particular, n satisfying the condition above is large enough to satisfy the requirement of Theorem 4 (since $r \leq R$). Therefore, by applying Theorem 4 we have that with probability at least $1-\delta$,

$$h_{\widehat{X},\Omega} \leq 11R\,n^{-\frac{1}{d}}(\log\tfrac{(2R)^d\,n}{r^d\,\delta})^{1/d} \leq \frac{r}{\max(1, 18(m-1)^2)}.$$

Step 2. Applying Theorem 5. In the previous step we provided a condition on n such that $h_{\widehat{X},\Omega}$ satisfies $h_{\widehat{X},\Omega} \leq \frac{r}{\max(1,18(m-1)^2)}$. By Proposition 1, Assumption 2 holds for the Sobolev kernel with smoothness s, for any $m \in \mathbb{N}$ since $m < s - d/2$. Then the conditions to apply Theorem 5 are satisfied. Applying Theorem 5 with $\lambda \geq 2\eta\max(1, \mathsf{MD}_m)$ and $\eta = \frac{3\max(1,18(m-1)^2)^m\,d^m}{m!}h^m_{\widehat{X},\Omega}$, we have

$$|\hat{c} - f_*| \leq 2\eta|f|_{\Omega,m} + \lambda\mathrm{Tr}(A_*) \leq 3\lambda(|f|_{\Omega,m} + \mathrm{Tr}(A_*)),$$

Thus, under this condition, we have with probability at least $1-\delta$,

$$|\hat{c} - f_*| \leq C_{m,s,d}R^m n^{-m/d}(\log\frac{2^d n}{\delta}),$$

where

$$C_{m,s,d} = 6\times 11^m \times \frac{\max(1, 18(m-1)^2)^m d^m}{m!}\max(1, \mathsf{MD}_m).$$

Step 3. Bounding the constant term $C_{m,s,d}$ in terms of m, s, d. Note that

$$\frac{\Gamma(m+d/2)}{\Gamma(d/2)} = (d/2)...(d/2+m-1) \leq (d/2+m-1)^{m-1}$$

and

$$\frac{\Gamma(s-d/2-m)}{\Gamma(s-d/2)} = \frac{1}{(s-d/2-m)....(s-d/2-1)} \leq \left(\frac{1}{s-d/2-m}\right)^{m-1},$$

which yields:

$$\mathsf{D}_m \leq (2\pi)^{d/4} \left(\frac{d/2 + m - 1}{s - d/2 - m} \right)^{(m-1)/2}.$$

Moreover, using the bound on M, we get

$$\mathsf{D}_m \mathsf{M} \leq 2^{s+1/2} \, (2\pi)^{3d/4} \left(\frac{d/2 + m - 1}{s - d/2 - m} \right)^{(m-1)/2}.$$

This yields the following bound for $C_{m,s,d}$:

$$C_{m,s,d} \leq \frac{6 \max(1, 18(m-1)^2)^m (11d)^m}{m!} \max \left(1, 2^{s+1/2} \, (2\pi)^{3d/4} \left(\frac{d/2 + m - 1}{s - d/2 - m} \right)^{(m-1)/2} \right).$$

□

F Global minimizer. Proofs

F.1 Proof of Remark 4

Proof Since f satisfies both Assumptions 1(b) and 4, denote by ζ the unique minimizer of f in Ω. Since ζ is a strict minimum by Assumption 1(b), there exists $\beta_1 > 0$ such that $\nabla^2 f(\zeta) \succeq \beta_1 I$. Thus, since $f \in C^2(\mathbb{R}^d)$, there exists a small radius $t > 0$ such that $\nabla^2 f(x) \succeq \frac{\beta_1}{2} I$ for all $x \in B_t(\zeta)$ and hence

$$\forall x \in \Omega \cap B_t(\zeta), \;\; f(x) - f_* = f(x) - f(\zeta) - \nabla f(\zeta)^\top (x - \zeta) \geq \tfrac{\beta_1}{4} \|x - \zeta\|^2. \tag{F.1}$$

Moreover, since f has no minimizer on the boundary of Ω and since ζ is the unique minimizer of f on Ω, f has no minimizer on $K = \overline{\Omega} \setminus B_t(x)$ which is a compact set. Denote by m the minimum of f on K. Since K is compact, this minimum is reached and since f does not reach its global minimum f_* on K, we have $m - f_* > 0$. Let R be a radius such that $\overline{\Omega} \subset B_R(\zeta)$, which exists since Ω is bounded. Then, since for any $x \in \overline{\Omega}$, $\|x - \zeta\| < R$, it holds for any $x \in K$:

$$f(x) - f_* = f(x) - m + m - f_* \geq m - f_* = \frac{2(m - f_*)}{2R^2} R^2 \geq \frac{2(m - f_*)}{2R^2} \|x - \zeta\|^2. \tag{F.2}$$

Thus, taking $\beta = \min(\frac{\beta_1}{2}, \frac{2(m-f_*)}{R^2})$ and combining Eqs. (F.1) and (F.2), it holds

$$\forall x \in \Omega, \;\; f(x) - f_* \geq \frac{\beta}{2} \|x - \zeta\|^2.$$

□

F.2 Proof of Theorem 7

Proof Let us divide the proof into four steps.

Step 1: Extending the parabola outside of Ω Since Ω is an open set containing ζ, there exists $t > 0$ such that the ball $B_t(\zeta) \subset \Omega$. Define $\delta = \frac{\beta-\nu}{2}t^2$. It holds :

$$\forall x \in \mathbb{R}^d \setminus \Omega,\ \frac{\beta}{2}\|x-\zeta\|^2 \geq \frac{\nu}{2}\|x-\zeta\|^2 + \delta. \tag{F.3}$$

Now define the following open set :

$$\widetilde{\Omega} = \left\{x \in \mathbb{R}^d\ :\ f(x) - f_* - \tfrac{\beta}{2}\|x-\zeta\|^2 > -\delta/2\right\}.$$

It is open since f is continuous. Moreover, it contains the closure of Ω which we denote with $\overline{\Omega}$ which is compact since it is closed and bounded in $\mathbb{R}^d$. Theorem 1.4.2 in [40] applied to $X = \widetilde{\Omega}$ and $K = \overline{\Omega}$ shows the existence of $\chi : \mathbb{R}^d \to \mathbb{R}$ such that $\chi \in C^\infty(\mathbb{R}^d)$, $\chi(x) \in [0,1]$, $\chi = 1$ on $\overline{\Omega}$ and $\chi = 0$ on $\mathbb{R}^d \setminus \widetilde{\Omega}$. Finally, define $\overline{p}_\nu(x) := \frac{\nu}{2}\|x-\zeta\|^2\chi(x)$. $\overline{p}_\nu$ satisfies the following properties :

- $\overline{p}_\nu \in C^\infty(\mathbb{R}^d)$;
- for all $x \in \overline{\Omega}$, $\overline{p}_\nu(x) = \frac{\nu}{2}\|x-\zeta\|^2 \leq \frac{\beta}{2}\|x-\zeta\|^2$;
- for all $x \in \mathbb{R}^d \setminus \widetilde{\Omega}$, $\overline{p}_\nu(x) = 0$;
- for all $x \in \widetilde{\Omega} \setminus \Omega$, $f(x) - f_* - \overline{p}_\nu(x) \geq \delta/2$.

