Proceedings of Machine Learning Research vol 178:1–23, 2022 35th Annual Conference on Learning Theory

Non-Convex Optimization with Certificates and Fast Rates Through Kernel Sums of Squares

Blake Woodworth BLAKE.WOODWORTH@INRIA.FR
Francis Bach FRANCIS.BACH@INRIA.FR
Alessandro Rudi ALESSANDRO.RUDI@INRIA.FR
Inria, Ecole Normale Supérieure, PSL Research University, Paris, France

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider potentially non-convex optimization problems, for which optimal rates of approximation depend on the dimension of the parameter space and the smoothness of the function to be optimized. In this paper, we propose an algorithm that achieves close to optimal a priori computational guarantees, while also providing a posteriori certificates of optimality. Our general formulation builds on infinite-dimensional sums-of-squares and Fourier analysis, and is instantiated on the minimization of periodic functions.

Keywords: Non-convex optimization, sum of squares, kernel methods.

1. Introduction

A well-designed optimization algorithm provides two important types of guarantees. First, it guarantees *a priori* that its output will achieve a certain degree of accuracy, with computational complexity that is hopefully adaptive to the specific properties of function to be optimized and possibly even optimal over a certain class of algorithms or functions to minimize. Second, it provides an *a posteriori* certificate, i.e., an explicit bound on the solution's accuracy that we can calculate once we run the algorithm. There are many examples of such well-designed optimization algorithms in the convex setting, which often use some form of convex duality (see, e.g., Nemirovski et al., 2010).

In this paper, our goal is to provide a well-designed algorithm for *non-convex* optimization,

$$c_* = \inf_{x \in \mathcal{X}} f(x), \tag{1}$$

with $\mathcal{X} \subseteq \mathbb{R}^d$ and f potentially non-convex. In general, this task is extremely difficult, and in the worst case the computational cost must be exponential in the dimension d. However, it is known (Novak, 2006) that in order for an algorithm to achieve the optimal computational complexity in solving Eq. (1), it must be adaptive to the degree of differentiability of f. That is, it should be able to overcome the curse of dimensionality in terms of the approximation error ε, when the function is very smooth. More precisely, if f is m-times differentiable, then the computational complexity for finding c_* with error ε should be $C_d\, \varepsilon^{-d/m}$, where C_d is exponential in d in the worst case, but the dependence on the accuracy scales as only $\varepsilon^{-d/m}$, which becomes quite mild once m approaches d. In this case, the curse of dimensionality is relegated just to C_d making it possible to efficiently solve non-convex problems to very high accuracy as long as d is relatively small, which has applications to tasks like hyperparameter tuning, industrial process optimization, and more.

Establishing well-designed algorithms for non-convex optimization is a difficult task. Many non-convex optimization algorithms in the literature lack a priori guarantees, a posteriori guarantees,

© 2022 B. Woodworth, F. Bach & A. Rudi.

or actually any guarantees at all. Many methods used in practice are based on heuristics, and can only guarantee convergence to a *local* rather than global minimum, or even just to a first-order stationary point. Some other algorithms have good a posteriori guarantees, but weak a priori bounds; for example, methods based on polynomial sum of squares (Lasserre, 2001; Parrilo, 2003) are not adaptive, a priori, to the smoothness and are therefore subject to the curse of dimensionality in terms of the accuracy ε. The new family of algorithms based on kernel sum of squares (Rudi et al., 2020) achieves quasi-optimal a priori guarantees, but without any certificate a posteriori.

Our Contributions

In this paper, we provide a general strategy to derive algorithms that compute a *lower bound* $\hat{c}$ of c_* with strong guarantees both a priori and a posteriori (see Corollary 11). As a particular example, we consider optimizing smooth, periodic functions f on $\mathcal{X} = [0,1]^d$ and derive an algorithm that: (*a priori*) approximates c_* with almost optimal error ε and computational complexity that scales well with ε, and (*a posteriori*) provide a certificate of accuracy, which is adaptive to the specific instance of the problem.

The a priori guarantee is useful since it shows that the proposed algorithm has nearly optimal complexity, which is adaptive to the smoothness of the function to be optimized. The a posteriori certificate is particularly useful because the accuracy of our estimate of c_* depends on the specific instance at hand, and may be much better than the worst-case, exponential-in-d constant would suggest. Indeed, better-than-worst-case performance is frequently observed in practice, but we need a certificate of accuracy in order to *know* when we are so lucky.

Notation and Basic Definitions

Throughout this paper, we will assume that the objective f is continuous, and that its infimum is attained on $\mathcal{X}$. For a function $h : \mathcal{X} \to \mathbb{R}$, will denote $\|h\|_{L^\infty(\mathcal{X})} = \sup_{x\in\mathcal{X}} |h(x)|$. We also define the Fourier transform of h as

$$\hat{h}(\omega) = \int_{\mathcal{X}} h(x) e^{-2\pi i \omega^\top x} dx \,,$$

and we define the inverse Fourier transform

$$h(x) = \int_{\Omega} \hat{h}(x) e^{2\pi i \omega^\top x} d\omega \,.$$

2. Deriving Well-Designed Algorithms for Smooth Non-Convex Optimization

Our approach begins with rewriting the problem (1) as finding the highest lower bound on f:

$$c_* = \max_{c\in\mathbb{R}} \; c \quad \text{such that} \quad f(x) \geq c \;\; \forall x \in \mathcal{X},$$

The inequality constraint $f - c \geq 0$ can be converted to an equality constraint by introducing a non-negative function $g \geq 0$:

$$c_* = \max_{c\in\mathbb{R},\; g\geq 0} \; c \quad \text{such that} \quad f(x) - c - g(x) = 0 \;\; \forall x \in \mathcal{X}.$$

Finally, we can rewrite this again in a penalized form

$$\tilde{c} \;\; = \;\; \max_{c\in\mathbb{R},\; g\geq 0} \; c \; - \; \|f - c - g\|_{L^\infty(\mathcal{X})}. \tag{2}$$

It may not yet be obvious what this accomplishes, but the following Lemma indicates its promise:

Lemma 1 *The problem (2) is a concave maximization problem with solution $\tilde{c} = c_*$, and for any feasible (c, g) such that $c \in \mathbb{R}$ and $g \geq 0$, c_* is lower-bounded by $c - \|f - c - g\|_{L^\infty(\mathcal{X})}$.*

Proof The objective $V(c, g) := c - \|f - c - g\|_{L^\infty(\mathcal{X})}$ is concave because it is a linear term c minus a convex term $\|\cdot\|_{L^\infty(\mathcal{X})}$ composed with a linear function of the optimization variables.

Since for all $x \in \mathcal{X}$, $f(x) - c - g(x) \geq -\|f - c - g\|_{L^\infty(\mathcal{X})}$, for any feasible (c, g), then $\forall x \in \mathcal{X}$, $f(x) \geq V(c, g)$. Thus, by minimizing with respect to x, $c_* \geq \tilde{c}$. In order to show the other inequality, we notice that $(c_*, f - c_*)$ is feasible and $V(c_*, f - c_*) = c_*$. ■

So, we have reduced the non-convex minimization problem (1) to a concave maximization problem (2), which is an improvement. However, whereas the original problem had a d-dimensional optimization variable, our new problem requires optimizing over the infinite-dimensional space of non-negative functions, and it is not clear how to do this. Furthermore, the quantity $\|f - c - g\|_{L^\infty(\mathcal{X})}$ is just as difficult to compute as $\inf_{x \in \mathcal{X}} f(x)$ would be. We will now describe several modifications leading to a more tractable optimization problem that maintains the same desirable properties.

A more tractable formulation. Specifically, our approach revolves around introducing a more tractable norm $\|\cdot\|_W$ on functions from $\mathcal{X}$ to $\mathbb{R}$, and restricting $g \in \mathcal{G}$, for $\mathcal{G}$ a tractable subset of all non-negative functions. We will specify W and $\mathcal{G}$ later. To provide guarantees on the method we only need them to satisfy the following:

Conditions 1 (Norms and models for non-negative functions)

1. *Let $\|\cdot\|_W$ be a norm on the space of real-valued functions over $\mathcal{X}$, such that $\|\cdot\|_{L^\infty(\mathcal{X})} \leq \|\cdot\|_W$. Denote by $\mathcal{W}$ the associated Banach space.*

2. *Let $\mathcal{G}$ be a convex subset of the set of non-negative functions, such that $\mathcal{G}$ is a closed subset of $\mathcal{W}$.*

Now, by restricting the problem in Eq. (2) on $\mathcal{G}$ and considering the norm $\|\cdot\|_W$, we obtain what can be a more tractable formulation,

$$\bar{c} = \max_{c \in \mathbb{R},\ g \in \mathcal{G}} \quad c - \|f - c - g\|_W. \tag{3}$$

It is, of course, not the case that *any* choice of $\|\cdot\|_W$ and $\mathcal{G}$ would make Eq. (3) easy to solve. However, we will later discuss examples where Eq. (3) is much easier to solve than Eq. (2). Regardless, we will see in the next Theorem, that this formulation comes with strong guarantees, expressing the error of the algorithm directly in terms of the approximation properties of the class of models $\mathcal{G}$ for non-negative functions.

Theorem 2 (Tightness of Eq. (3)) *Suppose that $f - c \in \mathcal{W}$ for any $c \in \mathbb{R}$, then*

$$c_* - q \leq \bar{c} \leq c_*, \qquad q = \min_{g \in \mathcal{G}} \|f - c_* - g\|_W.$$

Proof By construction $\bar{c} \leq c_*$, since $c - \|f - c - g\|_W \leq c - \|f - c - g\|_{L^\infty(\mathcal{X})} \leq c_*$ and $\mathcal{G} \subseteq \{g \mid g \geq 0\}$. Now the problem is well defined, since it corresponds to a maximization on the closed subset $\mathbb{R} \times \mathcal{G}$ of a concave and continuous objective function (for the topology inherited from $\mathcal{W}$). By setting $c = c_*$ and optimizing over g, the optimized objective is exactly $c_* - q$ with q as above, thus $c_* - q \leq \bar{c}$. Moreover,we have $q = 0$, i.e., $\bar{c} = c_*$ when there exists $g_* \in \mathcal{G}$ satisfying $\|f - c_* - g_*\|_W = 0$. ■

In the Theorem above we see that $\bar{c}$ is lower bound of c_* by construction. Moreover, the error $c_* - \bar{c}$ is bounded by q, the *approximation error* of $f - c_*$ with respect to the class of models for non-negative functions $\mathcal{G}$ that we consider, measured in the norm W. So far, this all holds for any $\|\cdot\|_W$ and $\mathcal{G}$; we continue by analyzing a specific choice.

