Non-parametric Models for Non-negative Functions

Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi

INRIA - Département d'Informatique de l'École Normale Supérieure
PSL Research University
Paris, France

July 7, 2020

Abstract

Linear models have shown great effectiveness and flexibility in many fields such as machine learning, signal processing and statistics. They can represent rich spaces of functions while preserving the convexity of the optimization problems where they are used, and are simple to evaluate, differentiate and integrate. However, for modeling non-negative functions, which are crucial for unsupervised learning, density estimation, or non-parametric Bayesian methods, linear models are not applicable directly. Moreover, current state-of-the-art models like generalized linear models either lead to non-convex optimization problems, or cannot be easily integrated. In this paper we provide the first model for non-negative functions which benefits from the same good properties of linear models. In particular, we prove that it admits a representer theorem and provide an efficient dual formulation for convex problems. We study its representation power, showing that the resulting space of functions is strictly richer than that of generalized linear models. Finally we extend the model and the theoretical results to functions with outputs in convex cones. The paper is complemented by an experimental evaluation of the model showing its effectiveness in terms of formulation, algorithmic derivation and practical results on the problems of density estimation, regression with heteroscedastic errors, and multiple quantile regression.

1 Introduction

The richness and flexibility of linear models, with the aid of possibly infinite-dimensional feature maps, allowed to achieve great effectiveness from a theoretical, algorithmic, and practical viewpoint in many supervised and unsupervised learning problems, becoming one of the workhorses of statistical machine learning in the past decades [16, 27]. Indeed linear models preserve convexity of the optimization problems where they are used. Moreover they can be evaluated, differentiated and also integrated very easily.

Linear models are adapted to represent functions with unconstrained real-valued or vector-valued outputs. However, in some applications, it is crucial to learn functions with constrained outputs, such as functions which are non-negative or whose outputs are in a convex set, possibly with additional constraints like an integral equal to one, such as in density estimation, regression of multiple quantiles [9], and isotonic regression [5]. Note that the convex pointwise constraints on the outputs of the learned function must hold everywhere and not only on the training points. In this context, other models have been considered, such as generalized linear models [21], at the expense of losing some important properties that hold for linear ones.

In this paper, we make the following contributions:

- We consider a class of models with non-negative outputs, as well as outputs in a chosen convex cone, which exhibit the same key properties of linear models. They can be used within empirical risk minimization with convex risks, preserving convexity. They are defined in terms of an arbitrary feature map and they can be evaluated, differentiated and integrated exactly.

- We derive a representer theorem for our models and provide a convex finite-dimensional dual formulation of the learning problem, depending only on the training examples. Interestingly, in the proposed formulation, the convex pointwise constraints on the outputs of the learned function are naturally converted to convex constraints on the coefficients of the model.

- We prove that the proposed model is a universal approximator and is strictly richer than commonly used generalized linear models. Moreover, we show that its Rademacher complexity is comparable with the one of linear models based on kernels.

- To show the effectiveness of the method in terms of formulation, algorithmic derivation and practical results, we express naturally the problems of density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression. We derive the corresponding learning algorithms for convex dual formulation, and compare it with standard techniques used for the specific problems on a few reference simulations.

2 Background

In a variety of fields ranging from *supervised learning*, to *Gaussian processes* [35], *inverse problems* [15], *scattered data approximation techniques* [34], and *quadrature methods* to compute multivariate integrals [3], prototypical problems can be cast as

$$f^* \in \arg\min_{f \in \mathcal{F}} \; L(f(x_1), \ldots, f(x_n)) + \Omega(f). \tag{1}$$

Here $L : \mathbb{R}^n \to \mathbb{R}$ is a (often convex) functional, $\mathcal{F}$ a class of real-valued functions, $x_1, \ldots, x_n$ a given set of points in $\mathcal{X}$, and Ω a suitable regularizer [27].

Linear models for the class of functions $\mathcal{F}$ are particularly suitable to solve such problems. They are classically defined in terms of a feature map $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{X}$ is the input space and $\mathcal{H}$ is a separable Hilbert space. Typically, $\mathcal{H} = \mathbb{R}^D$ with $D \in \mathbb{N}$, but $\mathcal{H}$ can also be infinite-dimensional. A linear model is determined by a parameter vector $w \in \mathcal{H}$ as

$$f_w(x) = \phi(x)^\top w, \tag{2}$$

leading to the space $\mathcal{F} = \{f_w \mid w \in \mathcal{H}\}$. These models are particularly effective for problems in the form Eq. (1) because they satisfy the following key properties.

P1. They preserve convexity of the loss function. Indeed, given $x_1, \ldots, x_n \in \mathcal{X}$, if $L : \mathbb{R}^n \to \mathbb{R}$ is convex, then $L(f_w(x_1), \ldots, f_w(x_n))$ is convex in w.

P2. They are universal approximators. Under mild conditions on ϕ and $\mathcal{H}$ (universality of the associated kernel function [22]) linear models can approximate any continuous function on $\mathcal{X}$. Moreover they can represent many classes of functions of interest, such as the class of polynomials, analytic functions, smooth functions on subsets of $\mathbb{R}^d$ or on manifolds, or Sobolev spaces [27].

P3. They admit a finite-dimensional representation. Indeed, there is a so-called *representer theorem* [14]. Let L be a possibly non-convex functional, $\mathcal{F} = \{f_w \mid w \in \mathcal{H}\}$, and assume Ω is an increasing function of $w^\top w$ (see [28] for more generality and details). Then, the optimal solution f^* of (1) corresponds to $f^* = f_{w^*}$, with $w^* = \sum_{i=1}^n \alpha_i \phi(x_i)$, and $\alpha_1, \dots \alpha_n \in \mathbb{R}$. Denoting by k the *kernel function* $k(x, x') := \phi(x)^\top \phi(x')$ for $x, x' \in \mathcal{X}$ (see, e.g., [27]), f^* can be rewritten as

$$f^*(x) = \sum_{i=1}^n \alpha_i k(x, x_i). \tag{3}$$

P4. They are differentiable/integrable in closed form. Assume that the kernel $k(x, x')$ is differentiable in the first variable. Then $\nabla_x f_{w^*}(x) = \sum_{i=1}^n \alpha_i \nabla_x k(x, x_i)$. Also the integral of f_{w^*} can be computed in closed form if we know how to integrate k. Indeed, for $p : \mathcal{X} \to \mathbb{R}$ integrable, we have $\int f_{w^*}(x)p(x)dx = \sum_{i=1}^n \alpha_i \int k(x, x_i)p(x)dx$.

Vector-valued models. By juxtaposing scalar-valued linear models, we obtain a vector valued linear model, i.e. $f_{w_1 \cdots w_p} : \mathcal{X} \to \mathbb{R}^p$ defined as $f_{w_1 \cdots w_p}(x) = (f_{w_1}(x), \dots, f_{w_p}(x)) \in \mathbb{R}^p$.

2.1 Models for non-negative functions or functions with constrained outputs

While linear models provide a powerful formalization for functions from $\mathcal{X}$ to $\mathbb{R}$ or $\mathbb{R}^p$, in some important applications arising in the context of unsupervised learning, non-parametric Bayesian methods, or graphical models, additional conditions on the model are required. In particular, we will focus on *pointwise output constraints*. That is, given $\mathcal{Y} \subsetneq \mathbb{R}^p$, we want to obtain functions satisfying $f(x) \in \mathcal{Y}$ *for all* $x \in \mathcal{X}$. A prototypical example is the problem of density estimation.

Example 1 (density estimation problem)**.** *The goal is to estimate the density of a probability ρ on $\mathcal{X}$, given some i.i.d. samples $x_1, \dots, x_n$. It can be formalized in terms of Eq.* (1) *(e.g., through maximum likelihood), with the constraint that f is a density, i.e., $f(x) \geqslant 0, \ \ \forall x \in \mathcal{X}$, and $\int_{\mathcal{X}} f(x)dx = 1$.*

Despite the similarity with Eq. (1), linear models cannot be applied because of the constraint $f(x) \geqslant 0$. Existing approaches to deal with the problem above are reported below, but lack some of the crucial properties **P1-4** that make linear models so effective for problems of the form Eq. (1).