The first, second and third properties are direct consequences of the properties of χ and the fact that $\nu < \beta$. The last property comes from combining Eq. (F.3) with the definition of $\widetilde{\Omega}$ and the fact that $\chi \in [0,1]$:

$$\begin{aligned}
\forall x \in \widetilde{\Omega} \setminus \Omega,\ f(x) - f_* - \overline{p}_\nu(x) &= f(x) - f_* - \chi(x)\tfrac{\nu}{2}\|x-\zeta\|^2 \\
&\geq f(x) - f_* - \tfrac{\nu}{2}\|x-\zeta\|^2 \\
&= \left(f(x) - f_* - \tfrac{\beta}{2}\|x-\zeta\|^2\right) \\
&\quad + \left(\tfrac{\beta}{2}\|x-\zeta\|^2 - \tfrac{\nu}{2}\|x-\zeta\|^2\right) \\
&\geq -\delta/2 + \delta = \delta/2.
\end{aligned}$$

Step 2: Extending $x \mapsto f(x) - \frac{\nu}{2}\|x-\zeta\|^2$ outside of Ω. Define $g(x) = f(x) - \overline{p}_\nu(x)$ on $\mathbb{R}^d$. Then g satisfies Assumption 1(b), g has exactly one minimizer in Ω which is ζ, and its minimum is $g(\zeta) = f_*$. Indeed, the fact that $g \in C^2(\mathbb{R}^d)$ comes from the fact that $f \in C^2(\mathbb{R}^d)$ by Assumption 1(b) on f and the fact that $\overline{p}_\nu \in C^\infty(\mathbb{R}^d)$. Moreover, $g \geq f_*$ on $\mathbb{R}^d$ and $g - f_* \geq \delta/2$ on $\partial\Omega$. Indeed, first note that since $\nu < \beta$, it holds

$$\forall x \in \Omega,\ g(x) = f(x) - \overline{p}_\nu(x) = f(x) - \tfrac{\nu}{2}\|x-\zeta\|^2 \geq f(x) - \tfrac{\beta}{2}\|x-\zeta\|^2 \geq f_*,$$

where the last inequality comes from Eq. (7.2). Second, since $\overline{p}_\nu = 0$ on $\mathbb{R}^d \setminus \widetilde{\Omega}$ and since f_* is the minimum of f, for any $x \in \mathbb{R}^d \setminus \widetilde{\Omega}$, $g(x) - f_* = f(x) - f_* \geq 0$. Finally, by the last point of the previous step, we see that $g(x) \geq f_* + \delta/2 > f_*$ for any $x \in \widetilde{\Omega} \setminus \Omega$. In particular, $g(x) \geq f_* + \delta/2$ for any $x \in \partial\Omega$. Since $g(\zeta) = f(\zeta) = f_*$, we see that f_* is the minimum of g on $\mathbb{R}^d$ and that this minimum is reached at ζ and is not reached on the boundary of Ω. The fact that ζ is the unique minimum on Ω comes from the fact that since $\nu < \beta$ and by Eq. (7.2) we have that for any $x \in \Omega \setminus \{\zeta\}$ the following holds

$$\begin{aligned} g(x) &= f(x) - \overline{p}_\nu(x) = f(x) - \tfrac{\nu}{2}\|x - \zeta\|^2 \\ &> f(x) - \tfrac{\beta}{2}\|x - \zeta\|^2 \geq f_*. \end{aligned} \tag{F.4}$$

The fact that this minimum is not reached on the boundary of Ω comes from the fact stated above that $g(x) \geq f_* + \delta/2$ for any $x \in \partial\Omega$. Finally, the fact that ζ is a strict minimum of g also comes from Eq. (F.4) which implies that $\nabla^2 g(\zeta) \succeq (\beta - \nu)I$ since g reaches a minimum in ζ, g is C^2 and $\nu < \beta$.

Note that g also satisfies Assumption 3 since f satisfies Assumption 3 and $\overline{p}_\mu \in C^\infty(\mathbb{R}^d) \subset C^2(\mathbb{R}^d) \cap \mathcal{H}$ by Assumption 2(a).

Step 3: Applying Corollary 1 **to** g. The previous point shows that g satisfies Assumptions 1(b) and 3 and that g has a unique minimum in Ω. Moreover, $\mathcal{H}$ satisfies Assumption 2. Hence, Corollary 1 to g and $\mathcal{H}$, the following holds : there exists $A_* \in \mathbb{S}_+(\mathcal{H})$ with $\operatorname{rank}(A_*) \leq d + 1$ such that $g(x) - f^* = \langle \phi(x), A_* \phi(x) \rangle$ for all $x \in \Omega$.

Step 4. Let p_0 be the maximum of Eq. (7.1). In Lemma 5 we have seen that the solution of Eq. (7.1) is $p_0 = f_*$. Since $A \succeq 0$ implies $\langle \phi(x), A\phi(x) \rangle \geq 0$ for all $x \in \Omega$, the problem in Eq. (7.1) is a relaxation of Eq. (7.3), where the constraint $f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top z - c = \langle \phi(x), A\phi(x) \rangle$ is substituted by $f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top z - c \geq 0, \forall x \in \Omega$. Then $p_0 \geq p^*$ if a maximum p^* exists for Eq. (7.3). Thus, if there exists A that satisfies the constraints in Eq. (7.3) for the value $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$ and $z_* = \zeta$, then $p_0 = p^*$ and (c_*, ζ, A) is a minimizer of Eq. (7.3).
The proof is concluded by noting that indeed there exists A that satisfies the constraints in Eq. (7.3) for the value $c_* = f_* + \frac{\nu}{2}\|\zeta\|^2$ and $z_* = \zeta$ and it is obtained by the previous step. □

F.3 Proof of Theorem 8

Proof The proof is a variation of the the one for Theorem 5, the main difference is that we take care of the additional term $z - \zeta$.

Step 0. The SDP problem in Eq. (7.4) **admits a solution**
(a) Under the constraints of Eq. (7.4), $c - \frac{\nu}{2}\|z\|^2$ cannot be larger than $\min_{i \in [n]} f(x_i)$. Indeed, for any $i \in [n]$, since $B \succeq 0$, the i-th constraint implies

$$f(x_i) - \tfrac{\nu}{2}\|x_i - z\|^2 - c + \frac{\nu}{2}\|z\|^2 = f(x_i) - \tfrac{\nu}{2}\|x_i\|^2 + \nu x_i^\top z - c = \Phi_i B \Phi_i \geq 0.$$

Hence, $f(x_i) \geq f(x_i) - \frac{\nu}{2}\|x_i - z\|^2 \geq c + \frac{\nu}{2}\|z\|^2$. Thus, since $B \succeq 0$, for any B, z, c satisfying the constraint, $c - \frac{\nu}{2}\|z\|^2 - \lambda \mathrm{Tr}(B) \leq \max_{i\in[i]} f(x_i)$.