The model for non-negative functions. We would like to use a class of models $\mathcal{G}$, that can approximate smooth non-negative functions using as few parameters as possible, while remaining tractable to optimize over. We consider the class of *PSD models* introduced by Marteau-Ferey et al. (2020) and defined as

$$g(x) = \phi(x)^* A \phi(x), \quad A \succeq 0,$$

for a suitable map $\phi : \mathcal{X} \to \mathbb{C}^n$ and $A \in \mathbb{C}^{n \times n}$ positive semidefinite, where ϕ^* denotes the conjugate transpose of ϕ. By the definition of positive semidefinite, $g(x) \geq 0$ for any $x \in \mathcal{X}$, and this is also a tractable class to optimize over since g is linear in the parameters A. The approximation properties of the model will depend on the choice of the feature map ϕ, and have been shown to give rates in the order of $n^{-m/d}$ for specific choices (Rudi and Ciliberto, 2021). Here, however, we need to study the approximation error with respect to our norm of choice $\|\cdot\|_W$, and we will see that different feature maps than the ones considered by Rudi et al. (2020) will lead to better rates.

The norms. We need norms that bound $\|\cdot\|_{L^\infty(\mathcal{X})}$ from above as tightly as possible, but are also easy to compute. We consider from this viewpoint two norms:

1. The "F norm": L^1 norm of the Fourier transform, i.e.,

$$\|u\|_F = \int_{\mathbb{R}^d} |\hat{u}(\omega)| d\omega,$$

2. The "S norm": The norm associated to a richer reproducing kernel Hilbert space such as the Sobolev space of exponent $(d+1)/2$. Let $S(\omega)$ non negative and integrable, we define

$$\|u\|_S^2 = C_S^2 \int_{\mathbb{R}^d} \frac{|\hat{u}(\omega)|^2}{S(\omega)} d\omega,$$

and $C_S^2 = \int_{\mathbb{R}^d} S(\omega) d\omega$. For example for the Sobolev case we set $S(\omega) = (1 + \|\omega\|^2)^{-(d+1)/2}$.

Lemma 3 *The norms above satisfy* $\|\cdot\|_{L^\infty(\mathbb{R}^d)} \leq \|\cdot\|_F \leq \|\cdot\|_S$, *for any non-negative and integrable* S. *Moreover,*

$$\|u\|_F = \min_{S \geq 0, S \in L^1(\mathbb{R}^d)} \|u\|_S.$$

Proof First, for any u with finite integrable Fourier transform, we have, by Hölder inequality,

$$|u(x)| = \left| \int \hat{u}(\omega) e^{2\pi i \langle x, \omega \rangle} d\omega \right| \leq \|\hat{u}\|_{L^1(\mathbb{R}^d)} \|e^{2\pi i \langle x, \cdot \rangle}\|_{L^\infty(\mathbb{R}^d)} \leq \|\hat{u}\|_{L^1(\mathbb{R}^d)} =: \|u\|_F, \qquad \forall x \in \mathbb{R}^d$$

from which we conclude that $\|\cdot\|_F$ always upper bounds $\|\cdot\|_{L^\infty(\mathbb{R}^d)}$. Analogously, note that, for any $S \geq 0$ and $S \in L^1(\mathbb{R}^d)$, and for any u such that $\|u\|_S$ is finite, we have, by Cauchy-Schwartz,

$$\|u\|_F^2 = \|\widehat{u}\|_{L^1(\mathbb{R}^d)}^2 = \left\| S^{1/2} \frac{\widehat{u}}{S^{1/2}} \right\|_{L^1(\mathbb{R}^d)}^2 \leq \|S^{1/2}\|_{L^2(\mathbb{R}^d)}^2 \left\| \frac{\widehat{u}}{S^{1/2}} \right\|_{L^2(\mathbb{R}^d)}^2 = C_S^2 \int \frac{|\hat{u}(\omega)|^2}{S(\omega)} = \|u\|_S^2.$$

Finally, for u with finite F-norm, $\|u\|_F = \|u\|_S$ when $S = |\widehat{u}|$ for which $S \geq 0, S \in L^1(\mathbb{R}^d)$. ■

The F norm and the S norm are of interest when we can assume that we have access to the Fourier transform of the target function f. In particular the norm S can be also computed in closed form in specific scenarios. For example, consider the case when f is of the form

$$f = \sum_{j=1}^{M} \beta_j h(x - x_i),$$

for some $\beta_1, \ldots, \beta_M \in \mathbb{R}$ and $x_1, \ldots, x_M \in \mathbb{R}^d$. This arises, e.g., for mixtures of Gaussians, and learning linear models or RBF networks. If we know the Fourier transform of h, then

$$\|f\|_S^2 := \sum_{i,j=1}^{M} \beta_i \beta_j H(x_i - x_j),$$

where H is the inverse Fourier transform of the function $S(\omega)|\hat{h}(\omega)|^2$.

Perhaps more interestingly, Lemma 3 shows that the F norm is weaker than the norm S for any S, meaning that the F norm allow us to automatically adapt to certain structure in the function. For example, suppose that $f(x) = h(Px)$ for some unknown $P \in \mathbb{R}^{d' \times d}$ with $d' \ll d$ and $h : \mathbb{R}^{d'} \to \mathbb{R}$. Using the S norm for a certain S that depends on P would allow us to adapt to the low-dimensional structure and depend on d' rather than d, but this requires knowing P. On the other hand, the F norm is always weaker, so we can take advantage of the low-dimensional structure automatically.

2.1. PSD Models for Periodic Functions on $[0, 1]^d$ and Their Approximation Properties

The goal of the section is to provide a self-contained introduction to the approximation properties of PSD models. In particular, we consider the problem of approximating smooth 1-periodic functions (which corresponds to $\mathcal{X} = [0, 1]^d$) using PSD models where ϕ is a subset of the Fourier basis. This setting, while already being of interest for practical applications, allows for an elementary proof which highlights the main conceptual steps of the derivation.

The main results of this section are Theorem 5 and, in particular, Theorem 7. With a more refined proof based on the same strategy, it is also possible to obtain results that hold for other scenarios beyond periodic functions on the torus and for more general maps ϕ. See, for example, Rudi and Ciliberto (2021) for the approximation of non-periodic C^m functions on subsets of $\mathbb{R}^d$ via PSD models based on a finite dimensional feature map defined with respect to the Gaussian kernel,

or Rudi et al. (2020) for a feature map defined with respect to any kernel that satisfy some algebraic property, such as the Sobolev kernel.

We have seen in Theorem 2 that the optimization error of Eq. (3) depends on the approximation error of the function f with respect to the class of models for non-negative functions. So, in our analysis there will be three main ingredients: a class of models $\mathcal{G}_t$ parametrized by its bandwidth, t, which will depend on the space of functions associated with a feature map ϕ_t; the space of functions where f lives, which we denote H_ρ; and the norm that we use to measure the approximation error, in our case, $\|\cdot\|_F$.

We start by introducing $\mathcal{G}_t$, parametrized by a bandwidth $t \in \mathbb{N}$. We associate each entry in $\phi_t(x)$ with an element of $\{k \in \mathbb{Z}^d : |k| \leq t\}$ where $|k| = \sum_j |k_j|$, with $n = \#(\{k \in \mathbb{Z}^d : |k| \leq t\}) = \binom{d+t}{t}$, i.e., $n = O(t^d)$. So for each $|k| \leq t$, we define the feature map $\phi_t : \mathcal{X} \to \mathbb{C}^n$ elementwise as $(\phi_t(x))_k = e_k(x)$ where e_k is the k-th Fourier component, i.e., $e_k(x) = e^{-2\pi i k^\top x}$. Consider the class of PSD models

$$\mathcal{G}_t = \{g_{A,t} \mid A \in \mathbb{C}^{n\times n}, A \succeq 0\}, \quad g_{A,t}(x) = \phi_t(x)^* A \phi_t(x), \tag{4}$$

We are thus considering the feature map ϕ_t associated with the classical band-limited space of functions. This choice is convenient for our analysis, but there are also many other choices of finite dimensional feature maps for PSD models that can have good approximation properties (Rudi and Ciliberto, 2021; Rudi et al., 2020).

We consider continuous, 1-periodic functions f on $\mathbb{R}^d$, i.e., functions satisfying $f(x+k) = f(x)$ for any $x \in \mathbb{R}^d$ and $k \in \mathbb{Z}^d$. We note that these can therefore be identified with continuous, periodic functions on the torus $[0,1]^d$. We now introduce H_ρ, where $\rho \in \ell_1(\mathbb{Z}^d)$ is a strictly positive summable sequence. The space is a separable Hilbert space of periodic functions defined as $H_\rho = \{f \in L^2(\mathcal{X}) \mid \|f\|_\rho < \infty\}$, where $\|f\|_\rho^2 = \sum_{k\in\mathbb{Z}^d} |\widehat{f}_k|^2/\rho_k < \infty$ and $\widehat{f}_k = \int_{[0,1]^d} e_k(x) f(x) dx$ is the Fourier series associated to f. One classical example of ρ is $\rho_k = (1 + \|k\|^2)^{-m}$, with $m > d/2$, corresponding to the Sobolev space $H^m_{2,\text{per}}$ of periodic functions whose derivatives up to order m are squared integrable (Wahba, 1990), or the space of periodic entire functions (of order 1), corresponding to $\rho_k = \exp(-\sigma\|k\|)$, for some $\sigma > 0$.

We will soon present a Theorem showing that the PSD models $\mathcal{G}_t$ can approximate functions of the form $f = \sum_{j=1}^q u_j^2$ for $q \in \mathbb{N}$ and $u_j \in H_\rho$, using a small t depending on the decreasing quantity

$$R_t^2 := \sum_{|k|>t} \rho_k.$$

First, we start with a Lemma concerning the norm $\|\cdot\|_F$ of the pointwise product of functions. This part of the proof is crucial and is handled differently in the other settings (e.g., Rudi et al., 2020).