Generalized linear models (GLM). Given a suitable map $\psi : \mathbb{R}^p \to \mathcal{Y}$, these models are of the form $f(x) = \psi(w^\top \phi(x))$. In the case of non-negative functions, common choices are $\psi(z) = e^z$, leading to the *exponential family*, or the positive part function $\psi(z) = \max(0, z)$. GLM have an expressive power comparable to linear models, being able to represent a wide class of functions, and admit a finite-dimensional representation [13] (they thus satisfy **P2** and **P3**). However, in general they do not preserve convexity of the functionals where they are used (except for specific cases, such as $L = -\sum_{i=1}^n \log z_i$ and $\psi(z) = e^z$ [21]). Moreover they cannot be integrated in closed form, except for specific ϕ, requiring some Monte Carlo approximations [25] (thus missing **P1** and **P4**). An elegant way to obtain a GLM-like non-negative model is via *non-parametric mirror descent* [36] (see, e.g., their Example 4). A favorable feature of this approach is that the map ψ is built implicitly according to the geometry of $\mathcal{Y}$. However, still the resulting model does not always satisfy **P3**, does not satisfy **P1** and **P4** , and is only efficient in small-dimensional input spaces.

Non-negative coefficients models (NCM). Leveraging the finite-dimensional representation of linear models in Eq. (3), the NCM models represent non-negative functions as $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$, with $\alpha_1, \dots \alpha_n \geqslant 0$, given a kernel $k(x, x') \geqslant 0$ for any $x, x' \in \mathcal{X}$, such as the Gaussian kernel $e^{-\|x-x'\|^2}$ or the Abel kernel $e^{-\|x-x'\|}$. By construction these models satisfy

P1, **P3**, **P4**. However, they do not satisfy **P2**. Indeed the fact that $\alpha_1, \dots, \alpha_n \geqslant 0$ does not allow cancellation effects and thus strongly constrains the set of functions that can be represented, as illustrated below.

Example 2. *The NCM model cannot approximate arbitrarily well a function with a width strictly smaller that the width of the kernel. Take $k(x, x') = e^{-\|x-x'\|^2}$ and try to approximate the function $e^{-\|x\|^2/2}$ on $[-1, 1]$. Independently of the chosen n or the chosen locations of the points $(x_i)_{i=1}^n$, it will not be possible to achieve an error smaller than a fixed constant (Appendix D for a simulation).*

Partially non-negative linear models (PNM). A partial solution to have a linear model that is pointwise non-negative is to require non-negativity only on the observed points $(x_i)_{i=1}^n$. That is, the model is of the form $w^\top \phi(x)$, with $w \in \{w \in \mathcal{H} \mid w^\top \phi(x_1) \geqslant 0, \dots, w^\top \phi(x_n) \geqslant 0\}$. While this model is easy to integrate in Eq. (1), this does not guarantee the non-negativity outside of a neighborhood of $(x_i)_{i=1}^n$. It is possible to enrich this construction with a set of points that cover the whole space $\mathcal{X}$ (i.e., a fine grid, if $\mathcal{X} = [-1, 1]^d$), but this usually leads to exponential costs in the dimension of $\mathcal{X}$ and is not feasible when $d \geqslant 4$.

3 Proposed Model for Non-negative Functions

In this section we consider a non-parametric model for non-negative functions and we show that it enjoys the same benefits of linear models. In particular, we prove that it satisfies at the same time all the properties **P1**, …, **P4**. As linear models, the model we consider has a simple formulation in terms of a feature map $\phi : \mathcal{X} \to \mathcal{H}$.

Let $\mathcal{S}(\mathcal{H})$ be the set of bounded Hermitian linear operators from $\mathcal{H}$ to $\mathcal{H}$ (set of symmetric $D \times D$ matrices if $\mathcal{H} = \mathbb{R}^D$ with $D \in \mathbb{N}$) and denote by $A \succeq 0$ the fact that A is a positive semi-definite operator (a positive semi-definite matrix, when $\mathcal{H}$ is finite-dimensional) [24, 18]. The model is defined for all $x \in \mathcal{X}$ as

$$f_A(x) = \phi(x)^\top A \phi(x), \qquad \text{where} \qquad A \in \mathcal{S}(\mathcal{H}), \;\; A \succeq 0. \tag{4}$$

The proposed model [1] is parametrized in terms of the operator (or matrix when $\mathcal{H}$ is finite dimensional) A, like in [8], but with an additional positivity constraint. Note that, by construction, it is linear in A and at the same time non-negative for any $x \in \mathcal{X}$, due to the positiveness of the operator A, as reported below (the complete proof in Appendix B.1).

Proposition 1 (Pointwise positivity and linearity in the parameters)**.** *Given $A, B \in \mathcal{S}(\mathcal{H})$ and $\alpha, \beta \in \mathbb{R}$, then $f_{\alpha A + \beta B}(x) = \alpha f_A(x) + \beta f_B(x)$. Moreover, $A \succeq 0 \;\Rightarrow\; f_A(x) \geqslant 0, \; \forall x \in \mathcal{X}$.*

An important consequence of linearity of f_A in the parameter is that, despite the pointwise non-negativity in x, it preserves **P1**, i.e., the convexity of the functional where it is used. First define the set $\mathcal{S}(\mathcal{H})_+$ as $\mathcal{S}(\mathcal{H})_+ = \{A \in \mathcal{S}(\mathcal{H}) \mid A \succeq 0\}$ and note that $\mathcal{S}(\mathcal{H})_+$ is convex [11].

Proposition 2 (The model satisfies **P1**)**.** *Let $L : \mathbb{R}^n \to \mathbb{R}$ be a jointly convex function and $x_1, \dots, x_n \in \mathcal{X}$. Then the function $A \mapsto L(f_A(x_1), \dots, f_A(x_n))$ is convex on $\mathcal{S}(\mathcal{H})_+$.*

Proposition 2 is proved in Appendix B.2. The property above provides great freedom in choosing the functionals to be optimized with the proposed model. However, when $\mathcal{H}$ has very high dimensionality or it is infinite-dimensional, the resulting optimization problem may be quite expensive. In the next subsection we provide a representer theorem and finite-dimensional representation for our model, that makes the optimization independent from the dimensionality of $\mathcal{H}$.

[1] Note that the model in Eq. (4) has already been considered in [4] with a similar goal as ours. However, this workshop publication has only be lightly peer-reviewed, the representer theorem they propose is incorrect, the optimization algorithm is based on an incorrect representation and inefficient at best. See Appendix F for details.

3.1 Finite-dimensional representations, representer theorem, dual formulation

Here we will provide a finite-dimensional representation for the solutions of the following problem,

$$\inf_{A \succeq 0} L(f_A(x_1), \dots, f_A(x_n)) + \Omega(A), \tag{5}$$

given some points $x_1, \dots, x_n \in \mathcal{H}$. However, the existence and uniqueness of solutions for the problem above depend crucially on the choice of the regularizer Ω as it happens for linear models when $\mathcal{H}$ is finite-dimensional [15]. To derive a representer theorem for our model, we need to specify the class of regularizers we are considering. In the context of linear models a typical regularizer is Tikhonov regularization, i.e., $\Omega(w) = \lambda w^\top w$, for $w \in \mathcal{H}$. Since the proposed model is expressed in terms of a symmetric operator (matrix, if $\mathcal{H}$ is finite-dimensional), the equivalent of the Tikhonov regularizer is a functional that penalizes the squared Frobenius norm of A, i.e., $\Omega(A) = \lambda \operatorname{Tr}(A^\top A)$, for $A \in \mathcal{S}(\mathcal{H})$ also written as $\Omega(A) = \lambda \|A\|_F^2$ [15]. However, since A is an operator, we can also consider different norms on its spectrum. From this viewpoint, an interesting regularizer corresponds to the *nuclear norm* $\|A\|_\star$, which induces sparsity on the spectrum of A, leading to low-rank solutions [23, 8]. In this paper, for the sake of simplicity we will present the results for the following regularizer, which is the matrix/operator equivalent of the *elastic-net* regularizer [38]:

$$\Omega(A) = \lambda_1 \|A\|_\star + \lambda_2 \|A\|_F^2, \quad \forall A \in \mathcal{S}(\mathcal{H}), \tag{6}$$

with $\lambda_1, \lambda_2 \geqslant 0$ and $\lambda_1 + \lambda_2 > 0$. Note that Ω is strongly convex as soon as $\lambda_2 > 0$; we will therefore take $\lambda_2 > 0$ in practice in order to have easier optimization. Recall the definition of the kernel $k(x, x') := \phi(x)^\top \phi(x')$, $x, x' \in \mathcal{X}$ [27]. We have the following theorem.