(b) There exists an admissible point. Indeed let (c_*, z_*, A_*) be the solution of Eq. (7.3) such that A_* has minimum trace norm (by Theorem 7, we know that this solution exists with $c_* = f_*$ and $z_* = \zeta$, under Assumptions 1 to 4). Then, by Lemma 3 applied to $g(x) = f(x) - \frac{\nu}{2}\|x\|^2 - \nu x^\top z_* - c_*$ and $A = A_*$, given $\widehat{X} = \{x_1, \ldots, x_n\}$ we know that there exists $\overline{B} \in \mathbb{S}_+(\mathbb{R}^n)$ satisfying $\mathrm{Tr}(\overline{B}) \leq \mathrm{Tr}(A_*)$ s.t. the constraints of Eq. (7.4) are satisfied for $c = c_*$ and $z = z_*$. Then $(c_*, z_*, \overline{B})$ is admissible for the problem in Eq. (7.4). Since there exists an admissible point for the constraints of Eq. (7.4) and its functional cannot be larger than $\max_{i\in[n]} f(x_i)$, then the SDP problem in Eq. (7.4) admits a solution [21].

Step 1. Consequences of existence of A_*. Let $(\hat{c}, \hat{z}, \hat{B})$ one minimizer of Eq. (7.4). The existence of the admissible point $(c_*, z_*, \overline{B})$ implies that

$$\hat{c} - \tfrac{\nu}{2}\|\hat{z}\|^2 - \lambda \mathrm{Tr}(\hat{B}) \geq c_* - \tfrac{\nu}{2}\|z_*\|^2 - \lambda \mathrm{Tr}(\overline{B}) \geq f_* - \lambda \mathrm{Tr}(A_*). \tag{F.5}$$

From which we derive,

$$\lambda \mathrm{Tr}(\hat{B}) - \lambda \mathrm{Tr}(A_*) \;\leq\; \Delta, \quad \Delta := \hat{c} - \tfrac{\nu}{2}\|\hat{z}\|^2 - f_*. \tag{F.6}$$

Step 2. L^∞ **bound due to the scattered zeros.** Note that the solution $(\hat{c}, \hat{z}, \hat{B})$ satisfies $\hat{g}(x_i) = \Phi_i^\top \hat{B} \Phi_i$ for $i \in [n]$, where the function $\hat{g}$ is defined as $\hat{g}(x) = f(x) - \frac{\nu}{2}\|x\|^2 + \nu x^\top \hat{z} - \hat{c}$ for $x \in \Omega$, moreover $h_{\widehat{X},\Omega} \leq \frac{r}{\max(1,18(m-1)^2)} = \frac{r}{18(m-1)^2}$ by assumption, since $m \geq 2$. Then we can apply Theorem 4 with $g = \hat{g}$, $\tau = 0$ and $B = \hat{B}$ obtaining for all $x \in \Omega$

$$f(x) - \tfrac{\nu}{2}\|x\|^2 + \nu x^\top \hat{z} - \hat{c} = \hat{g}(x) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \mathrm{Tr}(\hat{B})), \quad \eta = C_0 h^m_{\widehat{X},\Omega},$$

where C_0 is defined in Theorem 4 and $C_0 = 3\frac{(18d)^m (m-1)^{2m}}{m!}$ since $m \geq 2$. Since the inequality above holds for any $x \in \Omega$, by evaluating it in the global minimizer $\zeta \in \Omega$, we have $f(\zeta) = f_*$ and so

$$-\Delta - \tfrac{\nu}{2}\|\hat{z} - \zeta\|^2 = \hat{g}(\zeta) \geq -\eta(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \mathrm{Tr}(\hat{B})).$$

Now we bound $|\hat{g}|_{\Omega,m}$. Since $\hat{g}(x) = f(x) - p_{\hat{z},\hat{c}}(x)$, where $p_{\hat{z},\hat{c}}$ is a second degree polynomials defined as $p_{\hat{z},\hat{c}} = \frac{\nu}{2}\|x\|^2 - \nu x^\top \hat{z} + \hat{c}$, we have

$$|\hat{g}|_{\Omega,m} \leq |f|_{\Omega,m} + |p_{\hat{z},\hat{c}}|_{\Omega,m} \leq |f|_{\Omega,m} + \nu, \tag{F.7}$$

since for $m = 2$, we have $|p_{\hat{z},\hat{c}}|_{\Omega,2} = \sup_{i,j\in[d],x\in\Omega} |\frac{\partial^2 p_{\hat{z},\hat{c}}(x)}{\partial x_i \partial x_j}| = \nu$ and also $|p_{\hat{z},\hat{c}}|_{\Omega,m} = 0$ for $m > 2$. Then

$$\Delta \leq \Delta + \tfrac{\nu}{2}\|\hat{z} - \zeta\|^2 \leq \eta |f|_{\Omega,m} + \eta \mathsf{MD}_m \mathrm{Tr}(\hat{B}) + \eta\nu. \tag{F.8}$$

Conclusion. Combining Eq. (F.8) with Eq. (F.6), since $\frac{\nu}{2}\|\hat{z}-\zeta\|^2 \geq 0$ and since $\lambda \geq 2\mathsf{MD}_m\eta$ by assumption, we have

$$\tfrac{\lambda}{2}\mathrm{Tr}(\hat{B}) \leq (\lambda - \mathsf{MD}_m\eta)\mathrm{Tr}(\hat{B}) \leq \eta|f|_{\Omega,m} + \eta\nu + \lambda\mathrm{Tr}(A_*),$$

from which we obtain Eq. (7.7). Moreover, the inequality Eq. (7.6) is derived by bounding Δ from below as $\Delta \geq -\lambda\mathrm{Tr}(A_*)$ by Eq. (F.6), since $\mathrm{Tr}(\hat{B}) \geq 0$ by construction, and bounding it from above as

$$\Delta \leq 2\eta|f|_{\Omega,m} + 2\eta\nu + \lambda\mathrm{Tr}(A_*),$$

that is obtained by combining Eq. (F.8) with Eq. (7.7) and with the assumption $\mathsf{MD}_m\eta \leq \lambda/2$. Finally from Eq. (F.8) we obtain

$$\tfrac{\nu}{2}\|\hat{z}-\zeta\|^2 \leq |\Delta| \;+\; \eta|f|_{\Omega,m} + \eta\mathsf{MD}_m\mathrm{Tr}(\hat{B}) + \eta\nu,$$

from which we derive the bound $\frac{\nu}{2}\|\hat{z}-\zeta\|^2$ in Eq. (7.5), by bounding $|\Delta|$ and $\mathrm{Tr}(\hat{B})$ via Eq. (7.6) and Eq. (7.7). □

G Proofs for the extensions

G.1 Proof of Theorem 9

Proof Let $(\hat{c}, \hat{B})$ be a minimum trace-norm solution of Eq. (2.4). The minimum $p_{\lambda,n}$ of Eq. (2.4) then corresponds to $p_{\lambda,n} = \hat{c} - \lambda\mathrm{Tr}(\hat{B})$. Combining Eq. (8.1) with Eq. (5.7) from the proof of Theorem 5 and the fact that $\theta_2 \leq \lambda/8$, we have that

$$\tfrac{7}{8}\lambda\mathrm{Tr}(\tilde{B}) - \lambda\mathrm{Tr}(A_*) - \theta_1 \leq \tilde{\Delta}, \quad \tilde{\Delta} := \tilde{c} - f_*. \tag{G.1}$$