Lemma 4 *Let f, g be 1-periodic functions on $\mathcal{X}$ with $\|f\|_F, \|g\|_F < \infty$ and denote by $f \cdot g$ their pointwise product, i.e., $(f \cdot g)(x) = f(x)g(x)$. Then*

$$\|f \cdot g\|_F \leq \|f\|_F \|g\|_F.$$

Proof By the convolution property of Fourier series, $\widehat{(f \cdot g)}_k = \sum_{j\in\mathbb{Z}^d} \widehat{f}_j \widehat{g}_{k-j}$. By the Young inequality for the convolution of discrete sequences, we have for any two sequences $u, v \in \ell_1(\mathbb{Z}^d)$, $\sum_{j,k\in\mathbb{Z}^d} |\widehat{u}_j \widehat{v}_{k-j}| \leq (\sum_{k\in\mathbb{Z}^d} |u_k|)(\sum_{k\in\mathbb{Z}^d} |v_k|)$. The result is obtained by applying this inequality on the Fourier series of $f \cdot g$ and noting that $\sum_{k\in\mathbb{Z}^d} |\widehat{f}_k|$ is exactly $\|f\|_F$, and the same for g. ■

Now we are ready to state the first theorem that on the approximation error for the PSD models described above.

Theorem 5 *Let $f(x) = \sum_{j=1}^{T} u_j(x)^2$ for functions $u_j \in H_\rho$ and $T \in \mathbb{N}$, then*

$$\min_{g \in \mathcal{G}_t} \|f - g\|_F \;\leq\; C'_f \; R_t \,.$$

where $C'^2_f = \sum_{j=1}^{T} \|u_j\|_\rho \|u_j\|_F$.

Proof Denote by $u_{j,t}$ the function, $u_{j,t}(x) = \sum_{|k| \leq t} (\widehat{u}_j)_k e_k(x)$ (a low-pass filtered version of u), and by $v_{j,t}$ the n-dimensional vector $(v_{j,t})_k = \widehat{u}_{j,t}$, for any $|k| \leq t$. Now, define $\bar{A} \in \mathbb{C}^{n \times n}$ as

$$\bar{A} = \sum_{j=1}^{T} v_{j,t} v_{j,t}^*.$$

Since, by construction, $v_{j,t}^* \phi(x) = \sum_{|k| \leq t} (\widehat{u}_j)_k e_k(x) = u_{j,t}(x)$, then

$$g_{\bar{A},t}(x) = \phi_t(x)^* A \phi_t(x) = \sum_{j=1}^{T} \phi_t(x)^* (v_{j,t} v_{j,t}^*) \phi_t(x) = \sum_{j=1}^{T} (v_{j,t}^* \phi_t(x))^2 = \sum_{j=1}^{T} u_{j,t}(x)^2.$$

Now note that $u_j^2 - u_{j,t}^2 = (u_j + u_{j,t}) \cdot (u_j - u_{j,t})$, then, by using Lemma 4,

$$\|f - g_{\bar{A},t}\|_F = \Big\| \sum_{j=1}^{T} (u_j^2 - u_{j,t}^2) \Big\|_F \leq \sum_{j=1}^{T} (\|u_j\|_F + \|u_{j,t}\|_F) \|u_j - u_{j,t}\|_F.$$

We conclude noting that $\|u_{j,t}\|_F \leq \|u_j\|_F$ by construction and, by Cauchy-Schwartz,

$$\|u_j - u_{j,t}\|_F = \sum_{|k|>t} |\widehat{u}_j| = \sum_{|k|>t} \sqrt{\rho}_t \frac{|\widehat{u}_j|}{\sqrt{\rho}_t} \leq R_t \|u_{j,t}\|_\rho \leq R_t \|u_j\|_\rho.$$

Therefore, $\min_{g \in \mathcal{G}_t} \|f - g\|_F = \min_{A \in \mathbb{C}^{n \times n}, A \succeq 0} \|f - g_{A,t}\|_F \leq \|f - g_{\bar{A},t}\|_F \leq R_t C'_f$. ∎

The theorem above controls the approximation error of the PSD models of bandwidth t when the target function can be written in terms of a sum of squares of functions belonging to an H_ρ for a given ρ. In general it is not clear how to guarantee when a function f can be characterized as a sum of squares of functions in a given space. Luckily, in the case of m-times differentiable functions, there exists an easy geometrical characterization. We are going to use this fact, to specify the result above for the case when f is an m-times differentiable function. First, we need the following lemma, which is the adaptation to periodic functions of Theorem 2 of Rudi et al. (2020) (more specifically, of Corollary 2, page 23) .

Lemma 6 (Rudi et al. (2020)) *Let f be an $m+2$-times differentiable non-negative periodic function. Assume that the minimizers of f in $\mathcal{X}$ are finitely many and with strictly positive Hessian. Then, there exists $Q \in \mathbb{N}$ and $z_1, \ldots, z_Q$ periodic m-times differentiable functions, such that $f = \sum_{j=1}^{Q} z_j^2$.*

The proof of the lemma above is reported in Appendix A. Now we are ready to specialize Theorem 5 to the case of an m-times differentiable function.

Theorem 7 *Let f be an $(m + d/2 + 2)$-times differentiable periodic function, with $m > 0$ and let c_* be its global minimum. Assume that the minimizers of f in $\mathcal{X}$ are finitely many and with strictly positive Hessian. Then, for any $t \in \mathbb{N}$,*

$$\min_{g \in \mathcal{G}_t} \|f - c_* - g\|_F \;\leq\; C_f \, t^{-m},$$

where the constant C_f depends only on f, m, d.

The proof is self-contained and reported in Appendix B. It is obtained by first applying Lemma 6 on f, and Theorem 5 on the resulting characterization. To make this possible and to obtain a sharp rate, a crucial step is to show that the resulting functions belong to the space H_ρ for a specific ρ satisfying $\rho_k \propto |k|^{-2m-d}$, then deriving the bound on the associated residual R_t.

2.2. The Resulting Problem and the Associated A Priori Guarantees

Now the problem Eq. (3), with the PSD models (4) and the F norm, takes the following form

$$\bar{c} = \max_{c \in \mathbb{R},\ A \in \mathbb{C}^{n \times n}} c - \|f - c - g_A\|_F \quad \text{such that} \quad A \succeq 0, \tag{5}$$

and, combining Theorem 2 and Theorem 7 gives the following a priori guarantee

Corollary 8 *Let f be an $(m + d/2 + 2)$-times differentiable, 1-periodic function with $m > 0$, and let c_* be its global minimum. Also, let f have finitely many minimizers in $\mathcal{X}$, which each have strictly positive Hessian. Then, for any $t \in \mathbb{N}$*

$$0 \;\leq\; c_* - \bar{c} \;\leq\; C_f \, t^{-m}.$$

Expressing t with respect to n, the dimension of the matrix A, we have $t = O(n^{1/d})$. The bound above, then reads as

$$0 \;\leq\; c_* - \bar{c} \;\leq\; C' \, n^{-m/d}.$$

This shows that the solution, $\bar{c}$ is always a lower bound of the global minimum c_* and converges to c_* with a rate depending on the dimension of the matrix A and the degree of differentiability of f. E.g., when $m \geq d$, the error goes to zero as quick as n^{-1}. In the following section we see how to solve the optimization problem Eq. (5) in practice, by making use of the fact that $\|\cdot\|_F$, in the case of the torus, is a sum, which makes it easy to write Eq. (5) as a stochastic optimization objective.

3. Solving the Optimization Problem

We now describe the process of solving the optimization problem Eq. (3) in the specific case of the F norm and a PSD model $g_A(x) = \phi(x)^* A \phi(x)$ parametrized by positive semidefinite $A \in \mathbb{C}^{n \times n}$. For now, we consider an arbitrary feature map ϕ, but we will also contextualize our results in the specific case of ϕ_t, the map introduced in Section 2.1. A serious challenge to solving (5) is that computing $\|f - c - g_A\|_F$ or its subgradients exactly will typically be intractable because the F norm is the series:

$$\|f - c - g_A\|_F = \sum_{k \in \mathbb{Z}^d} |\hat{f}_k - c \mathbb{1}_{\{k=0\}} - \widehat{g_A}_k|.$$

To circumvent this issue, we recast the problem as a stochastic optimization objective. In particular, we introduce a probability measure, π, supported on $\mathbb{Z}^d$ and rewrite

$$\|f - c - g_A\|_F = \sum_{k\in\mathbb{Z}^d} \pi_k \frac{|\hat{f}_k - c\mathbb{1}_{\{k=0\}} - \widehat{g_A}_k|}{\pi_k} = \mathbb{E}_{k\sim\pi}\left[\frac{|\hat{f}_k - c\mathbb{1}_{\{k=0\}} - \widehat{g_A}_k|}{\pi_k}\right].$$

Written this way, we can now attack our objective using any number of methods from our stochastic optimization arsenal, such as projected stochastic gradient ascent.

To see how π should be chosen, we first note that $g_A(x) = \langle A,\, \phi(x)\phi(x)^*\rangle$, and use $M^{(k)} = \widehat{\phi\phi^*}_k \in \mathbb{C}^{n\times n}$ to denote the k-th Fourier component of $\phi\phi^*$, so that $\widehat{g_A}_k = \left\langle A,\, M^{(k)}\right\rangle$. Thus, our optimization problem now reads

$$\bar{c} = \max_{c\in\mathbb{R},\, A\in\mathbb{C}^{n\times n}} c - \mathbb{E}_{k\sim\pi}\left[\frac{|\hat{f}_k - c\mathbb{1}_{\{k=0\}} - \left\langle A,\, M^{(k)}\right\rangle|}{\pi_k}\right] \quad \text{such that } \; A \succeq 0.$$

Noting that c only appears in two terms, we can also eliminate this variable by solving

$$\max_c \; c - |\hat{f}_0 - c - \langle A,\, M^{(0)}\rangle| = \hat{f}_0 - \langle A,\, M^{(0)}\rangle.$$

Putting this all together, we want to solve the stochastic concave maximization problem

$$\bar{c} = \max_{A\succeq 0} \; \mathbb{E}_{k\sim\pi}[L_k(A)] \tag{6}$$

where

$$L_k(A) = \begin{cases} \frac{1}{\pi_0}\left(\hat{f}_0 - \left\langle A,\, M^{(0)}\right\rangle\right) & k = 0 \\ \frac{-1}{\pi_k}\left|\hat{f}_k - \left\langle A,\, M^{(k)}\right\rangle\right| & k \neq 0. \end{cases} \tag{7}$$

Using projected stochastic gradient ascent yields the following error guarantee:

Algorithm 1 PROJECTED STOCHASTIC GRADIENT ASCENT

1: Initialize $A_0 = 0$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: $\quad \tilde{A}_{t+1} = A_t + \eta\nabla L_{k_t}(A_t) \quad$ for $\; k_t \sim \pi$
4: $\quad A_{t+1} = \operatorname{argmin}_A \|A - \tilde{A}_{t+1}\|_{Frob.}$ s.t. $A \succeq 0$, $\|A\|_{Frob.} \leq R$.
5: **Return:** $\bar{A}_T = \frac{1}{T}\sum_{t=1}^T A_t$

Theorem 9 *Let $R \geq \|A^*\|_{Frob.}$ upper bound the norm of a maximizing A^*, and let $\bar{A}_T$ be the output of Algorithm 1, with constant stepsize $\eta = RT^{-1/2}/(1 + \sum_{k\in\mathbb{Z}^d}\|M^{(k)}\|_{Frob.})$ and $\pi_k \propto \|M^{(k)}\|_{Frob.} + (1 + \sum_{j=1}^d (2\pi k_j)^{d+1})^{-1}$. Then $\|\bar{A}_T\|_{Frob.} \leq R$, and for any $\delta \in (0,1)$, with probability $1-\delta$*

$$\mathbb{E}_{k\sim\pi}\left[L_k(\bar{A}_T)\right] \geq \bar{c} - \frac{20R\log(2/\delta)\left(1 + \sum_{k\in\mathbb{Z}^d}\|M^{(k)}\|_{Frob.}\right)}{\sqrt{T}}.$$

The proof, which we defer to Appendix C, simply requires proving that the functions L_k are Lipschitz-continuous and then appealing to Proposition 2.2 from Nemirovski et al. (2009). This result bounds, a priori, the optimization error incurred in trying to estimate A^* which realizes the maximum of (5). In Section 5, we combine this with Theorem 7 and the yet to be presented Theorem 10 to state our a priori guarantees.