Theorem 1 (Representer theorem, **P3**). *Let $L : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be lower semi-continuous and bounded below, and Ω as in Eq. (6). Then Eq. (5) has a solution A_* which can be written as*

$$\sum_{i,j=1}^{n} \mathbf{B}_{ij} \phi(x_i)\phi(x_j)^\top, \qquad \text{for some matrix } \mathbf{B} \in \mathbb{R}^{n \times n},\ \mathbf{B} \succeq 0. \tag{7}$$

A_* is unique if L is convex and $\lambda_2 > 0$. By Eq. (4), A_* corresponds to a function of the form

$$f_*(x) = \sum_{i,j=1}^{n} \mathbf{B}_{ij} k(x, x_i) k(x, x_j), \qquad \text{for some matrix } \mathbf{B} \in \mathbb{R}^{n \times n},\ \mathbf{B} \succeq 0.$$

The proof of the theorem above is in Appendix B.3, where it is derived for the more general class of spectral regularizers (this thus extends a result from [1], from linear operators between potentially different spaces to positive self-adjoint operators). A direct consequence of Thm. 1 is the following finite-dimensional representation of the optimization problem in Eq. (5). Denote by $\mathbf{K} \in \mathbb{R}^{n \times n}$ the matrix $\mathbf{K}_{i,j} = k(x_i, x_j)$ and assume w.l.o.g. that it is full rank (always true when the n observations are distinct and k is a *universal kernel* such as the Gaussian kernel [22]). Let $\mathbf{V}$ be the Cholesky decomposition of $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{V}^\top \mathbf{V}$. Define the finite dimensional model

$$\tilde{f}_{\mathbf{A}}(x) = \Phi(x)^\top \mathbf{A} \Phi(x), \qquad \mathbf{A} \in \mathbb{R}^{n \times n},\ \mathbf{A} \succeq 0, \tag{8}$$

where $\Phi : \mathcal{X} \to \mathbb{R}^n$, defined as $\Phi(x) = \mathbf{V}^{-\top} v(x)$, with $v(x) = (k(x, x_i))_{i=1}^n \in \mathbb{R}^n$, is the classical *empirical feature map*. In particular, $\tilde{f}_{\mathbf{A}} = f_A$ where A is of the form Eq. (7) with $\mathbf{B} = \mathbf{V}^{-1} \mathbf{A} \mathbf{V}^{-\top}$. We will say that $\tilde{f}_{\mathbf{A}}$ is a solution of Eq. (5) if the corresponding A is a solution of Eq. (5).

Proposition 3 (Equivalent finite-dimensional formulation in the primal)**.** *Under the assumptions of Thm. 1, the following problem has at least one solution, which is unique if $\lambda_2 > 0$ and L is convex :*

$$\min_{\mathbf{A} \succeq 0} L(\tilde{f}_{\mathbf{A}}(x_1), \dots, \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}). \tag{9}$$

Moreover, for any given solution $\mathbf{A}^* \in \mathbb{R}^{n \times n}$ of Eq. (9), the function $\tilde{f}_{\mathbf{A}^*}$ is a minimizer of Eq. (5). Finally, note that problems Eq. (5) and Eq. (9) have the same condition number if it is exists.

The proposition above (proof in Appendix B.4) characterizes the possibly infinite-dimensional optimization problem of Eq. (5) in terms of an optimization on $n \times n$ matrices. A crucial property is that the formulation in Eq. (9) *preserves convexity*, i.e., it is convex as soon as L is convex. To conclude, Appendix B.4 provides a construction for $\mathbf{V}$ valid for possibly rank-deficient $\mathbf{K}$. We now provide a finer characterization in terms of a dual formulation on only n variables.

Convex dual formulation. We have seen above that the problem in Eq. (5) admits a finite-dimensional representation and can be cast in terms of an equivalent problem on $n \times n$ matrices. Here, when L is convex, we refine the analysis and provide a dual optimization problem on only n variables. The dual formulation is particularly suitable when L is a sum of functions as we will see later. In the following theorem $[\mathbf{A}]_-$ corresponds to the negative part[2] of $\mathbf{A} \in \mathcal{S}(\mathbb{R}^n)$.

Theorem 2 (Convex dual problem)**.** *Assume L is convex, lower semi-continuous and bounded below. Assume Ω is of the form Eq. (6) with $\lambda_2 > 0$. Assume that the problem has at least a strictly feasible point, i.e., there exists $A_0 \succeq 0$ such that L is continuous in $(f_{A_0}(x_1), ..., f_{A_0}(x_n)) \in \mathbb{R}^n$ (this condition is satisfied in simple cases; see examples in Appendix B.5). Denoting with L^* the Fenchel conjugate of L (see [11]), problem Eq. (9) has the following dual formulation:*

$$\sup_{\alpha \in \mathbb{R}^n} -L^*(\alpha) - \tfrac{1}{2\lambda_2} \|[\mathbf{V} \operatorname{Diag}(\alpha) \mathbf{V}^\top + \lambda_1 \mathbf{I}]_-\|_F^2, \tag{10}$$

and this supremum is atteined. Moreover, if $\alpha^* \in \mathbb{R}^n$ is a solution of (10), a solution of (5) is obtained via (7), with $\mathbf{B} \in \mathbb{R}^{n \times n}$ defined as

$$\mathbf{B} = \lambda_2^{-1} \mathbf{V}^{-1} [\mathbf{V} \operatorname{Diag}(\alpha^*) \mathbf{V}^\top + \lambda_1 \mathbf{I}]_- \mathbf{V}^{-\top}. \tag{11}$$

The result above (proof in Appendix B.5) is particularly interesting when L can be written in terms of a sum of functions, i.e., $L(z_1, \dots, z_n) = \sum_{i=1}^n \ell_i(z_i)$ for some functions $\ell_i : \mathbb{R} \to \mathbb{R}$. Then the Fenchel dual is $L^*(\alpha) = \sum_{i=1}^n \ell_i^*(\alpha_i)$, where ℓ_i^* is the Fenchel dual of ℓ_i, and the optimization can be carried by using accelerated proximal splitting methods as FISTA [6], since $\|[\mathbf{V} \operatorname{Diag}(\alpha) \mathbf{V}^\top + \lambda_1 \mathbf{I}]_-\|_F^2$ is differentiable in α. This corresponds to a complexity of $O(n^3)$ per iteration for FISTA, due to the computation of Eq. (11), and can be made comparable with fast algorithms for linear models based on kernels [26], by using techniques from randomized linear algebra and Nyström approximation [17] (see more details in Appendix B.5).

4 Approximation Properties of the Model

The goal of this section is to study the approximation properties of our model and to understand its "richness", i.e., which functions it can represent. In particular, we will prove that, under mild

[2] Given the eigendecomposition $\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$ with $\mathbf{U} \in \mathbb{R}^{n \times n}$ unitary and $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ diagonal, then $[\mathbf{A}]_- = \mathbf{U} \mathbf{\Lambda}_- \mathbf{U}^\top$, with $\mathbf{\Lambda}_-$ diagonal, defined as $(\mathbf{\Lambda}_-)_{i,i} = \min(0, \mathbf{\Lambda}_{i,i})$ for $i = 1, \dots, n$.

assumptions on ϕ, (a) the proposed model satisfies the property **P2**, i.e., it is a *universal approximator* for non-negative functions, and (b) that it is strictly richer than the family of exponential models with the same ϕ. First, define the set of functions belonging to our model

$$\mathcal{F}_\phi^\circ = \{f_A \mid A \in \mathcal{S}(\mathcal{H}),\ A \succeq 0, \|A\|_\circ < \infty\},$$

where $\|\cdot\|_\circ$ is a suitable norm for $\mathcal{S}(\mathcal{H})$. In particular, norms that we have seen to be relevant in the context of optimization are the nuclear norm $\|\cdot\|_\star$ and the Hilbert-Schmidt (Frobenius) norm $\|\cdot\|_F$. Given norms $\|\cdot\|_a, \|\cdot\|_b$, we denote the fact that $\|\cdot\|_a$ is stronger (or equivalent) than $\|\cdot\|_b$ with $\|\cdot\|_a \trianglerighteq \|\cdot\|_b$ (for example, $\|\cdot\|_\star \trianglerighteq \|\cdot\|_F$). In the next theorem we prove that when the feature map is universal [22], such as the one associated to the Gaussian kernel $k(x,x') = \exp(-\|x-x'\|^2)$ or the Abel kernel $k(x,x') = \exp(-\|x-x'\|)$, then the proposed model is a universal approximator for non-negative functions over $\mathcal{X}$ (in particular, in the sense of *cc-universality* [22, 30], see Appendix B.6 for more details and the proof).