Analogously to Step 3 of the proof of Theorem 5, by applying Theorem 4 to Eq. (8.2) with $g(x) = f(x) - \tilde{c}$, $B = \tilde{B}$ and $\tau = \tau_1 + \tau_2\mathrm{Tr}(\tilde{B})$, we obtain for any $x \in \Omega$

$$f(x) - \tilde{c} \;\geq\; -2\tau_1 - 2\tau_2\mathrm{Tr}(\tilde{B}) \;-\; \eta(|g|_{\Omega,m} + \mathsf{MD}_m\mathrm{Tr}(\tilde{B})), \qquad \eta = C_0 h^m_{\hat{X},\Omega},$$

with C_0 defined in Theorem 4. Now evaluating the inequality above for $x = \zeta$, noting that $|g|_{\Omega,m} = |f|_{\Omega,m}$ since $m \geq 1$, and considering that by assumption $\tau_2 \leq \lambda/8$ and $\mathsf{MD}_m\eta \leq \lambda/2$ we derive

$$\tilde{\Delta} = -(f(\zeta) - \tilde{c}) \leq 2\tau_1 + \tfrac{3}{4}\lambda\mathrm{Tr}(\tilde{B}) + \eta|f|_{\Omega,m}. \tag{G.2}$$

The desired result is obtained by combining Eq. (G.2) and Eq. (G.1) as we did in Step 3 of Theorem 5. □

G.2 Proof of Corollary 2

Proof Define $\mathcal{H} = \{g \in C^s(\Omega) \ : \ \exists f \in C^s(\mathbb{R}^d), \ f|_\Omega = g\}$, endowed with the following norm :

$$\forall g \in \mathcal{H}, \ \|g\|_{\mathcal{H}} = \sup_{|\alpha| \leq s} \sup_{x \in \Omega} \|\partial^\alpha g(x)\|.$$

Note that this norm is well defined since for any $g \in \mathcal{H}$, since there exists $f \in C^s(\mathbb{R}^d)$ such that $g = f|_\Omega$, since all the derivatives of f are continuous hence bounded on Ω which is bounded, so are all the derivatives of g.

Now note that $\mathcal{H}$ satisfies Assumptions 2(a) to 2(c). Indeed, given $u, v \in \mathcal{H}$ the first assumption is satisfied as a simple consequence of the Leibniz formula, since for any $x \in \Omega$, $\partial^\alpha(u \cdot v)(x) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} \partial^\beta u(x) \partial^{\alpha-\beta} v(x)$ which in turn implies that for any $|\alpha| \leq s$ and $x \in \Omega$, $\|\partial^\alpha(u \cdot v)(x)\| \leq 2^{|\alpha|} \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$ and hence $\|u \cdot v\|_{\mathcal{H}} \leq 2^s \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$. Assumption 2(b) is trivially satisfied and Assumption 2(c) is a simple consequence of the dominated convergence theorem. Indeed, if $u \in \mathcal{H}$ and $\overline{u} \in C^s(\mathbb{R}^d)$ such that $\overline{u}|_\Omega = u$, define

$$\forall x, z \in \mathbb{R}^d, \ \overline{v}_z(x) = \int_0^1 (1-t)\overline{u}(z + t(x-z))dt.$$

$\overline{v}_z$ is in $C^s(\mathbb{R}^d)$ by dominated convergence, and $v_z = \overline{v}|_\Omega$ satisfies the desired property (in this case, there is no need to depend on r and one can simply take $g_{r,z} = v_z$).

Moreover, if $f \in C^{s+2}(\mathbb{R}^d)$, then in particular, for any $i, j \in [d]$, $\frac{\partial f}{\partial x_i \partial x_j} \in C^s(\mathbb{R}^d)$ and hence its restriction to Ω is in $\mathcal{H}$. Moreover, in that case, it is obvious that since $s \geq 0$, $f|_\Omega \in \mathcal{H}$. This shows that f satisfies Assumptions 1(b) and 3.

Therefore, Theorem 2 can be applied, and there exist $\tilde{w}_1, \ldots, \tilde{w}_p \mathcal{H}$, $p \in \mathbb{N}_+$, such that

$$\forall x \in \Omega, \ f(x) - f_* = \sum_{j \in [p]} w_j^2(x).$$

By definition of $\mathcal{H}$, taking $w_1, ..., w_p$ such that $w_j|_\Omega = \tilde{w}_j$, the corollary holds. □

G.3 Certificate of optimality for the global minimizer candidate of Eq. (7.4)

Theorem 12 (Certificate of optimality for Eq. (7.4)). *Let Ω satisfy Assumption 1(a) for some $r > 0$. Let k be a kernel satisfying Assumptions 2(a) and 2(d) for some $m \geq 2$. Let $\widehat{X} = \{x_1, \ldots, x_n\} \subset \Omega$ with $n \in \mathbb{N}$ such that $h_{\widehat{X},\Omega} \leq \frac{r}{18(m-1)^2}$. Let $f \in C^m(\Omega)$ and let $\hat{c} \in \mathbb{R}$, $\hat{z} \in \mathbb{R}^d$, $\hat{B} \in \mathbb{S}_+(\mathbb{R}^n)$ and $\tau \geq 0$ satisfying*

$$|f(x_i) - \tfrac{\nu}{2}\|x_i\|^2 + \nu x_i^\top \hat{z} - \hat{c} \ - \ \Phi_i^\top \hat{B} \Phi_i| \leq \tau, \quad i \in [n] \tag{G.3}$$

where Φ_i are defined in Sect. 2. Let $f_* = \min_{x\in\Omega} f(x)$ and $\hat{f} = \hat{c} - \frac{\nu}{2}\|\hat{z}\|^2$. Then,

$$|f(\hat{z}) - f_*| \le f(\hat{z}) - \hat{f} + 2\tau + C_1 h^m_{\widehat{X},\Omega}, \tag{G.4}$$

$$\tfrac{\nu}{2}\|\zeta - \hat{z}\|^2 \le f(\hat{z}) - \hat{f} + 2\tau + C_2 h^m_{\widehat{X},\Omega}. \tag{G.5}$$

and $C_1 = C_0(|f|_{\Omega,m} + \mathsf{MD}_m Tr(\hat{B}) + \mathsf{MD}_m \hat{C})$, $C_2 = C_0(|f|_{\Omega,m} + \nu + \mathsf{MD}_m Tr(\hat{B}))$, where $\hat{C} = \frac{\nu}{2}\|R^{-\top}(X - 1_n\hat{\zeta}^\top)\|^2$, with $X \in \mathbb{R}^{n\times d}$ the matrix whose i-th row corresponds to the point x_i and $1_n \in \mathbb{R}^n$ the vector where each element is 1. The constants C_0, defined in Theorem 4, and m, M, D_m, defined in Assumptions 2(a) and 2(d), do not depend on n, $\widehat{X}$, $h_{\widehat{X},\Omega}$, $\hat{c}$, $\hat{B}$ or f.