The computational complexity of each iteration of projected stochastic gradient ascent is dominated by the cost of the projection on line 4 of Algorithm 1. This projection can be implemented in time proportional to $O(n^3)$ by first computing the eigenvalue decomposition of $\tilde{A}_{t+1}$, zeroing-out any negative eigenvalues, and then projecting this truncated matrix onto the radius-R Frobenius norm ball.

4. A Posteriori Certification

Obviously, it is nice to know a priori that our estimate $\bar{A}_T$ will be close to attaining the optimum of (5). However, with this estimate in hand, what we really want is to compute a lower bound on c_*, so we need to actually evaluate $\mathbb{E}_{k\sim\pi}[L_k(\bar{A}_T)]$, which is non-trivial since π has infinite support.

Things are easier when f and $\phi\phi^*$ are *band-limited*, meaning that for some K, $|k| > K$ implies $\hat{f}_k = 0$ and $M^{(k)} = 0$. Specifically, we can choose $\pi_k \propto \|M^{(k)}\|_{Frob.}$, which is only supported on $\{k : |k| \leq K\}$, and then easily compute $\mathbb{E}_{k\sim\pi}[L_k(\bar{A}_T)]$ to obtain an exact lower bound on c_*.

However, if one or both of f and $\phi\phi^*$ are not band-limited, then we are forced to estimate the value of an infinite sum. One approach is to draw samples $k \sim \pi$ and estimate the value using a sample average, and under suitable conditions on f and the matrices $M^{(k)}$, this allows us to accurately estimate the value of the lower bound with high-probability. Alternatively, under stronger conditions on f and the matrices $M^{(k)}$, we can compute L_k for a finite set of k's and deterministically bound the contribution of the remaining, uncomputed terms. The following Theorem indicates the accuracy of these methods:

Theorem 10 *Let f satisfy the conditions of Theorem 7 with $m > d/2$. Then for any K and $k_1, \ldots, k_K \overset{i.i.d.}{\sim} \pi$, for any $\delta \in (0,1)$ and $A \succeq 0$, with probability $1-\delta$,*

$$c_* \geq \bar{c} \geq \hat{c}_{1-\delta} := \frac{1}{K}\sum_{i=1}^{K} L_{k_i}(A) - Err_{1-\delta} \geq \mathbb{E}_{k\sim\pi}[L_k(A)] - 2Err_{1-\delta}$$

$$Err_{1-\delta} := (\sqrt{d+1}\|f\|_{C^{d+1}(\mathcal{X})} + \|A\|_{Frob.})\sqrt{\frac{2\log(2/\delta)}{K}}\Big(1 + \sum_{k\in\mathbb{Z}^d}\|M^{(k)}\|_{Frob.}\Big).$$

where $\|f\|_{C^{d+1}(\mathcal{X})} = \max_{1\leq j\leq d}\max_{1\leq q\leq d+1}\|\frac{\partial^q}{\partial x_j^q}f\|_{L^\infty(\mathcal{X})}$. In addition, for any K, the following holds deterministically:

$$c_* \geq \bar{c} \geq \hat{c}_1 := \sum_{k:|k|\leq K}\pi_k L_k(A) - Err_1 \geq \mathbb{E}_{k\sim\pi}[L_k(A)] - 2Err_1$$

$$Err_1 := \sum_{k:|k|>K}\Big[|\hat{f}_k| + \|A\|_{Frob.}\|M^{(k)}\|_{Frob.}\Big]$$

The proof, which we defer to Appendix D, analyzes $\hat{c}_{1-\delta}$ and $\hat{c}_1$ separately. For the former, we first show that $L_k(A)$ is bounded for each k, and then apply Hoeffding's inequality. For the latter, we

decompose the sum over $k \in \mathbb{Z}^d$ into those ks with $|k| \leq K$ and those ks with $|k| > K$, and then upper bound this second portion of the sum.

The Theorem shows that the sample average has additive error that decays with $1/\sqrt{K}$ with high probability. Furthermore, this lower bound is tractable given enough knowledge of our feature map for us to upper bound $\sum_{k \in \mathbb{Z}^d} \|M^{(k)}\|_{Frob.}$. For the deterministic lower bound on c_*, we need to have some control over how quickly $\|M^{(k)}\|_{Frob.}$ decays with increasing $|k|$, but if the feature map is chosen so that this decay is (eventually) rapid, then this lower bound can be tight. In the particular case of $\mathcal{G}_t$ introduced in Section 2.1, we show in Appendix E that we can bound $\sum_{k \in \mathbb{Z}^d} \|M^{(k)}\|_{Frob.} \leq n(4e)^{2d} t^{2d}$, and for $K \geq 2t$, $\sum_{k:|k|>K} \|M^{(k)}\|_{Frob.} = 0$. Therefore, $\hat{c}_{1-\delta}$ can provide a tight approximation of $\bar{c}$ using $K \gg n(4e)^{2d} t^{2d}$ samples, and $\hat{c}_1$ can once the maximum bandwidth is set $K \geq 2t$, as long as the Fourier coefficients $\hat{f}_k$ decay sufficiently quickly. Of course, for both the stochastic or deterministic estimate, we require information about the objective—whether it is a bound on $\|f\|_{C^{d+1}(\mathcal{X})}$ or on the decay of the high-frequency Fourier components of f. Although these particular pieces of information might not be strictly necessary, it is necessary to know *something* about f in order to have any hope of success. Furthermore, our results hold in terms of any upper bound on $\|f\|_{C^{d+1}(\mathcal{X})}$ or the high-frequency Fourier components, even very crude ones.

With Theorem 10, we can use the solution returned by our optimization algorithm to compute a lower bound on c_*, one that holds with high probability and one that holds deterministically. However, to actually compute a certificate of the accuracy of our lower bound, we also need an upper bound on c_*. Getting *some* upper bound on c_* is as easy as evaluating $f(x)$ at any point x, although most xs will not be close to minimizing f, so this may not give us much information about c_*. Of course, there are many better ways, and the bulk of the non-convex optimization literature is devoted to designing algorithms for computing approximate minimizers of f, i.e. upper bounds on c_*. Upper bounds for c_* are easier to produce, $f(x_0)$ for any point $x_0 \in \mathcal{X}$ is a valid upper bound. We can use the point x_0 produced for example by the method proposed by Rudi et al. (2020, Section 7), that converges provably to a global minimizer with a rate that avoids the curse of dimensionality. In our experiments (in low dimensions), we simply compute $f(x_1), \ldots, f(x_N)$ for N random points and upper bound $c_* \leq \min_i f(x_i)$, which allows for tight enough certificates.

5. A Priori and A Posteriori Guarantees

In the previous sections, we have described a method for estimating c_* in the case of periodic functions on $[0, 1]^d$, and all the pieces are in place to state our method's a priori and a posteriori guarantees. To summarize so far:

1. Theorem 2 shows that the solution of the relaxed problem (3), $\bar{c}$, is a lower bound on c_* which is tight up to the error of approximating $f - c_*$ with the class of non-negative functions, $\mathcal{G}$.
2. In Theorem 7, we bound this approximation error for smooth, periodic functions with respect to $\mathcal{G}_t$, the class of PSD models defined in (4) with the band-limited kernel ϕ_t.
3. But, we need to actually solve (3) defined using the F norm and $\mathcal{G}_t$. So, in Theorem 9, we bound the optimization error of the solution returned by projected stochastic gradient ascent.
4. However, our optimization algorithm returns the parameters $\bar{A}_T$ of a PSD model, and to compute a lower bound on c_* we need to actually evaluate the value of the objective at $\bar{A}_T$. So, finally, Theorem 10 bounds the estimation error when using $\bar{A}_T$ to estimate lower bounds $\hat{c}_{1-\delta}$ and $\hat{c}_1$ on c_* that holds with high probability and deterministically, respectively.

Therefore, our a priori guarantees amount to combining (Approximation Error) + (Optimization Error) + (Estimation Error). On the other hand, given any PSD model parameters, A, we can evaluate an a posteriori bound on the error by upper bounding $c_* \leq f(x)$ for any x and lower bounding c_* using Theorem 10. The following Corollary summarizes these guarantees:

Corollary 11 *For the F norm and family of PSD models $\mathcal{G}_t$ defined using ϕ_t, under the conditions of Theorems 7, 9, and 10, let $\hat{c}_{1-\delta}(\bar{A}_T)$ and $\hat{c}_1(\bar{A}_T)$ be lower bound estimates defined in Theorem 10. Then for any $\delta \in (0,1)$, we provide the following a priori guarantee with probability $1-2\delta$:*

$$c_* \geq \hat{c}_{1-\delta}(\bar{A}_T) \geq c_* - C_f t^{-m} - C_d n t^{2d}\left(\frac{20R\log(2/\delta)}{\sqrt{T}} + \frac{2(\sqrt{d+1}\|f\|_{C^{d+1}(\mathcal{X})} + R)\sqrt{2\log(2/\delta)}}{\sqrt{K}}\right),$$

with $C_d = (4e)^{2d}$. At the same time, given any point x and parameters A for the PSD model, we guarantee a posteriori that $f(x) \geq c_* \geq \hat{c}_1(A)$ and $f(x) \geq c_* \geq \hat{c}_{1-\delta}(A)$ with probability $1-\delta$.