Theorem 3 (Universality, **P2**). *Let $\mathcal{H}$ be a separable Hilbert space, $\phi : \mathcal{X} \to \mathcal{H}$ a universal map [22], and $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Then $\mathcal{F}_\phi^\circ$ is a universal approximator of non-negative functions over $\mathcal{X}$.*

The fact that the proposed model can approximate arbitrarily well any non-negative function on $\mathcal{X}$, when ϕ is universal, makes it a suitable candidate in the context of nonparametric approximation/interpolation or learning [34, 33] of non-negative functions. In the following theorem, we give a more precise characterization of the functions contained in $\mathcal{F}_\phi^\circ$. Denote by $\mathcal{G}_\phi$ the set of linear models induced by ϕ, i.e., $\mathcal{G}_\phi = \{w^\top \phi(\cdot) \mid w \in \mathcal{H}\}$ and by $\mathcal{E}_\phi$ the set of *exponential models* induced by ϕ,

$$\mathcal{E}_\phi \ = \ \{\, e^f \mid f(\cdot) = w^\top \phi(\cdot),\ \ w \in \mathcal{H}\,\}.$$

Theorem 4 ($\mathcal{F}_\phi^\circ$ strictly richer than the exponential model). *Let $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Let $\mathcal{X} = [-R,R]^d$, with $R > 0$. Let ϕ such that $W_2^m(\mathcal{X}) = \mathcal{G}_\phi$, for some $m > 0$, where $W_2^m(\mathcal{X})$ is the* Sobolev space *of smoothness m [2]. Let $x_0 \in \mathcal{X}$. The following hold:*

(a) $\mathcal{E}_\phi \subsetneq \mathcal{F}_\phi^\circ$;

(b) the function $f_{x_0}(x) = e^{-\|x-x_0\|^{-2}} \in C^\infty(\mathcal{X})$ satisfies $f_{x_0} \in \mathcal{F}_\phi^\circ$ and $f_{x_0} \notin \mathcal{E}_\phi$.

Thm. 4 shows that if ϕ is rich enough, then the space of exponential models is strictly contained in the space of functions associated to the proposed model. In particular, the proposed model can represent functions that are exactly zero on some subset of $\mathcal{X}$ as showed by the example f_{x_0} in Thm. 4, while the exponential model can represent only strictly positive functions, by construction. Discussion on the condition $W_2^m(\mathcal{X}) = \mathcal{G}_\phi$, proof of Thm. 4 and its generalization to $\mathcal{X} \subseteq \mathbb{R}^d$ are in App. B.7. Here we note only that the condition $W_2^m(\mathcal{X}) = \mathcal{G}_\phi$ is quite mild and satisfied by many kernels such as the Abel kernel $k(x,x') = \exp(-\|x-x'\|)$ [34, 7]. We conclude with a bound on the *Rademacher complexity* [10] of $\mathcal{F}_\phi^\circ$, which is a classical component for deriving generalization bounds [29]. Define $\mathcal{F}_{\phi,L}^\circ = \{f_A \mid A \succeq 0, \|A\|_\circ \leqslant L\}$, for $L > 0$. Thm. 5 shows that the Rademacher complexity of $\mathcal{F}_{\phi,L}^\circ$ depends on L and not on the dimensionality of $\mathcal{X}$, as for regular kernel methods [10].

Theorem 5 (Rademacher complexity of $\mathcal{F}_\phi^\circ$). *Let $\|\cdot\|_\circ \trianglerighteq \|\cdot\|_F$ and $\sup_{x\in\mathcal{X}} \|\phi(x)\| \leqslant c < \infty$. Let $(x_i)_{i=1}^n$ be i.i.d. samples, $L \geqslant 0$. The Rademacher complexity of $\mathcal{F}_{\phi,L}^\circ$ on $(x_i)_{i=1}^n$ is upper bounded by $\frac{2Lc^2}{\sqrt{n}}$ (proof in Appendix B.8).*

5 Extensions: Integral Constraints and Output in Convex Cones

In this section we cover two extensions. The first one generalizes the optimization problem in Eq. (5) to include linear constraints on the integral of the model, in order to deal with problems like density estimation in Example 1. The second formalizes models with outputs in convex cones, which is crucial when dealing with problems like multivariate quantile estimation [9], detailed in Sec. 6.

Constraints on the integral and other linear constraints. We can extend the definition of the problem in Eq. (5) to take into account constraints on the integral of the model. Indeed by linearity of integration and trace, we have the following (proof in Appendix B.9).

Proposition 4 (Integrability in closed form, **P4**). *Let $A \in \mathcal{S}(\mathcal{H})$ with A bounded and ϕ uniformly bounded. Let $p : \mathcal{X} \to \mathbb{R}$ be an integrable function. There exists a trace class operator $W_p \in \mathcal{S}(\mathcal{H})$ such that $\int_{\mathcal{X}} f_A(x)p(x)dx = \operatorname{Tr}(AW_p)$ and $W_p = \int_{\mathcal{X}} \phi(x)\phi(x)^\top p(x)dx$.*

The result can be extended to derivatives and more general linear functionals on f_A (see Appendix B.9). In particular, note that if we consider the *empirical feature map* Φ in Eq. (8), which characterizes the optimal solution of Eq. (5), by Thm. 1, we have that W_p is defined explicitly as $W_p = \mathbf{V}^{-\top}\mathbf{M}_p\mathbf{V}^{-1}$ with $(\mathbf{M}_p)_{i,j} = \int k(x, x_i)k(x, x_j)p(x)dx$, for $i, j = 1, \ldots, n$ and it is computable in closed form. Then, assuming an equality and an inequality constraint on the integral w.r.t. two functions p and q and two values $c_1, c_2 \in \mathbb{R}$, the resulting problem takes the following finite-dimensional form

$$\begin{aligned} \min_{\mathbf{A} \in \mathcal{S}(\mathbb{R}^n)} \quad & L(\tilde{f}_{\mathbf{A}}(x_1), \ldots, \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}), \\ \text{s.t.} \quad & \mathbf{A} \succeq 0, \ \operatorname{Tr}(\mathbf{A}W_p) = c_1, \ \operatorname{Tr}(\mathbf{A}W_q) \leqslant c_2. \end{aligned} \tag{12}$$

Representing function with outputs in convex polyhedral cones. We represent a vector-valued function with our model as the juxtaposition of p scalar valued models, with $p \in \mathbb{N}$, as follows

$$f_{A_1 \cdots A_p}(x) = (f_{A_1}(x), \ldots, f_{A_p}(x)) \in \mathbb{R}^p, \qquad \forall\, x \in \mathcal{X}.$$

We recall that a convex polyhedral cone $\mathcal{Y}$ is defined by a set of inequalities as follows

$$\mathcal{Y} = \{y \in \mathbb{R}^p \mid c^{1\top} y \geqslant 0, \ldots, c^{h\top} y \geqslant 0\}, \tag{13}$$

for some $c^1, \ldots, c^h \in \mathbb{R}^p$ and $h \in \mathbb{N}$. Let us now focus on a single constraint $c^\top y \geqslant 0$. Note that, by definition of positive operator (i.e., $A \succeq 0$ implies $v^\top Av \geqslant 0$ for any A), we have that $\sum_{s=1}^p c_s A_s \succeq 0$ implies $\phi(x)^\top (\sum_{s=1}^p c_s A_s)\phi(x) \geqslant 0$ for any $x \in \mathcal{X}$, which, by linearity of the inner product and the definition of $f_{A_1 \cdots A_p}$ is equivalent to $c^\top f_{A_1 \cdots A_p}(x) \geqslant 0$. From this reasoning we derive the following proposition (see complete proof in Appendix B.10).