Proof We divide the proof in two steps

Step 1. First note that

$$\hat{g}(x) := f(x) - \tfrac{\nu}{2}\|x\|^2 + \nu x^\top \hat{z} - \hat{c} = f(x) - \tfrac{\nu}{2}\|x - \hat{z}\|^2 - \hat{f}.$$

By applying Theorem 4 with $g = \hat{g}$ and $B = \hat{B}$ we have that for any $x \in \Omega$ $f(x) - \frac{\nu}{2}\|x - \hat{z}\|^2 - \hat{f} = \hat{g}(x) \ge -\varepsilon - 2\tau$, where $\varepsilon = C_0(|\hat{g}|_{\Omega,m} + \mathsf{MD}_m \mathrm{Tr}(\hat{B}))h^m_{\widehat{X},\Omega}$ and C_0 is defined in Theorem 4. In particular this implies that

$$f(\zeta) - \hat{f} - \tfrac{\nu}{2}\|x - \hat{z}\|^2 \ge -\varepsilon - 2\tau,$$

from which Eq. (G.5) is obtained by considering that $f(\hat{z}) \ge f(\zeta)$ since ζ is a minimizer of f. To conclude the proof of Eq. (G.5) note that $|\hat{g}|_{\Omega,m} \le |f|_{\Omega,m} + \nu$ since $m \ge 2$.

Step 2. Now to obtain Eq. (G.4) we need to do a slightly different construction. Let $u_j(x) = e_j^\top(x - \hat{z})$ for any $x \in \Omega$. Note that since u_j is the restriction to Ω of a C^∞ function on $\mathbb{R}^d$, by Assumption 2(a), $u_j \in \mathcal{H}$. Moreover, note that $\frac{\nu}{2}\|x - \hat{z}\|^2 = \frac{\nu}{2}\sum_{j=1}^d u_j(x)^2$. Take $\hat{u}_j \in \mathbb{R}^n$ defined as $\hat{u}_j = V^* u_j$ and note that

$$\Phi_i^\top \hat{u}_j = \langle V\phi(x_i), V^* u_j\rangle = \langle V^* V\phi(x_i), u_j\rangle = \langle P\phi(x_i), u_j\rangle = u_j(x_i).$$

Then, defining $\hat{G} = \frac{\nu}{2}\sum_{i=1}^d \hat{u}_j\hat{u}_j^\top \in \mathbb{S}_+(\mathbb{R}^n)$ we have

$$\tfrac{\nu}{2}\|x_i - \hat{z}\|^2 = \Phi_i^\top \hat{G}\Phi_i, \qquad \forall i \in [n].$$

Substituting $-\frac{\nu}{2}\|x_i\|^2 + \nu x_i^\top \hat{z}$ with $\frac{\nu}{2}\|\hat{z}\|^2 - \Phi_i^\top \hat{G}\Phi_i$ in the inequality in Eq. (G.3), we obtain

$$|f(x_i) - \hat{f} - \Phi_i^\top(\hat{B} + \hat{G})\Phi_i^\top| \le \tau, \quad \forall i \in [n].$$

By applying Theorem 4 with $g(x) = f(x) - \hat{f}$ and $B = \hat{B} + \hat{G}$ we have that $f(x) - \hat{f} \ge -\varepsilon - 2\tau$ for all $x \in \Omega$, where $\varepsilon = C' h^m_{\widehat{X},\Omega}$ with $C' = C_0(|g|_{\Omega,m} + \mathsf{MD}_m \mathrm{Tr}(\hat{B} + \hat{G}))$.

In particular, $f(\zeta) - \hat{f} \geq -\varepsilon - 2\tau$, from which Eq. (G.4) is obtained considering that $f(\hat{z}) \geq f_*$ since ζ is a minimizer of f.

Finally, note that $|g|_{\Omega,m} \leq |f|_{\Omega,m}$ since $m \geq 1$. The proof is concluded by noting that using the definition of V we have $\hat{u}_j = R^{-\top}\hat{v}_j$ with $\hat{v}_j \in \mathbb{R}^n$ corresponding to $\hat{v}_j = (u_j(x_1), \ldots, u_j(x_n))$ for $j \in [d]$ and that $\mathrm{Tr}(\hat{G}) = \frac{\nu}{2}\sum_{j\in[d]} \|\hat{u}_j\|^2$. In particular, some basic linear algebra leads to $\mathrm{Tr}(\hat{G}) = \frac{\nu}{2}\|R^{-\top}(X - 1_n\hat{z}^\top)\|^2$. □

H Details on the algorithmic setup used in the benchmark experiments

In this section, we explain exactly the algorithmic setup which we used to perform the experiments in Sect. 10.1. In all the following problems, the set Ω on which we will minimize the function will be a hyper-rectangle. Given a hyper-rectangle R, we will identify it with its center $c_R \in \mathbb{R}^d$ and its width $w_R \in \mathbb{R}^d$, such that $R = \prod_{i=1}^d ((c_R)_i - (w_R)_i/2, (c_R)_i + (w_R)_i/2)$.

Algorithm 2 Finding a minimizer given points X

function FINDMINIMIZER(f, X, $k(\cdot,\cdot)$, $\lambda_{\min}$, $\lambda_{\max}$, ε)
  $K = (k(x_i, x_j))_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$
  Φ such that $\Phi^\top\Phi = K$ (cholesky decomposition)
  $f_X = (f(x_i))_{1\leq i\leq n} \in \mathbb{R}^n$

  function SCALARFUNCTION(t)
    $\lambda = e^t$
    $\widehat{\alpha}$ solution to Eq. (H.1) with λ, ε, Φ, f_X
    $\widehat{x} = \sum_{i=1}^n \widehat{\alpha}_i x_i$
    $\widehat{f} = f(\widehat{x})$
    return $\widehat{f}, \widehat{x}$
  end function
  $\widehat{f}, \widehat{x}$ = MINIMIZESCALAR(SCALARFUNCTION ,$t_{\min} = \log(\lambda_{\min})$, $t_{\max} = \log(\lambda_{\max})$)
  return $\widehat{f}, \widehat{x}$
end function

We start by defining Algorithm 2 whose main goal is to find a global minimizer as described in the previous sections given sample points $(x_1, ..., x_n)$. Recall that the algorithm introduced in Sects. 6 and 7.1 computes a minimizer by solving problem:

$$\widehat{\alpha} = \underset{\substack{\alpha\in\mathbb{R}^n \\ \alpha^\top 1_n = 1}}{\mathrm{argmin}} \sum_{i=1}^n \alpha_i f(x_i) - \frac{\varepsilon}{n}\log\det\left(\Phi^\top \mathrm{Diag}(\alpha)\Phi + \lambda I\right) + \frac{\varepsilon}{n}\log\frac{\varepsilon}{n} - \varepsilon, \quad \text{(H.1)}$$

where Φ satisfies $\Phi^\top\Phi = K$ for $K = (k(x_i, x_j))_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$, and choosing $\widehat{x}$ as the approximation of the minimizer, defined by

$$\widehat{x} = \sum_{i=1}^n \widehat{\alpha}_i x_i. \quad \text{(H.2)}$$

However, the kernel k, and the hyper-parameters λ also have to be chosen. Therefore, Algorithm 2 will use as inputs: 1) The function f to minimize; 2) the evaluation points $x_i,\ 1 \le i \le n$, summarized in a matrix $X \in \mathbb{R}^{n\times d}$; 3) the kernel k; 4) two parameters $\lambda_{\min}$ and $\lambda_{\max}$ such that we can choose λ in $[\lambda_{\min}, \lambda_{\max}]$; 5) The paramter ε, which controls the log barrier. For simplicity, we hide the hyperparameters linked to the solving of Eq. (H.1) through a Newton method, as explained in Sect. 7.1. Algorithm 2 automatically selects the hyperparameter λ by minimizing the function wich to λ associates the function value of the resulting $\widehat{x}$ on a log scale (ScalarFunction). This function is minimized in the range $[\lambda_{\min}, \lambda_{\max}]$ through the function MinimizeScalar. Hence, the number of function evaluations inherent to running this algorithm is $n+n_{\min}$ where $n_{\min}$ is a minimum number of evaluations (equal to 10).