The Corollary follows immediately by combining Theorems 7, 9, and 10. Since $t = O(n^{1/d})$, by choosing $T = K = O(n^{6+2m/d})$, we have

$$c_* \geq \hat{c}_{1-\delta}(\bar{A}_T) \geq c_* - C' n^{-m/d},$$

when n is the dimension of the matrix $\bar{A}_T$. In this case the algorithm has a complexity that is $O(Tn^3 + Kn^2) = O(n^{9+2m/d})$. In particular, for the class of $m + d/2 +$ 2-times differentiable functions with $m > d/2$, we achieve a bound $c_* \geq \hat{c}_{1-\delta}(\bar{A}_T) \geq c_* - C' n^{-1}$, with a computational cost of $O(n^{11})$. There is a lot of room for improvement in the constants of the exponents, but the considered algorithm shows that it is possible to obtain the global optimum of a function with both a posteriori guarantees and an a priori error rate that is adaptive to the degree of differentiability of the function to minimize and that avoids the curse of dimensionality for very smooth functions.

6. Empirical Evaluation

Finally, we apply our method to two simple non-convex optimization problems in one and two dimensions. The results are summarized in Figures 1 and 2, and all of the details of the experiments are deferred to Appendix F, in which we describe a new feature map ϕ, and describe a more practical algorithm for solving (5) based on reparametrizing $A = UU^*$ (Burer and Monteiro, 2003).

7. Discussion

Convex duality. Following Rudi et al. (2020), we can also provide a dual interpretation to the use of PSD models. Indeed, the minimimization problem we solve can be written in the form

$$\inf_{\mu \text{ probability measure}} \int_{\mathcal{X}} f(x) d\mu(x)\,,$$

which we can reformulate as

$$\inf_{\mu \text{ signed measure}} \int_{\mathcal{X}} f(x) d\mu(x) \text{ such that } \int_{\mathcal{X}} d\mu(x) = 1 \text{ and } \int_{\mathcal{X}} \Phi(x)\Phi(x)^* d\mu(x) \succcurlyeq 0.$$

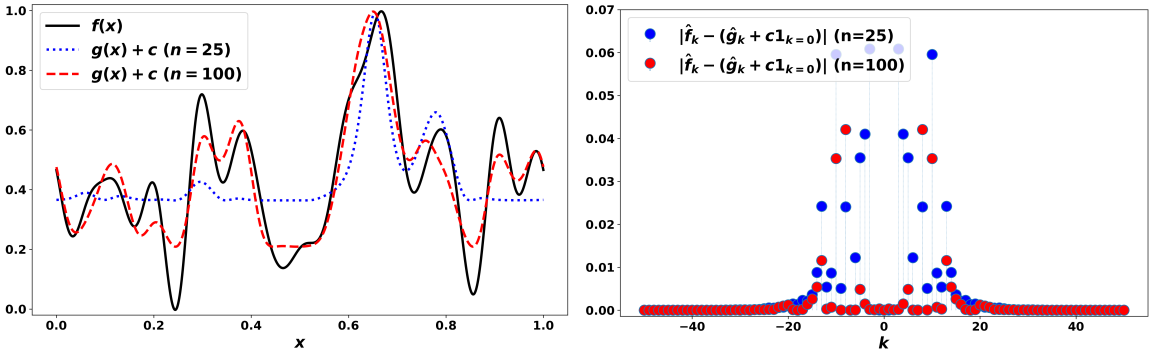

Figure 1: For $f : \mathbb{R} \to \mathbb{R}$ as described in Appendix F, we use PSD models using feature maps of dimension $n = 25$ and $n = 100$. To the left, we see that for $n = 25$, the model $g(x) + c$ does not approximate f well, so our a posteriori error guarantee is 0.24. But, when $n = 100$, the model approximates f much better, and our a posteriori guarantee is 0.01. To the right, we plot the absolute difference between $\hat{f}_k$ and $\widehat{(g + c)}_k$ for small ks, which is what drives the difference in performance between $n = 25$ and $n = 100$.

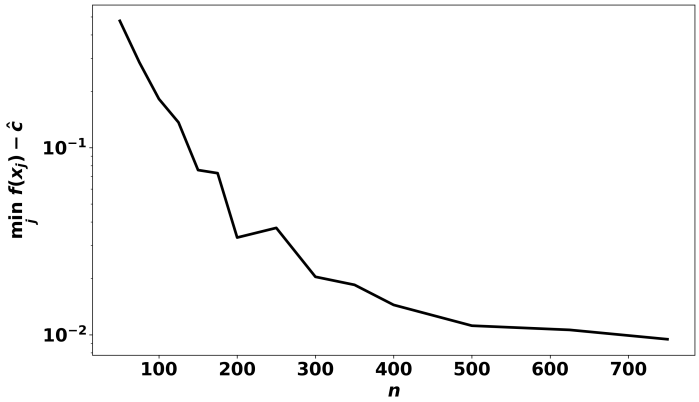

Figure 2: For $f : \mathbb{R}^2 \to \mathbb{R}$ as described in Appendix F, we plot the a posteriori error guarantee of our algorithm's estimate vs. n, i.e $\min_j f(x_j) - \hat{c}$ which is the difference between the minimum value of f achieved on a random grid of points, which upper bounds c_*, and our estimate of the minimum, $\hat{c}_1$, which lower bounds c_*. The error is at most 10% of the function's range with $n = 150$, and can be made less than 1% with $n = 750$.

Since we expect the solution of the original problem to consist of Dirac measures supported on global minimizers, we are free to add constraints that are satisfied by these Diracs, such as,

$$\int_{\mathcal{X}} |d\mu(x)| \leqslant 1 \text{ or } \Omega(\mu) \leqslant 1,$$

for any norm Ω on signed measure that is larger than the total variation norm. The first constraint leads to a dual problem

$$\sup_{c\in\mathbb{R},\ B\succcurlyeq 0} c - \left\| f - c1 - \phi(\cdot)^\top B\phi(\cdot)\right\|_\infty,$$

while the second one leads to

$$\sup_{c\in\mathbb{R},\ B\succcurlyeq 0} c - \Omega^*\big(f - c1 - \phi(\cdot)^\top B\phi(\cdot)\big).$$

It turns out that the dual of the S norm and of F norm dominate the total variation norm, and thus have this dual interpretation. Thus, our method for obtaining a posteriori certificates directly extends to optimization problems that are defined through probabilty measures and already tackled by kernel sum-of-squares, such as optimal transport (Vacher et al., 2021), or optimal control (Berthier et al., 2021).

Comparison to previous work on kernel sums-of-square. Compared to Rudi et al. (2020), in this work the subsampling is different: the constraint $\int_{\mathcal{X}} \phi(x)\phi(x)^* d\mu(x) \succcurlyeq 0$ is replaced by the projection on the span of $\phi(x^{(1)}, \dots, \phi(x^{(n)})$ being a positive semidefinite matrix. This is a relaxation in the dual, which still leads to a lower bound for the optimization problem.

This also suggests a candidate optimal solution when applied to the torus. Indeed, at optimality, we expect μ to be close to a Dirac at x_0, and then (in 1D for simplicity), $\hat{\mu}_1$ should be close to $e^{-2i\pi x_0}$, and we can read off a candidate minimizer as the argument of the first Fourier coefficients (we could imagine using more than one).

Acknowledgments

This work was supported by the French government under the management of the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grants SEQUOIA 724063 and REAL 947908).

References

Eloïse Berthier, Justin Carpentier, Alessandro Rudi, and Francis Bach. Infinite-dimensional sums-of-squares for optimal control. Technical Report 2110.07396, arXiv, 2021.

Richard A. Brualdi. *Introductory Combinatorics. Fifth Edition.* Pearson, 2009.

Samuel Burer and Renato D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Loukas Grafakos. *Classical Fourier Analysis*, volume 2. Springer, 2008.

Jean-Bernard Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. *Advances in Neural Information Processing Systems*, 33:12816–12826, 2020.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Arkadi Nemirovski, Shmuel Onn, and Uriel G. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

Erich Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349. Springer, 2006.

Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003.

Alessandro Rudi and Carlo Ciliberto. PSD representations for effective probability models. *Advances in Neural Information Processing Systems*, 34, 2021.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. *arXiv preprint arXiv:2012.11978*, 2020.

Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Learning Theory*, pages 4143–4173, 2021.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Appendix A. Proof of Lemma 6

Proof The proof of the adaptation of Corollary 2 of Rudi et al. (2020) to the periodic setting is organized in four steps that are summarized in this paragraph. By translation $\tilde{f}(x) = f(x - b)$ for a suitable vector $b \in \mathcal{X}$, we show that all the global minima are contained in a closed set $\mathcal{A}$ strictly contained in an open set Ω strictly contained in the closed set $[0, 1]^d$. Then Corollary 2 of Rudi et al. (2020) shows that there exist q functions $u_1, \dots u_q \in C^{m+2}(\mathbb{R}^d)$ such that $\tilde{f}(x) = \sum_{j=1}^{q} u_j(x)^2$ for any $x \in \Omega$. Then we build two bump functions such that $0 \leq \eta, \nu \leq 1$ and $\eta^2 + \nu^2 = 1$ on $\mathbb{R}^d$, moreover $\eta = 0$ on $\mathcal{X} \setminus \Omega$ and 1 on $\mathcal{A}$ (so $\nu = 1$ on $\mathcal{X} \setminus \Omega$ and 0 on $\mathcal{A}$). Then we create a new function $\tilde{v} = \sqrt{\tilde{f}} \cdot \nu$ and we show that its periodic extension $\tilde{v}_{\text{per}}$ satisfies $\tilde{v}_{\text{per}} \in C^{m+2}(\mathbb{R}^d)$. We prove the same for $\tilde{u}_{j,\text{per}}$, the periodic extension of $\tilde{u}_j = u_j \cdot \nu$. The result is obtained by noting that $\tilde{u}_{1,\text{per}}, \dots, \tilde{u}_{q,\text{per}}, \tilde{v}_{\text{per}}$ are m-times differentiable periodic functions that satisfy $f(x) = \tilde{v}_{\text{per}}(x + b)^2 + \sum_j \tilde{u}_{j,\text{per}}(x + b)^2$ for any $x \in \mathbb{R}^d$.