Proposition 5. *Let $\mathcal{Y}$ be defined as in Eq. (13). Let $A_1, \ldots, A_p \in \mathcal{S}(\mathcal{H})$. Then the following holds*

$$\sum_{s=1}^p c_s^t A_s \succeq 0 \quad \forall t = 1, \ldots, h \qquad \Rightarrow \qquad f_{A_1 \cdots A_p}(x) \in \mathcal{Y} \quad \forall x \in \mathcal{X}.$$

Note that the set of constraints on the l.h.s. of the equation above defines in turn a convex set on $A_1, \ldots, A_p$. This means that we can use it to constrain a convex optimization problem over the

space of the proposed vector-valued models as follows

$$\min_{A_1,\dots,A_p \in \mathcal{S}(\mathcal{H})} \quad L(f_{A_1 \cdots A_p}(x_1), \dots, f_{A_1 \cdots A_p}(x_n)) \; + \; \sum_{s=1}^{p} \Omega(A_s) \tag{14}$$

$$\text{s.t.} \quad \sum_{s=1}^{p} c_s^t A_s \succeq 0, \quad \forall\, t = 1, \dots, h.$$

By Proposition 5, the function $f_{A_1^* \cdots A_p^*}$, where $(A_1^*, \dots, A_p^*)$ is the minimizer above, will be a function with output in $\mathcal{Y}$. Moreover, the formulation above admits a finite-dimensional representation analogous to the one for non-negative functions, as stated below (see proof in Appendix B.11)

Theorem 6 (Representer theorem for model with output in convex polyhedral cones)**.** *Under the assumptions of Thm. 1, the problem in Eq. (14) admits a minimizer* $(A_1^*, \cdots, A_p^*)$ *of the form*

$$A_s^* = \sum_{i,j=1}^{n} [\mathbf{B}_s]_{i,j} \phi(x_i)\phi(x_j)^\top \implies (f_*(x))_s = \sum_{i,j=1}^{n} [\mathbf{B}_s]_{i,j} k(x_i, x) k(x_j, x), \quad s = 1, ..., p,$$

where $f_* := f_{(A_1^*, \dots, A_p^*)}$ *is the corresponding function and the* $\mathbf{B}_s \in \mathcal{S}(\mathbb{R}^n)$ *are symmetric* $n \times n$ *matrices which satisfy the conic constraints* $\sum_{s=1}^{p} c_s^t \mathbf{B}_s \succeq 0,\ t = 1, ..., h.$

Remark 1 (Efficient representations when the ambient space of $\mathcal{Y}$ is high-dimensional)**.** *When* $p \gg h$*, or when* $\mathcal{Y}$ *is a polyhedral cone with* $\mathcal{Y} \subset \mathcal{G}$ *and* $\mathcal{G}$ *an infinite-dimensional space, it is still possible to have an efficient representation of functions with output in* $\mathcal{Y}$ *by using the representation of* $\mathcal{Y}$ *in terms of* conical hull *[11], i.e.,* $\mathcal{Y} = \{\sum_{i=1}^{t} \alpha_i y_i \mid \alpha_i \geqslant 0\}$ *for some* $y_1, \dots, y_t$ *and* $t \in \mathbb{N}$*. In particular, given* $A_1, \dots, A_t \succeq 0$*, the model* $f_{A_1 \dots A_t}(x) = \sum_{i=1}^{t} f_{A_i}(x) y_i$ *satisfies* $f_{A_1 \dots A_t}(x) \in \mathcal{Y}$ *for any* $x \in \mathcal{X}$*. Moreover it is possible to derive a representer theorem as Thm. 6.*

Remark 2. *By extending this approach, we believe it is possible to model (a) functions with output in the cone of positive semidefinite matrices, (b) convex functions. We leave this for future work.*

6 Numerical Simulations

In this section, we provide illustrative experiments on the problems of density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression. We derive the algorithm according to the finite-dimensional formulation in Eq. (12) for non-negative functions with constraints on the integral, and to Eq. (14) with the finite-dimensional representation suggested by Thm. 6. Optimization is performed applying FISTA [6] on the dual of the resulting formulations. More details on implementation and the specific formulations are given below and in Appendix E. The algorithms are compared with careful implementations of Eq. (1) with the models presented in Sec. 2.1, i.e., partially non-negative models (PNM), non-negative coefficients models (NCM) and generalized linear models (GLM). For all methods we used $\Omega(A) = \lambda \left(\|A\|_* + \frac{0.01}{2} \|A\|_F^2 \right)$ or $\Omega(w) = \lambda \|w\|^2$. We used the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2))$ with width σ. Full cross-validation has been applied to each model independently, to find the best λ (see Appendix E).

Density estimation. This problem is illustrated in Example 1. Here we considered the *log-likelihood loss* as a measure of error, i.e., $L(z_1, \dots, z_n) = -\frac{1}{n} \sum_{i=1}^{n} \log(z_i)$, which is jointly convex and with an efficient proximal operator [12]. We recall that the problems are constrained to output a function whose integral is 1. In Fig. 1, we show the experiment on $n = 50$ i.i.d. points

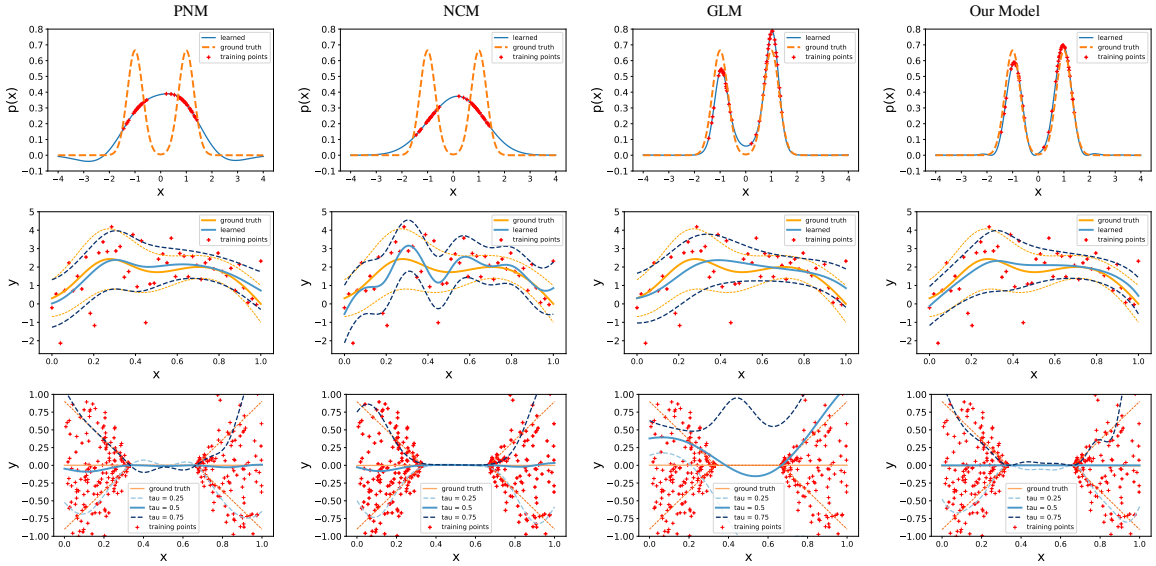


Figure 1: Details in Sec. 6. *(top)* density estimation, *(center)* regression with Gaussian heteroscedastic errors. *(bottom)* multiple quantile regression. Shades of blue: estimated curves. Orange: ground truth. Models: *(left)* PNM, *(center-left)* NCM, *(center-right)*, GLM, *(right)* Our model.

sampled from $\rho(x) = \frac{1}{2}\mathcal{N}(-1, 0.3) + \frac{1}{2}\mathcal{N}(1, 0.3)$ and where for all the models we used $\sigma = 1$, to illustrate pictorially the main interesting behaviors. Instead in Appendix E, we perform a multivariate experiment in $d = 10$ and $n = 1000$, where we cross-validated σ for each algorithm and show the same effects more quantitatively. Note that PNM *(left)* is non-negative on the training points, but it achieves negative values on the regions not covered by examples. This effect is worsened by the constraint on the integral that borrows areas from negative regions to reduce the log-likelihood on the dataset. NCM *(center-left)* produces a function whose integral is one and that is non-negative everywhere, but the poor approximation properties of the model do not allow to fit the density of interest (see Example 2). GLM *(center-right)* produces a function that is non-negative and approximates quite well ρ, however, the obtained function does not sum to one, but to 0.987, since the integral constraint can be enforced only approximately via Monte Carlo sampling (GLM does not satisfy **P4**). Estimating the integral is easy in low dimensions but becomes soon impractical in higher dimensions [25]. Finally the proposed model *(right)* leads to a convex problem and produces a non-negative function whose integral is 1 and that fits the density ρ quite well.