In our experiments, we use $\varepsilon = 10^{-3}$, $\lambda_{\min} = 10^{-12}$, $\lambda_{\max} = 1$ and we use either the Brent method or simply a grid search with a maximum number of 100 points. This minimization does not have to be very precise. The full algorithm we use is an iterative scheme and is written down in Algorithm 3, computing a sequence (x_k) of approximations of a minimizer of f by iteratively reducing the size of the hyper-rectangle from which the points used in Algorithm 2 are sampled. More precisely, we start from points sampled from a hyper-rectangle with center $x_0 = c_\Omega$ and with width $w_0 = w_\Omega$ (that is the hyper-rectangle Ω) to form $m-1$ samples which, together with x_0, form the points $\widetilde{X}_0 \in \mathbb{R}^{m\times d}$ used to compute the first approximation of the minimizer using FindMinimizer : x_1. Then at each step k, we use the last approximation of the minimizer x_k as the new center of the hyper-rectangle, with width w_k which is set through the predefined function Contraction as $w_k = \text{Contraction}(k)w_0$. As for the first iteration, we then form the concatenation $\widetilde{X}_k \in \mathbb{R}^{m\times d}$ of $m-1$ samples from this hyper-rectangle plus x_k. In order to keep track of the previous points (as a kind of momentum), we apply FindMinimizer with $X_k = [\widetilde{X}_k, \widetilde{X}_{k-1}, \widetilde{X}_{k-2}]$, that is keeping the two last set of points as well as the ones sampled for the k-th step.

Algorithm 3 Converging to the minimum

function FindMinimizerIter(f, Ω, m, N, $k_{(\cdot)}(\cdot,\cdot)$, Contraction)
  $\varepsilon = 10^{-3}, \lambda_{\min} = 10^{-12}, \lambda_{\max} = 1$
  $\widetilde{X}_{-2}, \widetilde{X}_{-1} = [], []$
  $x_0, w_0 = c_\Omega, w_\Omega$
  for $k = 0$ to $N-1$ **do**
    $w_k = \text{Contraction}(k)\times w_0$
    $\sigma_k = \|w_k\|/2$
    $\widetilde{X}_k = [x_k^\top, \text{Uniform}(x_k, w_k, m-1)^\top]^\top$
    $X_k = [\widetilde{X}_{k-2}^\top, \widetilde{X}_{k-1}^\top, \widetilde{X}_k^\top]^\top$
    $f_{k+1}, x_{k+1} = \text{FindMinimizer}(f, X_k, k_{\sigma_k}, \lambda_{\min}, \lambda_{\max}, \varepsilon)$
  end for
  return f_N, x_N
end function

The function FindMinimizerIter uses the following parameters: 1) a kernel function $x, x', \sigma \mapsto k_\sigma(x, x')$ such that σ is a parameter to adapt to the the typical width of the data; 2) the initial hyper-rectangle Ω; 3) the function f; 4) the contraction function

CONTRACTION to set the width of the successive hyper-rectangles; 5) the number m of new points sampled and used at each iteration; 6) the number N of iterations. In our implementation, we use the following parameters.

– For $\sigma > 0$ and $x, y \in \mathbb{R}^d$, we will use the following kernel, which is a mix between the Gaussian (very regular functions) and the Abel kernel (Sobolev functions of order $s = (d+1)/2$ functions), plus a small term 0.01 which allows to handle the constant component of a function more easily.

$$k_\sigma(x, y) = 0.01 + \exp(-\|x - y\|^2/(2\sigma^2)) + \exp(-\|x - y\|/\sigma). \tag{H.3}$$

– We will use the following contraction function, which depends on the dimension as well as the number of iterations :

$$\text{CONTRACTION}(k) = \max\left(\left(1 + \tfrac{1}{d}\right)^{-k}, \tfrac{1}{1+k^{0.6}}\right). \tag{H.4}$$

– The number N of iterations will be set to $N = 200$ unless stated otherwise.
– m will be specified in the experiments : indeed, the higher the dimension, the larger m has to be in order to get meaningful results. Note that one actually uses $n = 3m$ points (from the third iteration onwards) to form the optimization problem, hence the dimension of the SDP solved with the Newton method will be $3m$.

Remark 6 It is equivalent to minimize a function f and minimize the function $\frac{f}{f+c}$ for a positive constant c. This allows to minimize a function in $[0, 1]$ instead of minimizing a real-valued function: however, this also makes higher derivatives behave differently than those of the original function. In practice, instead of minimizing f directly, we minimize $\frac{f}{f+c}$, where c is chosen such that $\frac{f}{f+c}$ will be spread evenly over $[0, 1]$, typically by selecting c as a quantile of the $(f(x_i))_{1\leq i\leq n}$ (we choose the 0.25 quantile). We performed experiments by comparing this renormalization method with simply minimizing f, and this yields slightly better results.

H.1 Additional experiments for global optimization

See Table 4.

Table 4 Complete results of our algorithm on functions on $\mathbb{R}^d$ for $d \geq 2$

	d	iters thresh	Final absolute error	fevs/iter		d	iters thresh	Final absolute error	fevs/iter
FreudensteinRoth	2	5	1.55E−10	21	BoxBetts	3	10	0.00E+00	26
Gulf	3	5	1.22E−03	26	Wolfe	3	1	0.00E+00	26
Mishra09	3	1	2.47E−25	26	Ratkowsky02	3	2	2.86E−06	26
Hartmann3	3	5	0.00E+00	26	Meyer	3	NaN	3.72E+09	26
HelicalValley	3	5	7.67E−09	26	Colville	4	32	1.87E−03	31
Corana	4	1	0.00E+00	31	Shekel07	4	20	9.54E−07	31
PowerSum	4	2	3.26E−04	31	Ratkowsky01	4	90	3.69E+02	31
MieleCantrell	4	3	9.03E−13	31	Powell	4	6	2.85E−07	31
Shekel10	4	18	0.00E+00	31	Shekel05	4	18	1.91E−06	31
BiggsExp04	4	12	7.88E−05	31	Gear	4	2	1.18E−09	31
Kowalik	4	15	4.87E−05	31	DeVilliersGlasser01	4	4	1.06E+03	31
DeVilliersGlasser02	5	NaN	2.28E+03	36	Dolan	5	2	3.78E−13	36
BiggsExp05	5	3	2.64E−03	36	Trid	6	10	0.00E+00	41
Watson	6	11	1.09E−03	41	Hartmann6	6	8	0.00E+00	41
LennardJones	6	2	0.00E+00	41	Thurber	7	125	9.70E+03	46
Xor	9	NaN	6.99E−03	56	Paviani	10	23	1.03E−04	61
Cola	17	68	3.35E−01	96	CrossInTray	2	10	0.00E+00	21
Bukin04	2	16	1.58E−10	21	Matyas	2	1	1.80E−16	21
BartelsConn	2	6	0.00E+00	21	RotatedEllipse01	2	3	2.63E−12	21
Branin01	2	5	0.00E+00	21	OddSquare	2	7	2.38E−07	21