Step 1. Translating the functions, applying Corollary 2 of Rudi et al. (2020). Let $s \in \mathbb{N}$ and $M = \{x_1, \dots, x_s\}$ be the set of minimizers on $[0,1)^d$. First, we work with a translated periodic function $\tilde{f} = f(x-b)$, where $b = (\tau/2, \dots, \tau/2) \in \mathbb{R}^d$ and τ is the minimum distance of a point in M from $[0,1]^d \setminus [0,1)^d$ (note that $\tau > 0$ by construction). Let $\mathcal{A} = [2\tau/3, 1-2\tau/3]^d$ and $\Omega = \cup_{x \in (\tau/2, 1-\tau/2)^d} B_{\tau/6}(x)$, where $B_{\tau/6}(x)$ is the open ball of radius $\tau/6$ centered in x. Note that $\mathcal{A} \subset \Omega \subset \mathcal{X}$ and that $\mathcal{A}, \mathcal{X}$ are closed, while Ω is open. Now in the translated version $\tilde{f}$ all the zeros are in $\mathcal{A}$. By applying Corollary 2 of Rudi et al. (2020) on $\tilde{f}$ and Ω, we obtain that there exists $q \in \mathbb{N}$ and $u_1, \dots, u_q \in C^m(\mathbb{R}^d)$ such that

$$\tilde{f}(x) = \sum_{j=1}^{q} u_j(x)^2, \quad \forall x \in \Omega.$$

Step 2. Building the bump functions. Let now α, β be two infinitely differentiable non-negative functions on $\mathbb{R}^d$ such that $\alpha = 0$ on $\mathbb{R}^d \setminus \Omega$ and is strictly positive on Ω, while $\beta = 0$ on $\mathcal{A}$, strictly positive on $\mathbb{R}^d \setminus \mathcal{A}$. Since $\alpha^2 + \beta^2 > 0$ on $\mathbb{R}^d$, the function $\sqrt{\cdot} \in C^\infty((0,\infty))$ and $\alpha, \beta \in C^\infty(\mathbb{R}^d)$, then $\eta = \alpha/\sqrt{\alpha^2+\beta^2}$ and $\nu = \beta/\sqrt{\alpha^2+\beta^2}$ are $C^\infty(\mathbb{R}^d)$ and satisfy $\eta = 0$ on $\mathbb{R}^d \setminus \Omega$, $\eta = 1$ on $\mathcal{A}$, analogously $\nu = 1$ on $\mathbb{R}^d \setminus \Omega$ and 0 on $\mathcal{A}$, moreover $\eta^2 + \nu^2 = 1$ on $\mathbb{R}^d$.

Step 3. Construction of $\tilde{v}_{\text{per}}$ and $\tilde{u}_{j,\text{per}}$. Let $C = [\tau, 1-\tau]^d$. By construction, $\{x_1, \dots, x_s\} \subset C$ and so $\tilde{f} > 0$ on the set $\mathcal{X} \setminus C$. Then $(\tilde{f})^{1/2} \in C^{m+2}(\mathcal{X} \setminus C)$ since it is the composition of $\sqrt{\cdot} \in C^\infty((0,\infty))$ and $\tilde{f}$ that is $\tilde{f} > 0$ on $\mathcal{X} \setminus C$. Then $\tilde{v} = (\tilde{f})^{1/2} \cdot \nu \in C^{m+2}(\mathcal{X})$, since $(\tilde{f})^{1/2}$ is $m+2$ times differentiable on the set $\mathcal{X} \setminus C$ and ν is infinitely differentiable and 0 on $\mathcal{A} \supset\supset C$. Denote by $\tilde{v}_{\text{per}}$ the periodic extension of $\tilde{v}$, i.e. $\tilde{v}_{\text{per}}(x+k) = \tilde{v}(x)$ for any $x \in \mathcal{X}$ and $k \in \mathbb{Z}^d$.

We now prove that $\tilde{v}_{\text{per}}$ satisfies $\tilde{v}_{\text{per}} \in C^{m+2}(\mathbb{R}^d)$. First note that it is $m+2$ times differentiable in the interior of each cube $\mathcal{X}+k$ for $k \in Z^d$, since $\tilde{v}$ has this property on the interior of $\mathcal{X}$. Moreover it has the same property also in a neighbourhood of the set $S + k$ with $S = [0,1]^d \setminus (0,1]^d$ and $k \in \mathbb{Z}^d$. Indeed, let $B_{\tau/6}(x+k)$ be the open ball of radius $\tau/6$ around $x+k$, with $x \in S$ and $k \in \mathbb{Z}^d$. On $B_{\tau/6}(x+k)$ the function $\tilde{v}$ is equal to $(\tilde{f})^{1/2}$, which in that region is $m+2$ times differentiable, since we are in a translation of $\mathcal{X} \setminus \Omega$.

Define now $\tilde{u}_j = u_j \cdot \eta$ and denote by $\tilde{u}_{j,\text{per}}$ its periodic extension. Similary to the case of $\tilde{v}_{\text{per}}$, since $\tilde{u}_j$ is identically 0 on $\mathcal{X} \setminus \Omega$ and it is m-times differentiable on $\mathcal{X}$, we can prove that the periodic extension of $\tilde{u}_j$ satisfies $\tilde{u}_{j,\text{per}} \in C^m(\mathbb{R}^d)$.

Step 4. Conclusion. Note that $\mathbb{R}^d = \cup_{k \in \mathbb{Z}^d} \{\mathcal{X} + k\}$. Then, by expanding the definitions, we have that for all $x \in \mathcal{X}, k \in \mathbb{Z}^d$,

$$\begin{aligned}
\tilde{v}_{\text{per}}(x+k)^2 + \sum_{j=1}^{q} \tilde{u}_{j,\text{per}}(x+k)^2 &= \tilde{v}_j(x)^2 + \sum_{j=1}^{q} \tilde{u}_j(x)^2 \\
&= ((\tilde{f})^{1/2}(x)\nu(x))^2 + \sum_{j=1}^{q} (u_j(x)\eta(x))^2 = \tilde{f}(x)\nu(x)^2 + \eta(x)^2 \sum_{j=1}^{q} u_j(x)^2 \\
&= \begin{cases} \tilde{f}(x) & x \in \mathcal{X} \setminus \Omega \\ \tilde{f}(x)\nu(x)^2 + \eta(x)^2 \sum_{j=1}^{q} u_j(x)^2 & x \in \Omega \setminus \mathcal{A} \\ \sum_{j=1}^{q} u_j(x)^2 & x \in \mathcal{A} \end{cases} \\
&= \tilde{f}(x),
\end{aligned}$$

where in the last two steps we use the fact that $\sum_j u_j(x)^2 = \tilde{f}(x)$ on $x \in \Omega$ and the fact that $\eta^2 + \nu^2 = 1$ everywhere and, in particular, $\eta = 0$ on $\mathcal{X} \setminus \Omega$, and on $\mathcal{A}$, while $\nu = 1$ on $\mathcal{X} \setminus \Omega$ and 0 on $\mathcal{A}$. The proof is concluded by taking the translated version of by the vector b. I.e. $z_j(x) = \tilde{u}_{j,\text{per}}(x+b)$ for any $x \in \mathbb{R}^d$ and $j = 1, \dots, q$ and moreover $z_{q+1}(x) = \tilde{v}_{\text{per}}(x+b)$ for any $x \in \mathbb{R}^d$, and $Q = q + 1$. ■

Appendix B. Proof of Theorem 7

Proof First, let $s = m + d/2$. Note that, by applying Lemma 6 to f, we have that there exist $Q \in \mathbb{N}$ functions $u_1, \dots, u_Q$, that are periodic, s-times differentiable and which provide a new characterization of f as $f = \sum_{j=1}^{Q} u_j^2$. The desired result is obtained by applying Theorem 5 to this characterization of f. To apply Theorem 5, we need to find a suitable $\rho \in \ell_1(\mathbb{Z}^d)$ such that H_ρ contains $u_1, \dots, u_Q$. In particular, we choose $\rho_k = (1 + \sum_{j=1}^{d} (2\pi k_j)^s)^{-2}$. We prove now that $u_j \in H_\rho$ and we characterize the resulting convergence rate.

Denote $i = \sqrt{-1}$ and by $\partial_j^q u$ the function $\partial_j^q u = \frac{\partial^q}{\partial x_j^q} u$, for all $j \in \{1, \dots, d\}$ and $q \in \{0, \dots, s\}$, and u an s-times differentiable periodic function. Denote by $\|u\|_{C^s(\mathcal{X})}$ the following norm $\|u\|_{C^s(\mathcal{X})} = \max_{1 \le j \le d} \max_{0 \le q \le s} \|\partial_j^q u\|_{L^\infty(\mathcal{X})}$. With the notation we are using of the Fourier series we have $\widehat{(\partial_j^q u)}_k = (2\pi i k_j)^q \widehat{u}_k$ (see, e.g., Prop. 3.1.2 of Grafakos, 2008). Now, by the Plancherel's identity (Prop 3.1.16 of Grafakos, 2008) and the fact that $\mathcal{X}$ has volume equal to 1,

$$\sum_{k \in \mathbb{Z}^d} (2\pi k_j)^{2q} |\widehat{u}_k|^2 = \sum_{k \in \mathbb{Z}^d} |(2i\pi k_j)^q \widehat{u}_k|^2 = \int_{\mathcal{X}} |\partial_j^q u|^2 dx \le \|\partial_j^q u\|^2_{L^\infty(\mathcal{X})} \le \|u\|^2_{C^s(\mathcal{X})} < \infty,$$

where the norm $\|u\|_{C^s(\mathcal{X})}$ is finite, since, for any u that is periodic and s-times differentiable, we have that $\frac{\partial^q}{\partial x_j^q} u$, with $q \le s$, is also continuous and periodic, so uniformly bounded on $\mathcal{X}$. Now, since $(\sum_{j=0}^{d} a_j)^2 \le c_d \sum_{j=0}^{d} a_j^2$ for any $a_j \ge 0$, with $c = d + 1$, by expanding the definition of ρ_k,

$$\sum_{k \in \mathbb{Z}^d} \frac{|\widehat{u}_k|^2}{\rho_k} = \sum_{k \in \mathbb{Z}^d} \Big| \widehat{u}_k + \sum_{j=1}^{d} (2\pi k_j)^s \widehat{u}_k \Big|^2 \le c \Big(\sum_{k \in \mathbb{Z}^d} |\widehat{u}_k|^2 + \sum_{j=1}^{d} \sum_{k \in \mathbb{Z}^d} |(2i\pi k_j)^s \widehat{u}|^2 \Big) \le c \|u\|^2_{C^s(\mathcal{X})}. \tag{8}$$

This proves that the functions $u_1, \dots, u_Q$, that are periodic and s-times differentiable, belong to the space H_ρ. Now we can apply Theorem 5, which gives $\min_{g \in \mathcal{G}_t} \|f - g\|_F \le C'_f R_t$, for any $t \in \mathbb{N}$, where now $C'^2_f = c \sum_{j=1}^{Q} \|u_j\|_F \|u_j\|_{C^m(\mathcal{X})}$ and R_t is bound as follows.