Heteroscedastic Gaussian process estimation. The goal is to estimate $\mu : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}_+$ determining the conditional density ρ of the form $\rho(y|x) = (2\pi v(x))^{-1/2} \exp(-(y - \mu(x))^2/(2v(x)))$ from which the data are sampled. The considered functional corresponds to the negative log-likelihood, i.e., $L = \sum_{i=1}^{n} \frac{1}{2} \log v(x_i) + (y_i - \mu(x_i))^2/(2v(x_i))$ that becomes convex in η, θ via the so called *natural parametrization* $\eta(x) = \mu(x)/v(x)$ and $\theta(x) = 1/v(x)$ [20]. We used a linear model to parametrize η and the non-negative models for θ. The experiment on the same model of [20, 37] is reported in Fig. 1. Modeling θ via PNM *(left)* leads to a convex problem and reasonable performance. In particular, the fact that $\theta = 0$ corresponds to $v = +\infty$ prevents the model for θ from crossing zero. NCM *(center-left)* leads to a convex problem, but very sensitive to the kernel width σ and with poor approximation properties. GLM *(center-right)* leads to a non-convex problem and we need to restart the method randomly to have a reasonable convergence. Our model *(right)* leads to a convex problem and produces a non-negative function for θ, that fits well the observed data.

Multiple quantile regression. The goal here is to estimate multiple quantiles of a given conditional distribution $P(Y|x)$. Given $\tau \in (0,1)$, q_τ defined by $P(Y > q_\tau(x)|x) = \tau$ is the τ-quantile of ρ. By construction $0 < \tau_{-h} \leqslant \ldots \leqslant \tau_h < 1$ implies $q_{\tau_{-h}}(x) \leqslant \ldots \leqslant q_{\tau_h}(x)$. If we denote by $\mathbf{q} : \mathcal{X} \to \mathbb{R}^{2h+1}$ the list of quantiles, we have by construction $\mathbf{q}(x) \in \mathcal{Y}$ where $\mathcal{Y}$ is a convex cone $\mathcal{Y} = \{y \in \mathbb{R}^h \mid y_{-h} \leqslant \ldots \leqslant y_h\}$. To regress quantiles, we used the pinball loss L_τ (convex, non-smooth) considered in [19, 31], obtaining $L = \sum_{j=-h}^{h} \sum_{i=1}^{n} L_{\tau_j}(f(x_i), y_i)$. In Fig. 1, we used $\tau_{-1} = \frac{1}{4}, \tau_0 = \frac{1}{2}, \tau_1 = \frac{3}{4}$. Using PNM, *(left)* the ordering is enforced by explicit constraints on the observed dataset [32, 9]. The resulting problem is convex. However, in regions with low density of points, PNM quantiles do not respect their natural order. To enforce the order constraint, a fine grid covering the space would be needed as in [32]. For NCM, GLM and our model, we represented the quantiles as $q_{\tau_{\pm j}} = q_{\tau_0} \pm \sum_{i=1} v_{\pm i}$ where the v's are non-negative functions and q_{τ_0}, with $\tau_0 = \frac{1}{2}$, is the median and is modeled by a linear model. NCM *(center-left)* leads to a convex problem and quantiles that respect the ordering, but the estimation is very sensitive to the chosen σ and has poor approximation properties. GLM *(center-right)* leads to a non-convex non-differentiable problem, with many local minima, which is difficult to optimize with standard techniques (see Appendix E). GLM does not succeed in approximating the quantiles. Our model *(right)* leads to a convex optimization problem that approximates the quantiles relatively well and preserves their natural order everywhere.

Acknowledgments

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

References

[1] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(Mar):803–826, 2009.

[2] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Elsevier, 2003.

[3] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(1):714–751, 2017.

[4] J. Andrew Bagnell and Amir-massoud Farahmand. Learning positive functions in a Hilbert space. In *NIPS Workshop on Optimization (OPT2015)*, pages 3240–3255, 2015.

[5] Richard E. Barlow and Hugh D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

[6] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[7] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

[8] Mathieu Blondel, Akinori Fujino, and Naonori Ueda. Convex factorization machines. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 19–35. Springer, 2015.

[9] Howard D. Bondell, Brian J. Reich, and Huixia Wang. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

[10] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[11] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[12] Caroline Chaux, Patrick L. Combettes, Jean-Christophe Pesquet, and Valérie R. Wajs. A variational formulation for frame-based inverse problems. *Inverse Problems*, 23(4):1495–1518, 2007.

[13] Elliott Ward Cheney and William Allan Light. *A Course in Approximation Theory*, volume 101. American Mathematical Soc., 2009.

[14] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

[15] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media, 1996.

[16] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer, 2001.

[17] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[18] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[19] Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

[20] Quoc V. Le, Alex J. Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496, 2005.

[21] P. McCullagh and J. A. Nelder. Generalized linear models, 1989.

[22] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[23] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[24] Michael Reed. *Methods of Modern Mathematical Physics: Functional Analysis*. Elsevier, 1980.

[25] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[26] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898, 2017.

[27] Bernhard Schölkopf and Alexander Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[28] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[29] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[30] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (Jul):2389–2410, 2011.

[31] Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

[32] Ichiro Takeuchi, Quoc Le, Timothy Sears, and Alexander Smola. Nonparametric quantile regression. *Journal of Machine Learning Research*, 01 2005.

[33] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

[34] Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

[35] Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*, volume 2. MIT Press, 2006.

[36] Yingxiang Yang, Haoxiang Wang, Negar Kiyavash, and Niao He. Learning positive functions with pseudo mirror descent. In *Advances in Neural Information Processing Systems*, pages 14144–14154, 2019.

[37] Ming Yuan and Grace Wahba. Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & Probability Letters*, 69(1):11–20, 2004.

[38] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B (Statistical Methodology)*, 67(2):301–320, 2005.
biblio.bib

A Notation and basic definitions

- $\mathcal{H}$ is a separable Hilbert space.
- $\mathcal{X}$ is a Polish space (we will require explicitly compactness in some theorems).
- $\phi : \mathcal{X} \to \mathcal{H}$ is a continous map. We also assume it to be uniformly bounded i.e.

$$\sup_{x \in \mathcal{X}} \|\phi(x)\| \leqslant c$$

  for some $c \in (0, \infty)$, if not differently stated.
- $k(x, x') := \phi(x)^\top \phi(x')$ is the *kernel* function associated to the feature map ϕ [SS02, BTA11].

B Proofs and additional discussions

B.1 Proof of Proposition 1

In this section, let us extend the definition in Eq. (4) to any operator $A \in \mathcal{S}(\mathcal{H})$, without the implied positivity restriction (in Eq. (4), we ask that $A \succeq 0$) :

$$\forall A \in \mathcal{S}(\mathcal{H}),\ \forall x \in \mathcal{X},\ f_A(x) := \phi(x)^\top A \phi(x). \tag{4bis}$$

Proof of Proposition 1.
To prove linearity, let $A, B \in \mathcal{S}(\mathcal{H})$ and $\alpha, \beta \in \mathbb{R}$. Since $\mathcal{S}(\mathcal{H})$ is a vector space,

$$\alpha A + \beta B \in \mathcal{S}(\mathcal{H}).$$

Let $x \in \mathcal{X}$. By definition for the first equality and linearity for the second,

$$f_{\alpha A + \beta B}(x) = \phi(x)^\top (\alpha A + \beta B)\phi(x) = \alpha \phi(x)^\top A \phi(x) + \beta \phi(x)^\top B \phi(x).$$

Finally, since by definition, $f_A(x) = \phi(x)^\top A \phi(x)$ and $f_B(x) = \phi(x)^\top B \phi(x)$, it holds :

$$f_{\alpha A + \beta B}(x) = \alpha \phi(x)^\top A \phi(x) + \beta \phi(x)^\top B \phi(x) = \alpha f_A(x) + \beta f_B(x).$$

Since this holds for all $x \in \mathcal{X}$, this shows $f_{\alpha A + \beta B} = \alpha f_A + \beta f_B$.