Table 4 continued

	d	iters thresh	Final absolute error	fevs/iter		d	iters thresh	Final absolute error	fevs/iter
Ursem04	2	4	0.00E+00	21	Ripple01	2	NaN	9.52E−02	21
Brent	2	2	2.77E−09	21	Schaffer02	2	1	2.64E−14	21
DropWave	2	27	0.00E+00	21	NeedleEye	2	1	0.00E+00	21
Schwefel22	2	5	1.17E−08	21	XinSheYang01	2	2	1.05E−08	21
Pinter	2	1	5.40E−16	21	Penalty01	2	3	5.10E−17	21
Langermann	2	5	0.00E+00	21	Salomon	2	7	1.81E−08	21
VenterSobiezcczanskiSobieski	2	1	0.00E+00	21	Schaffer03	2	14	8.38E−07	21
Shubert04	2	7	− 1.91E−06	21	Price01	2	5	2.40E−05	21
Giunta	2	3	0.00E+00	21	Cigar	2	2	1.94E−08	21
Bukin02	2	4	0.00E+00	21	YaoLiu09	2	2	0.00E+00	21
Vincent	2	8	0.00E+00	21	Qing	2	8	3.04E−05	21
WayburnSeader01	2	2	7.58E−11	21	Levy13	2	5	7.61E−13	21
Schaffer04	2	7	− 3.58E−07	21	Brown	2	3	4.37E−18	21
Ackley01	2	8	6.50E−09	21	CrossLegTable	2	NaN	9.98E−01	21
Schwefel36	2	4	0.00E+00	21	CosineMixture	2	9	0.00E+00	21
Quadratic	2	2	0.00E+00	21	Exponential	2	1	0.00E+00	21
NewFunction01	2	NaN	1.18E−01	21	HolderTable	2	5	0.00E+00	21
TestTubeHolder	2	NaN	1.98E−02	21	Ursem03	2	5	0.00E+00	21
Sphere	2	1	7.73E−16	21	Levy03	2	3	1.19E−18	21
Schaffer01	2	1	4.44E−15	21	Rastrigin	2	7	1.07E−14	21

Table 4 continued

	d	iters thresh	Final absolute error	fevs/iter		d	iters thresh	Final absolute error	fevs/iter
McCormick	2	2	0.00E+00	21	SixHumpCamel	2	5	−4.77E−07	21
RotatedEllipse02	2	3	4.94E−15	21	Branin02	2	NaN	1.25E+00	21
Alpine01	2	28	7.44E−09	21	Quintic	2	10	2.25E−05	21
Schwefel26	2	9	−5.45E−07	21	SineEnvelope	2	3	7.83E−15	21
Stochastic	2	5	3.99E−07	21	ZeroSum	2	NaN	1.00E+00	21
UrsemWaves	2	NaN	9.08E−01	21	RosenbrockModified	2	11	0.00E+00	21
Penalty02	2	68	1.68E−06	21	XinSheYang02	2	7	1.21E−11	21
Ackley03	2	8	0.00E+00	21	YaoLiu04	2	7	5.22E−09	21
Schwefel21	2	3	9.83E−09	21	Zimmerman	2	5	7.46E−03	21
Decanomial	2	2	3.95E−09	21	Alpine02	2	3	−3.81E−06	21
Ursem01	2	3	0.00E+00	21	NewFunction02	2	NaN	2.01E−01	21
Levy05	2	6	0.00E+00	21	Chichinadze	2	2	7.63E−06	21
Trefethen	2	55	6.91E−06	21	Hansen	2	28	0.00E+00	21
DeckkersAarts	2	13	0.00E+00	21	Michalewicz	2	4	−3.34E−06	21
Treccani	2	1	1.87E−15	21	Bukin06	2	2	3.42E−04	21
XinSheYang03	2	2	0.00E+00	21	Ackley02	2	23	0.00E+00	21
Csendes	2	NaN	NAN	21	Schwefel20	2	8	4.63E−08	21
Whitley	2	3	2.95E−11	21	XinSheYang04	2	49	0.00E+00	21
Step	2	1	0.00E+00	21	WayburnSeader02	2	4	6.14E−09	21

Table 4 continued

	d	iters thresh	Final absolute error	fevs/iter		d	iters thresh	Final absolute error	fevs/iter
Mishra02	2	1	0.00E+00	21	StyblinskiTang	2	6	0.00E+00	21
ElAttarVidyasagarDutta	2	47	0.00E+00	21	Plateau	2	1	0.00E+00	21
Price03	2	31	3.10E−16	21	HimmelBlau	2	2	1.17E−14	21
DixonPrice	2	6	6.07E−14	21	Parsopoulos	2	6	3.35E−16	21
Price02	2	17	0.00E+00	21	Mishra08	2	2	3.99E−16	21
Beale	2	3	5.06E−19	21	Sodp	2	1	1.62E−23	21
Schwefel06	2	15	3.02E−08	21	Zirilli	2	6	− 5.96E−08	21
Mishra01	2	1	0.00E+00	21	Damavandi	2	NaN	NAN	21
ThreeHumpCamel	2	1	1.68E−19	21	Mishra07	2	1	1.97E−29	21
Pathological	2	2	2.22E−14	21	Ripple25	2	NaN	1.66E−01	21
Hosaki	2	2	0.00E+00	21	Exp2	2	2	2.62E−15	21
Trigonometric02	2	158	0.00E+00	21	Judge	2	2	0.00E+00	21
AMGM	2	1	0.00E+00	21	CrownedCross	2	NaN	3.35E−02	21
JennrichSampson	2	6	7.63E−06	21	Schwefel01	2	1	3.67E−16	21
Adjiman	2	5	0.00E+00	21	Katsuura	2	3	0.00E+00	21
Bohachevsky1	2	6	1.24E−14	21	Shubert01	2	7	− 1.53E−05	21
Zacharov	2	2	4.89E−16	21	Mishra06	2	5	2.38E−07	21
Deb01	2	2	0.00E+00	21	Infinity	2	1	2.51E−34	21