Since, $(\sum_{j=0}^{d} a_j^s)^2 \ge c_s (\sum_{j=0}^{d} a_j)^{2s}$ for any $a_j \ge 0$, with $c_s = (d+1)^{-2(s-1)}$, (see, e.g., page 11 of Grafakos, 2008), and the cardinality of the set of vectors in $\mathbb{Z}^d$ summing up to a given number corresponds to $\#\{k \mid |k| = r\} = \binom{r+d-1}{d-1} \le C_d r^{d-1}$ for any $r, d \in \mathbb{N}$ (e.g., page 52 of Brualdi, 2009), with $C_d = (2e)^{k-1}$, we have

$$R_t^2 = \sum_{|k| > t} \rho_k \le \sum_{|k| > t} \frac{1}{1 + c_s |k|^{2s}} = \sum_{r > t} \frac{\#\{k \mid |k| = r\}}{1 + c_s r^{2s}} \le \sum_{r > t} \frac{C_d r^{d-1}}{1 + c_s r^{2s}} \le \frac{C_d}{(s-d)\, c_s} t^{-(2s-d)},$$

where, in the last step, we used the fact that $\frac{C_d r^{d-1}}{1 + c_s r^{2s}} \le \frac{C_d}{c_s} r^{-(2s-d+1)}$, moreover, $\sum_{r > t} r^{-(2s-d+1)} \le \int_t^\infty x^{-(2s-d+1)} dx = \frac{t^{-(2s-d)}}{2s-d}$. The final constant is then $C_f^2 = C'^2_f C_d / ((s-d) c_s)$. ■

Appendix C. Proof of Theorem 9

Proof When L_k is concave and G-Lipschitz w.r.t. the Frobenius norm for all k, then the average of the iterates of projected stochastic gradient ascent with constant stepsize $\eta = B/(G\sqrt{T})$ applied to a problem of the form in Eq. (6) will have error bounded by (Nemirovski et al., 2009, Proposition 2.2)

$$\bar{c} - \mathbb{E}_{k\sim\pi}[L_k(\bar{A}_T)] \leq a_0 \cdot \frac{GR\log(2/\delta)}{\sqrt{T}} \tag{9}$$

with probability at least $1-\delta$. For our particular objective, it is easy to see that for all $k \in \mathbb{Z}^d$,

$$\sup_{A,A'} \frac{|L_k(A) - L_k(A')|}{\|A - A'\|_{Frob.}} \leq \frac{\|M^{(k)}\|_{Frob.}}{\pi_k}.$$

Therefore, with our choice of $\pi_k \propto \|M^{(k)}\|_{Frob.} + (1 + \sum_{j=1}^d (2\pi k_j)^{d+1})^{-1}$, we can bound the parameter of Lipschitz continuity for all k by

$$G \leq \sum_{k\in\mathbb{Z}^d} \left[\|M^{(k)}\|_{Frob.} + \left(1 + \sum_{j=1}^d (2\pi k_j)^{d+1}\right)^{-1} \right] \leq 1 + \sum_{k\in\mathbb{Z}^d} \|M^{(k)}\|_{Frob.}$$

Plugging this into Eq. (9) completes the proof. ■

Appendix D. Proof of Theorem 10

Proof First, we prove the high-probability bound. We begin by arguing that $L_k(A)$ is bounded. Let $\mu_k = (1 + \sum_{j=1}^d (2\pi k_j)^{d+1})^{-1}$ so that $\pi_k \propto \|M^{(k)}\|_{Frob.} + \mu_k$. For each $k \neq 0$, we have

$$\begin{aligned}
|L_k(A)| &= \frac{1}{\pi_k}\left|\hat{f}_k - \left\langle A,\, M^{(k)}\right\rangle\right| \\
&\leq \frac{\sum_{k'\in\mathbb{Z}^d}\left[\|M^{(k')}\|_{Frob.} + \mu_{k'}\right]}{\|M^{(k)}\|_{Frob.} + \mu_k}\left(|\hat{f}_k| + \|A\|_{Frob.}\|M^{(k)}\|_{Frob.}\right) \\
&\leq \left(1 + \sum_{k'\in\mathbb{Z}^d}\|M^{(k')}\|_{Frob.}\right)\left(\frac{|\hat{f}_k|}{\mu_k} + \|A\|_{Frob.}\right).
\end{aligned}$$

Now we need to bound $|\hat{f}_k|/\mu_k$. Note that, by applying Eq. (8), with $u = f$, $s = d+1$ and $\rho_k = \mu_k^2$, we have that for any $k \in \mathbb{Z}^d$,

$$\frac{|f_k|}{\mu_k} \leq \left(\sum_{k\in\mathbb{Z}^d}\frac{|f_k|^2}{\mu_k^2}\right)^{1/2} \leq \sqrt{d+1}\|f\|_{C^{d+1}(\mathcal{X})},$$

where $\|f\|_{C^{d+1}(\mathcal{X})} = \max_{j=1,\ldots,d} \max_{q=1,\ldots,d+1} \|\frac{\partial^q}{\partial x_j^q} f\|_{L^\infty(\mathcal{X})}$. The result then follows by Hoeffding's inequality: for any $\delta \in (0,1)$

$$\mathbb{P}\Bigg(\left|\mathbb{E}_{k\sim\pi}[L_k(A)] - \frac{1}{K}\sum_{i=1}^{K} L_{k_i}(A)\right| \geq (\sqrt{d+1}\|f\|_{C^{d+1}(\mathcal{X})} + \|A\|_{Frob.})\sqrt{\frac{2\log(2/\delta)}{K}}\Big(1 + \sum_{k\in\mathbb{Z}^d}\|M^{(k)}\|_{Frob.}\Big)\Bigg) \leq \delta.$$

Rearranging and noting that $\mathbb{E}_{k\sim\pi}[L_k(A)] \leq \bar{c} \leq c_*$, completes the first half of the proof.

For the second set of bounds, we note that

$$\begin{aligned}
\bar{c} &\geq \mathbb{E}_{k\sim\pi}[L_k(A)] \\
&= \sum_{k:|k|\leq K} \pi_k L_k(A) - \sum_{k:|k|>K} \left|\hat{f}_k - \left\langle A,\, M^{(k)}\right\rangle\right| \\
&\geq \sum_{k:|k|\leq K} \pi_k L_k(A) - \sum_{k:|k|>K} \Big(|\hat{f}_k| + \|A\|_{Frob.}\|M^{(k)}\|_{Frob.}\Big).
\end{aligned}$$

This completes the proof. ∎

Appendix E. Bound on $\|M^{(k)}\|_{Frob.}$ for ϕ_t

The k_1, k_2 entry of $(\phi\phi^*)(x)$ is equal to $e_{k_1}(x)e_{k_2}(x)^*$, so

$$\begin{aligned}
[M^{(k)}]_{k_1,k_2} &= \int_{[0,1]^d} e_{k_1}(x)e_{k_2}(x)^* e^{-2\pi i x^\top k} dx \\
&= \int_{[0,1]^d} e^{-2\pi i x^\top (k+k_1-k_2)} dx = \mathbb{1}_{\{k=k_2-k_1\}}.
\end{aligned}$$

Therefore, the entries of $M^{(k)}$ are bounded by 1, and if $|k| \geq 2t$, then $M^{(k)} = 0$. Therefore, we can bound $\|M^{(k)}\|_{Frob.} \leq n\mathbb{1}_{\{|k|\leq 2t\}}$. Since $\#\{|k| \leq 2t\} = \sum_{r=0}^{2t} \#\{|k| \leq r\} = \sum_{r=0}^{2t} \binom{2r+d-1}{d-1} = \binom{2t+d}{d}$ (see, e.g., Theorem 2.5.1 of Brualdi, 2009),

$$\sum_{k\in\mathbb{Z}^d} \|M^{(k)}\|_{Frob.} \leq n\binom{2t+d}{2t} \leq n(4e)^{2d}t^{2d}.$$

Appendix F. Experimental Details

Here, we describe how we applied our approach to two simple non-convex optimization problems to show its promise. As described, we can lower bound c_* by solving the stochastic concave maximization problem Eq. (6) using an algorithm like projected stochastic gradient ascent. However, each iteration requires projecting the algorithm's iterate onto the PSD cone, which is computationally expensive.

Therefore, following Burer and Monteiro (2003), we reparametrize $A = UU^*$, which is always positive semidefinite, using new parameters $U \in \mathbb{C}^{n\times n}$, yielding the *unconstrained* objective

$$\bar{c} = \max_{U\in\mathbb{C}^{n\times n}} \mathbb{E}_{k\sim\pi}[L_k(UU^*)]. \tag{10}$$

Due to the non-linear reparametrization, the objective is no longer concave, but just as Burer and Monteiro (2003) exhibit for linear SDPs, we find that stochastic gradient ascent on U succeeds for our problem when we optimize a smooth surrogate for our non-differentiable objective. Specifically, for $k \neq 0$, we replace

$$L_k(A) = \frac{-1}{\pi_k}\left|\hat{f}_k - \left\langle A,\, M^{(k)}\right\rangle\right| \to \tilde{L}_k(A) = \frac{-1}{\pi_k}\sqrt{(\alpha\pi_k)^2 + \left|\hat{f}_k - \left\langle A,\, M^{(k)}\right\rangle\right|^2},$$

where α is small scalar. Choosing α larger makes the objective smoother, but makes $\tilde{L}_k$ a worse approximation of L_k. In our experiments, we tune α, along with the other hyperparameters—including the stepsize, η and the number of iterations, T—with cross validation.

A random non-convex objective. We constructed a family of non-convex periodic functions on $[0,1]$ and $[0,1]^2$ to test our algorithm. The functions are defined in terms of their Fourier series, with $\hat{f}_k \sim \mathcal{N}(0, 1/(1+|k|)^2) + i\cdot\mathcal{N}(0, 1/(1+|k|)^2)$ for each k with $|k| \leq 15$ in the 1D case and $|k| \leq 4$ in the 2D case. We then adjust the Fourier components so that they satisfying the necessary property $\hat{f}_{k^*} = \hat{f}_k^*$. The value of f itself is then computed on a grid of points, $x_1, \ldots, x_N$, and is rescaled by dividing by $\max_i f(x_i) - \min_j f(x_j)$ so that f's range is of order 1.