To prove the non-negativity, assume now that $A \succeq 0$. By definition of of positive semi-definiteness,

$$\forall h \in \mathcal{H},\ h^\top A h \geqslant 0.$$

In particular, for any $x \in \mathcal{X}$, the previous inequality applied to $h = \phi(x)$ yields

$$f_A(x) = \phi(x)^\top A \phi(x) \geqslant 0.$$

Hence, $f_A \geqslant 0$.

B.2 Proof of Proposition 2

Recall the definition of f_A for any $A \in \mathcal{S}(\mathcal{H})$ in Eq. (4bis). We have the lemma:

Lemma 1 (Linearity of evaluations)**.** *Let $x_1, \dots, x_n \in \mathcal{X}$. Then the map*

$$A \in \mathcal{S}(\mathcal{H}) \mapsto (f_A(x_i))_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$$

is linear from $\mathcal{S}(\mathcal{H})$ to $\mathbb{R}^n$.

Proof. This just follows from the fact that the definition of $f_A(x_i)$, $f_A(x_i) := \phi(x_i)^\top A \phi(x_i)$, is linear in A.

Proof of Proposition 2. Let $L : \mathbb{R}^n \to \mathbb{R}$ be a jointly convex function and $x_1, \dots, x_n \in \mathcal{X}$. The function $A \in \mathcal{S}(\mathcal{H}) \mapsto L(f_A(x_1), \dots, f_A(x_n))$ can be written $L \circ R$, where

$$R : A \in \mathcal{S}(\mathcal{H}) \mapsto (f_A(x_i))_{1 \leqslant i \leqslant n} \in \mathbb{R}^n.$$

Since L is convex, and R is linear by Lemma 1, their composition is convex.

Moreover, since $\mathcal{S}(\mathcal{H})_+$ is a convex subset of $\mathcal{S}(\mathcal{H})$, the restriction of $A \in \mathcal{S}(\mathcal{H}) \mapsto L(f_A(x_1), \dots, f_A(x_n))$ on $\mathcal{S}(\mathcal{H})_+$ is also convex.

B.3 Proof of Thm. 1

In this section, we prove Thm. 1 for a more general class of spectral regularizers.

B.3.1 Compact operators and spectral functions

In this section, we briefly introduce compact self-adjoint operators and the spectral theory of compact self-adjoint operators. For more details, see for instance [GGK04]. We start by defining a compact self-adjoint operator (see Section2.16 of [GGK04]) and stating its main properties:

Definition 1 (compact operators)**.** *Let $\mathcal{H}$ be a separable Hilbert space. A bounded self-adjoint operator $A \in \mathcal{S}(\mathcal{H})$ is said to be compact if its range is included in a compact set. We denote with $\mathcal{S}_\infty(\mathcal{H})$ the set of compact self adjoint operators on $\mathcal{H}$. It is a closed subspace of $\mathcal{S}(\mathcal{H})$ for the operator norm and the closure of the set of finite rank operators.*

Proposition 6 (Spectral theorem [GGK04])**.** *Let $\mathcal{H}$ be a separable Hilbert space and let A be a compact self adjoint operator on $\mathcal{H}$. Then there exists a spectral decomposition of A, i.e., an orthonormal system $(u_k) \in \mathcal{H}$ of eigenvectors of A and corresponding eigenvalues (σ_k) such that for all $h \in \mathcal{H}$, it holds*

$$Ah = \sum_k \sigma_k u_k^\top h \; u_k =: \left(\sum_k \sigma_k u_k u_k^\top \right) h.$$

Moreover, if σ_k is an infinite sequence, it converges to zero.
Furthermore, we say that the orthonormal system (u_k) of eigenvectors of A and the corresponding eigenvalues (σ_k) is a basic system *of eigenvectors of A if all the σ_k are non zero. In this case, if P_0 denotes the orthogonal projection on $\mathrm{Ker}(A)$, then it holds*

$$\forall h \in \mathcal{H}, \; h = \Pi_0 \, h + \sum_k u_k u_k^\top \, h$$

In what follows, to simplify notations, we will usually write $A = U\,\mathrm{Diag}(\sigma)U^\top$ in order to denote a basic system of eigenvectors of A. Moreover, if A is positive semi-definite, we will assume that the eigenvalues are sorted in decreasing order, i.e., $\sigma_{k+1} \leqslant \sigma_k$.

Definition 2 (Spectral function on $\mathcal{S}_\infty(\mathcal{H})$ [GGK04])**.** *Let $q : \mathbb{R} \to \mathbb{R}$ be a lower semi-continuous function such that $q(0) = 0$. Let $\mathcal{H}$ be any separable Hilbert space. For any $A \in \mathcal{S}_\infty(\mathcal{H})$ and any basic system $A = U\,\mathrm{Diag}(\sigma)U^\top$, we define the spectral function q*

$$q(A) = U\,\mathrm{Diag}(q(\sigma)))U^T = \sum_k q(\sigma_k) u_k u_k^\top.$$

B.3.2 Classes of regularizers

Let us now state our main assumption on regularizers.

Assumption 1 (Assumption on regularizers)**.** *For any $A \in \mathcal{S}(\mathcal{H})$, $\Omega(A)$ is of of the form*

$$\Omega(A) = \begin{cases} \mathrm{Tr}(q(A)) = \sum_k q(\sigma_k) & \text{if } A = U\,\mathrm{Diag}(\sigma)U^\top \in \mathcal{S}_\infty(\mathcal{H}),\ \sum_k q(\sigma_k) < \infty \\ +\infty & \text{otherwise,} \end{cases}$$

where $q : \mathbb{R} \to \mathbb{R}_+$ is:

- *non-decreasing on $\mathbb{R}_+$ with $q(0) = 0$;*
- *lower semi-continuous;*
- $q(\sigma) \underset{|\sigma| \to +\infty}{\longrightarrow} +\infty$.

Note that in this case, Ω is defined on $\mathcal{S}(\mathcal{H})$ for any Hilbert space $\mathcal{H}$.

Remark 3. *$\Omega(A) = \lambda_1 \|A\|_\star + \frac{\lambda_2}{2}\|A\|_F^2$ satisfies Assumption 1, with $q(\sigma) = \lambda_1\,|\sigma| + \lambda_2\,\sigma^2$.*

Lemma 2 (Properties of Ω)**.** *Let Ω satisfying Assumption 1. Then the following properties hold.*

(i) For any separable Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and any linear isometry $O : \mathcal{H}_1 \to \mathcal{H}_2$, i.e., such that $O^*O = I_{\mathcal{H}_1}$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H}_1),\ \Omega(OAO^*) = \Omega(A).$$

(ii) For any separable Hilbert space $\mathcal{H}$ and any orthogonal projection $\Pi \in \mathcal{S}(\mathcal{H}_1)$, i.e., satisfying $\Pi = \Pi^*,\ \Pi^2 = \Pi$, it holds

$$\forall A \succeq 0,\ \Omega(\Pi A \Pi) \leqslant \Omega(A).$$

(iii) For any finite dimensional Hilbert space $\mathcal{H}_n$,

$$\Omega \text{ is lower semi-continuous (l.s.c)}, \qquad \Omega(A) \underset{\|A\|_{op} \to +\infty}{\longrightarrow} +\infty$$

where we denoted by $\|\cdot\|_{op}$ the operator norm.

Proof.