Table 4 continued

	d	iters thresh	Final absolute error	fevs/iter		d	iters thresh	Final absolute error	fevs/iter
Sargan	2	4	4.92E−13	21	Mishra03	2	NaN	1.33E−01	21
Bird	2	3	0.00E+00	21	Mishra05	2	11	4.66E−04	21
GoldsteinPrice	2	3	0.00E+00	21	Mishra10	2	1	0.00E+00	21
EggCrate	2	5	1.43E−16	21	Schwefel02	2	2	1.47E−13	21
Cube	2	21	5.98E−09	21	Rosenbrock	2	35	5.15E−05	21
Bohachevsky2	2	2	4.82E−14	21	DeflectedCorrugatedSpring	2	8	0.00E+00	21
Step2	2	1	0.00E+00	21	Deb03	2	4	0.00E+00	21
Wavy	2	1	0.00E+00	21	Trigonometric01	2	NaN	5.83E−01	21
Tripod	2	20	8.92E−08	21	Weierstrass	2	16	3.52E−06	21
Shubert03	2	28	0.00E+00	21	MultiModal	2	1	6.00E−26	21
Price04	2	3	1.06E−20	21	Deceptive	2	7	2.97E−04	21
Schwefel04	2	2	1.28E−17	21	Leon	2	12	8.62E−15	21
CarromTable	2	10	0.00E+00	21	Rana	2	NaN	1.86E+00	21
Bohachevsky3	2	6	1.67E−15	21	PenHolder	2	6	0.00E+00	21
Zettl	2	2	0.00E+00	21	Keane	2	1	2.25E−36	21
Mishra11	2	6	3.16E−30	21	Mishra04	2	NaN	1.78E−01	21

References

1. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis, vol. 1349. Springer, Berlin (2006)
2. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, Berlin (2013)
3. Ivanov, V.V.: On optimum minimization algorithms in classes of differentiable functions. In: Doklady Akademii Nauk, vol. 201, pp. 527–530. Russian Academy of Sciences, Moscow (1971)
4. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems: Standard Information for Functionals, vol. 12. European Mathematical Society, Helsinki (2008)
5. Osborne, M.A., Garnett, R., Roberts, S.J.: Gaussian processes for global optimization. In: International Conference on Learning and Intelligent Optimization (LION3), pp. 1–15 (2009)
6. Lasserre, J.-B.: Global optimization with polynomials and the problem of moments. SIAM J. Optim. **11**(3), 796–817 (2001)
7. Laurent, M.: Sums of squares, moment matrices and optimization over polynomials. In: Emerging Applications of Algebraic Geometry, pp. 157–270. Springer, Berlin (2009)
8. Lasserre, J.-B.: Moments, Positive Polynomials and Their Applications, vol. 1. World Scientific, Singapore (2010)
9. Marteau-Ferey, U., Bach, F., Rudi, A.: Non-parametric models for non-negative functions. Adv. Neural. Inf. Process. Syst. **33**, 12816–12826 (2020)
10. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer, Berlin (2011)
11. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces. Elsevier, Amsterdam (2003)
12. Lasserre, J.-B., Toh, K.-C., Yang, S.: A bounded degree SOS hierarchy for polynomial optimization. EURO J. Comput. Optim. **5**(1–2), 87–117 (2017)
13. Marx, S., Pauwels, E., Weisser, T., Henrion, D., Lasserre, J.: Semi-algebraic approximation using Christoffel–Darboux kernel. Technical report arXiv:1904.01833 (2019)
14. Nie, J.: Optimality conditions and finite convergence of Lasserre's hierarchy. Math. Program. **146**(1–2), 97–121 (2014)
15. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**(3), 337–404 (1950)
16. Paulsen, V.I., Raghupathi, M.: An Introduction to the Theory of Reproducing Kernel Hilbert Spaces, vol. 152. Cambridge University Press, Cambridge (2016)
17. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, Berlin (2008)
18. Wendland, H.: Scattered Data Approximation, vol. 17. Cambridge University Press, Cambridge (2004)
19. Del Moral, P., Niclas, A.: A Taylor expansion of the square root matrix function. J. Math. Anal. Appl. **465**(1), 259–266 (2018)
20. Narcowich, F.J., Ward, J.D., Wendland, H.: Refined error estimates for radial basis function interpolation. Constr. Approx. **19**(4), 541–564 (2003)
21. Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
22. Penrose, M.: Random Geometric Graphs, vol. 5. Oxford University Press, Oxford (2003)
23. Nemirovski, A.: Interior point polynomial time methods in convex programming. Lecture notes (2004)
24. Bach, F.: Breaking the curse of dimensionality with convex neural networks. J. Mach. Learn. Res. **18**(1), 629–681 (2017)
25. Lasserre, J.-B.: A sum of squares approximation of nonnegative polynomials. SIAM Rev. **49**(4), 651–669 (2007)
26. Lasserre, J.-B.: A new look at nonnegativity on closed sets and polynomial optimization. SIAM J. Optim. **21**(3), 864–885 (2011)
27. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
28. Montaz Ali, M., Khompatraporn, C., Zabinsky, Z.B.: A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. J. Global Optim. **31**(4), 635–672 (2005)
29. Jamil, M., Yang, X.S.: A literature survey of benchmark functions for global optimisation problems. Int. J. Math. Model. Numer. Optim. **4**(2), 150 (2013)
30. Molga, M., Smutnicki, C.: Test functions for optimization needs (2005)
31. Henrion, D., Lasserre, J.-B., Löfberg, J.: Gloptipoly 3: moments, optimization and semidefinite programming. Optim. Methods Softw. **24**, 10 (2007)

32. Lasserre, J.-B.: The moment-SOS hierarchy and the Christoffel–Darboux kernel. Technical Report. arXiv:2011.08566 (2020)
33. Nesterov, Y.: Squared functional systems and optimization problems. In: High Performance Optimization, pp. 405–440. Springer, Berlin (2000)
34. Slot, L., Laurent, M.: Near-optimal analysis of Lasserre's univariate measure-based bounds for multivariate polynomial optimization. Math. Program. **188**, 443–460 (2020)
35. Zhou, D.-X.: Derivative reproducing properties for kernel methods in learning theory. J. Comput. Appl. Math. **220**(1–2), 456–463 (2008)
36. Bach, F.: Sharp analysis of low-rank kernel matrix approximations. In: Conference on Learning Theory, pp. 185–209 (2013)
37. Rudi, A., Camoriano, R., Rosasco, L.: Less is more: Nyström computational regularization. Adv. Neural. Inf. Process. Syst. **28**, 1657–1665 (2015)
38. Rudi, A., Rosasco, L.: Generalization properties of learning with random features. Adv. Neural. Inf. Process. Syst. **30**, 3215–3225 (2017)
39. Bach, F.: On the equivalence between kernel quadrature rules and random feature expansions. J. Mach. Learn. Res. **18**(1), 714–751 (2017)
40. Hörmander, L.: The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis. Springer, Berlin (2015)
41. Brenner, S., Scott, R.: The Mathematical Theory of Finite Element Methods, vol. 15. Springer, Berlin (2007)
42. Weidmann, J.: Linear Operators in Hilbert Spaces, vol. 68. Springer, Berlin (1980)
43. Bhatia, R.: Matrix Analysis, vol. 169. Springer, Berlin (2013)
44. Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W.: NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge (2010)
45. Sickel, W.: Superposition of functions in Sobolev spaces of fractional order. A survey. Banach Center Publ. **27**, 481–497 (1992)