A different feature map The feature map that we use for our experiments has the form $\phi_{n,\rho}(x) = \tilde{\phi}_{n,\rho}(x[1]) \circ \tilde{\phi}_{n,\rho}(x[2]) \circ \cdots \circ \tilde{\phi}_{n,\rho}(x[d])$, where $n \in \mathbb{N}$ and $\rho \in (0,1)$ are hyperparameters to be chosen later, $\tilde{\phi}_{n,\rho} : \mathbb{R} \to \mathbb{C}^n$, and $\circ$ denotes the hadamard product, so the feature map decomposes over the coordinates of x. To define $\tilde{\phi}_{n,\rho}$, we sample n points $x_1, \ldots, x_n$ uniformly at random from $[0,1]^d$, and set

$$\tilde{\phi}_{n,\rho}(x[i])[j] = \varphi_\rho(x[i] - x_j[i]), \qquad \varphi_\rho(x) = \sum_{k\in\mathbb{Z}} \rho^{|k|} e^{2\pi i k x}.$$

The function φ is chosen so that its Fourier components $\hat{\varphi}_k = \rho^{|k|}$ decay exponentially quickly with k. In Appendix F.1 below, we show how to compute the matrices $M^{(k)}$ that are needed to implement our algorithm, and we bound $\|M^{(k)}\|_{Frob.}$, which is needed to compute the a posteriori guarantees. For this feature map, larger n allows for a more expressive, but more computationally expensive PSD model and Figures 1 and 2 demonstrate the effect of n on our a posteriori accuracy guarantees in one and two dimensions, respectively. The parameter ρ, which we choose using cross-validation, clearly affects the Fourier components of the PSD model that we learn, with smaller ρ making them decay more quickly with k.

F.1. Analysis of the Feature Map

In what follows, we will drop the subscripts n and ρ and consider these hyperparameters to be fixed and arbitrary. We recall the definition

$$\tilde{\phi}(x[i])[j] = \varphi(x[i] - x_j[i]), \qquad \varphi(x) = \sum_{k\in\mathbb{Z}} \rho^{|k|} e^{2\pi i k x}.$$

We further note that

$$\begin{aligned}\varphi(x) &= \sum_{k\in\mathbb{Z}} \rho^{|k|} e^{2i\pi kx} = -1 + 2\cdot \mathrm{Re}\Big(\sum_{k\in\mathbb{N}} \rho^k e^{2i\pi kx}\Big) \\ &= -1 + 2\cdot \mathrm{Re}\Big(\frac{1}{1-\rho e^{2i\pi x}}\Big) = -1 + 2\frac{1-\rho\cos 2\pi x}{1+\rho^2-2\rho\cos 2\pi x} \\ &= \frac{1-\rho^2}{1+\rho^2-2\rho\cos 2\pi x}. \end{aligned} \tag{11}$$

In this Appendix, we show how to compute $M^{(k)}$, the kth Fourier component of $\phi\phi^*$, which is needed to implement our algorithm. Since $\phi(x) = \tilde{\phi}(x[1]) \circ \cdots \circ \tilde{\phi}(x[d])$ decomposes across coordinates, this essentially boils down to computing the 1D version d times and multiplying across dimensions.

So, for now we focus on the case $d = 1$, and attempt to compute

$$[M^{(k)}]_{ij} = [\widehat{\phi_i\phi_j}]_k.$$

Since $\phi_i(x) = \phi(x - x_i)$ and $\phi_j(x) = \phi(x - x_j)$, we need to know how to compute the k-th Fourier coefficient of $x \mapsto \varphi(x-y)\varphi(x-z)$. We have:

$$\begin{aligned}\varphi(x-y)\varphi(x-z) &= \sum_{n,m\in\mathbb{Z}} \rho^{|n|+|m|} e^{2i\pi n(x-y)+2i\pi m(x-z)} \\ &= \sum_{n,m\in\mathbb{Z}} \rho^{|n|+|m|} e^{-2i\pi ny-2i\pi mz} e^{2i\pi(n+m)x}.\end{aligned}$$

For simplification and by symmetry, we can consider $T(y,x) = \varphi(x-y)\varphi(x)$, so that $\phi(x - y)\phi(x - z) = T(y - z, x - z)$. Thus, this k-th Fourier coefficient is simply

$$\hat{T}(y)_k = \sum_{n+m=k} \rho^{|n|+|m|} e^{-2i\pi ny} = \sum_{n\in\mathbb{Z}} \rho^{|n|+|n-k|} e^{-2i\pi ny}.$$

Moreover, we will need to compute $e^{-2ik\pi z}\hat{T}(y - z)_k$. We directly have $\hat{T}(y)_{-k} = \hat{T}(y)_k^*$, so we can consider $k \geqslant 0$ and

$$\begin{aligned}\hat{T}(y)_k &= \sum_{n=-\infty}^{0} \rho^{k-2n} e^{-2i\pi ny} + \sum_{n=1}^{k-1} \rho^k e^{-2i\pi ny} + \sum_{n=k}^{+\infty} \rho^{2n-k} e^{-2i\pi ny} \\ &= \sum_{n=0}^{+\infty} \rho^{k+2n} e^{2i\pi ny} + \rho^k \sum_{n=1}^{k-1} e^{-2i\pi ny} + \sum_{n=k}^{+\infty} \rho^{2n-k} e^{-2i\pi ny} \\ &= \begin{cases} \rho^k\Big(\frac{1}{1-\rho^2 e^{2i\pi y}} + \frac{e^{-2i\pi y}-e^{-2i\pi ky}}{1-e^{-2i\pi y}} + \frac{e^{-2i\pi ky}}{1-\rho^2 e^{-2i\pi y}}\Big) & y \neq 0 \\ \rho^k\Big(k + \frac{1+\rho^2}{1-\rho^2}\Big) & y = 0. \end{cases}\end{aligned}$$

Therefore,

$$e^{-2ik\pi z}\hat{T}(y-z)_k = \begin{cases} \rho^k\Big(\frac{e^{-2ik\pi z}}{1-\rho^2 e^{2i\pi(y-z)}} + \frac{e^{-2i\pi(y+kz)}-e^{-2i\pi(ky+z)}}{e^{-2i\pi z}-e^{-2i\pi y}} + \frac{e^{-2i\pi ky}}{1-\rho^2 e^{-2i\pi(y-z)}}\Big) & y \neq z \\ \rho^k e^{-2ik\pi z}\Big(k + \frac{1+\rho^2}{1-\rho^2}\Big) & y = z. \end{cases}$$

Therefore, we have that the i,jth entry of $M^{(k)}$ for $k \geq 0$ is given by

$$[M^{(k)}]_{ij} = \begin{cases} \rho^k \left(\frac{e^{-2ik\pi x_j}}{1-\rho^2 e^{2i\pi(x_i - x_j)}} + \frac{e^{-2i\pi(x_i + kx_j)} - e^{-2i\pi(kx_i + x_j)}}{e^{-2i\pi x_j} - e^{-2i\pi x_i}} + \frac{e^{-2i\pi k x_i}}{1-\rho^2 e^{-2i\pi(x_i - x_j)}} \right) & x_i \neq x_j \\ \rho^k e^{-2i\pi x_i} \left(k + \frac{1+\rho^2}{1-\rho^2} \right) & x_i = x_j, \end{cases}$$

and for $k < 0$, we have $M^{(k)} = M^{(-k)^*}$. This allows us to compute $[M^{(k)}]_{ij}$ in the 1D case, which is the above function of k, x_i, and x_j; denote this function $h(k, x_i, x_j)$. For the multidimensional case, since the feature map decomposes over coordinates, we simply have

$$[M^{(k)}]_{ij} = \prod_{a=1}^{d} h(k, x_i[a], x_j[a]).$$

With this in hand, we can implement our algorithm.

Special cases and bounds. Now, we try to control $\|M^{(k)}\|_{Frob.}$, which is needed to compute the a posteriori error guarantees.

First, we note that in the special case $k = 0$, we get:

$$[M^{(0)}]_{ij} = \frac{1 - \rho^4}{1 + \rho^4 - 2\rho^2 \cos 2\pi(x_i - x_j)}.$$

Moreover, we have:

$$\hat{T}(0)_k = \rho^{|k|} \left[|k| + \frac{1+\rho^2}{1-\rho^2} \right].$$

We also have, since φ is always non-negative (see (11)):

$$|\hat{T}(y)_k| = \left| \int_0^1 e^{-2ik\pi x} \varphi(x)\varphi(x-y)dx \right| \leqslant \int_0^1 \varphi(x)\varphi(x-y)dx = \hat{T}(0)_k.$$

Therefore, in 1D

$$|[M^{(k)}]_{ij}| = |e^{-2i\pi k x_j} \hat{T}(x_i - x_j)_k| \leq |\hat{T}(x_i - x_j)_k| \leq \hat{T}(0)_k = \rho^{|k|} \left[|k| + \frac{1+\rho^2}{1-\rho^2} \right]$$

Therefore, we can upper bound in 1D

$$\|M^{(k)}\|_{Frob.} = \sqrt{\sum_{i,j=1}^{n} [M^{(k)}]_{ij}^2} \leq n\rho^{|k|} \left[|k| + \frac{1+\rho^2}{1-\rho^2} \right].$$

In multiple dimensions, we can further bound

$$\begin{aligned}
\|M^{(k)}\|_{Frob.} &\leq n^d \prod_{i=1}^{d} \rho^{|k_i|}\left[|k_i| + \frac{1+\rho^2}{1-\rho^2}\right] \\
&\leq n^d \rho^{|k|}\left[\frac{|k|}{d} + \frac{1+\rho^2}{1-\rho^2}\right]^d \\
&= \left(n\frac{1-\rho^2}{1+\rho^2}\right)^d \rho^{|k|}\left[1 + \frac{|k|\frac{1-\rho^2}{1+\rho^2}}{d}\right]^d \\
&\leq \left(n\frac{1-\rho^2}{1+\rho^2}\right)^d \rho^{|k|} e^{|k|\frac{1-\rho^2}{1+\rho^2}} \\
&= \left(n\frac{1-\rho^2}{1+\rho^2}\right)^d \left(\rho e^{\frac{1-\rho^2}{1+\rho^2}}\right)^{|k|} \\
&= \zeta\tilde{\rho}^{|k|},
\end{aligned}$$

where ζ is a constant independent of k and $\tilde{\rho} < 1$. Therefore, $\|M^{(k)}\|_{Frob.}$ decays exponentially quickly as $|k|$ increases, which ensures that $\sum_{k\in\mathbb{Z}^d}\|M^{(k)}\|_{Frob.}$ is finite and not too large, and that $\sum_{k:|k|>K}\|M^{(k)}\|_{Frob.}$ goes to zero as K increases, which can allow for a tight a posteriori guarantee using Theorem 10.