(i) Write $A = \sum_k \sigma_k u_k u_k^\top$ where the (u_k) form a basic system of eigen-vectors for A. The $(v_k) = (Ou_k)$ form a basic system of eigen-vectors for OAO^*, as

$$OAO^* = \sum_k \sigma_k v_k v_k^\top, \qquad \sigma_k \neq 0.$$

Hence, by definition, $q(OAO^*) = \sum_k q(\sigma_k) v_k v_k^\top$. By definition of the trace, we have

$$\Omega(OAO^*) = \sum_k q(\sigma_k) = \Omega(A).$$

(ii) Let A be a compact self-adjoint semi-definite operator. Let $A = U \operatorname{Diag}(\sigma) U^\top$ be a basic system of eigenvectors of A, where the σ_k are positive and in decreasing order. Define $B = U \operatorname{Diag}(\sqrt{\sigma}) U^\top$ and note that in this case, $A = B^2 = B^*B$. Using Exercise 23 of [GGK04], we have that for any orthogonal projection operator Π and any index k, $\sigma_k(\Pi B^* B \Pi) \leqslant \sigma_k(B^*B)$ and hence $\sigma_k(\Pi A \Pi) \leqslant \sigma_k(A)$. Since q is non decreasing, it holds $q(\sigma_k(\Pi A \Pi)) \leqslant q(\sigma_k(A))$ and hence

$$\Omega(\Pi A \Pi) = \sum_k q(\sigma_k(\Pi A \Pi)) \leqslant \sum_k q(\sigma_k(A)) = \Omega(A).$$

(iii) Let $\mathcal{H}_n$ be a finite dimensional Hilbert space and let $\|\cdot\|_{op}$ be the operator norm on $\mathcal{S}(\mathcal{H}_n)$. If q is continuous, then $A \in \mathcal{H}_n \mapsto q(A)$ is continuous and hence Ω is continuous (since the trace is continuous in finite dimensions). Now assume q is lower semi-continuous, and define for $n \in \mathbb{N}$, $q_n(t) := \inf_{s \in \mathbb{R}} q(s) + n|t - s|$. We have $q_n \geqslant 0$, $q_n(0) = 0$ q_n is uniformly continuous and q_n is an increasing sequence of functions such that $q_n \to q$ point-wise. Now it is easy to see that $\operatorname{Tr}(q(A)) = \sup_n \operatorname{Tr}(q_n(A))$ and hence Ω is lower semi-continuous as a supremum of continuous functions.
The fact that Ω goes to infinity is a direct consequence of the fact that q goes to infinity, by Assumption 1.

Remark 4. *The three conditions of the previous lemma are in fact the only conditions needed in the proof. We could loosen Assumption 1 to satisfy only these three properties.*

B.3.3 Finite-dimensional representation and existence of a solution

Fix $n \in \mathbb{N}$, a loss function $L : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, a separable Hilbert space $\mathcal{H}$, a regularizer Ω on $\mathcal{S}(\mathcal{H})$ a feature map $\phi : \mathcal{X} \to \mathcal{H}$ and points $(x_1, ..., x_n) \in \mathcal{X}^n$.

Recall the problem in Eq. (5):

$$\inf_{A \succeq 0} L(f_A(x_1), \ldots, f_A(x_n)) + \Omega(A). \tag{5}$$

Define $\mathcal{H}_n$ to be the finite-dimensional subset of $\mathcal{H}$ spanned by the $\phi(x_i)$, i.e.,

$$\mathcal{H}_n := \operatorname{span}(\phi(x_i))_{1 \leqslant i \leqslant n} = \left\{ \sum_{i=1}^n \alpha_i \phi(x_i) \ : \ \alpha \in \mathbb{R}^n \right\}.$$

Define Π_n is the orthogonal projection on $\mathcal{H}_n$, i.e.,

$$\Pi_n \in \mathcal{S}(\mathcal{H}), \ \Pi_n^2 = \Pi_n, \ \operatorname{range}(\Pi_n) = \mathcal{H}_n.$$

Define $\mathcal{S}_n(\mathcal{H})_+$ to be the following subspace of $\mathcal{S}(\mathcal{H})_+$:

$$\mathcal{S}_n(\mathcal{H})_+ := \Pi_n \mathcal{S}(\mathcal{H})_+ \Pi_n = \{\Pi_n A \Pi_n \ : \ A \in \mathcal{S}(\mathcal{H})_+\}.$$

Proposition 7. *Let L be a lower semi-continuous function which is bounded below, and assume Ω satisfies Assumption 1. Then Eq. (5) has a solution A^* which is in $\mathcal{S}_n(\mathcal{H})_+$.*

Proof. In this proof, denote by J the function defined by

$$\forall A \in \mathcal{S}(\mathcal{H}),\ J(A) := L(f_A(x_1), ..., f_A(x_n)) + \Omega(A).$$

Our goal is to prove that the problem $\inf_{A\in\mathcal{S}(\mathcal{H})_+} J(A)$ has a solution which is in $\mathcal{S}_n(\mathcal{H})_+$, i.e., of the form $\Pi_n A \Pi_n$ for some $A \in \mathcal{S}(\mathcal{H})_+$.

1. Let us start by fixing $A \in \mathcal{S}(\mathcal{H})_+$.
First note that since Π_n is the orthogonal projection on $\mathrm{span}(\phi(x_i))_{1\leqslant i\leqslant n}$, in particular $\Pi_n\phi(x_i) = \phi(x_i)$ for all $1 \leqslant i \leqslant n$. Thus, for any $1 \leqslant i \leqslant n$,

$$f_A(x_i) = \phi(x_i)^\top A\phi(x_i) = \phi(x_i)^\top \Pi_n A\Pi_n\phi(x_i) = f_{\Pi_n A\Pi_n}(x_i).$$

Here, the first and last equalities come from the definition of f_A and $f_{\Pi_n A\Pi_n}$. Thus,

$$J(A) = L(f_{\Pi_n A\Pi_n}(x_1), ..., f_{\Pi_n A\Pi_n}(x_n)) + \Omega(A).$$

Now since Ω satisfies Assumption 1, by the second point of Lemma 2, it holds $\Omega(\Pi_n A\Pi_n) \leqslant \Omega(A)$, hence

$$J(\Pi_n A\Pi_n) \leqslant J(A).$$

This last inequality combined with the fact that $\mathcal{S}_n(\mathcal{H})_+ = \Pi_n\mathcal{S}(\mathcal{H})_+\Pi_n \subset \mathcal{S}(\mathcal{H})_+$ show that

$$\inf_{A\in\mathcal{S}_n(\mathcal{H})_+} J(A) = \inf_{A\succeq 0} J(A). \tag{15}$$

2. Let us now show that $\inf_{A\in\mathcal{S}_n(\mathcal{H})_+} J(A)$ has a solution. Let us exclude the case where $J = +\infty$, in which case $A = 0$ can be taken to be a solution.

Let V_n be the injection $V_n : \mathcal{H}_n \hookrightarrow \mathcal{H}$. Note that $V_nV_n^* = \Pi_n$ and $V_n^*V_n = I_{\mathcal{H}_n}$. These simple facts easily show that

$$\mathcal{S}_n(\mathcal{H})_+ = V_n\mathcal{S}(\mathcal{H}_n)_+V_n^* = \left\{V_n\tilde{A}V_n^* \ :\ \tilde{A} \in \mathcal{S}(\mathcal{H}_n)_+\right\}.$$

Thus, our goal is to show that $\inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+} J(V_nAV_n^*)$ has a solution.

By the first point of Lemma 2, since $V_n^*V_n = I_{\mathcal{H}_n}$, it holds

$$\forall \tilde{A} \in \mathcal{S}(\mathcal{H}_n),\ \Omega(V_n\tilde{A}V_n^*) = \Omega(\tilde{A}) \implies J(V_n\tilde{A}V_n^*) = L(f_{V_n\tilde{A}V_n^*}(x_1), ..., f_{V_n\tilde{A}V_n^*}(x_n)) + \Omega(\tilde{A}).$$

Let $\tilde{A}_0 \in \mathcal{S}(\mathcal{H}_n)_+$ be a point such that $J_0 := J(V_n\tilde{A}_0V_n^*) < \infty$. Let c_0 be a lower bound for L. By the third point of Lemma 2, there exists a radius R_0 such that for all $\tilde{A} \in \mathcal{S}(\mathcal{H}_n)$,

$$\|\tilde{A}\|_F > R_0 \implies \Omega(\tilde{A}) > J_0 - c_0.$$

Since c_0 is a lower bound for L, this implies

$$\inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+} J(V_n\tilde{A}V_n^*) = \inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+,\ \|\tilde{A}\|_F\leqslant R_0} J(V_n\tilde{A}V_n^*).$$

Now since L is lower semi-continuous, Ω is lower semi-continuous by the last point of Lemma 2, and $\tilde{A} \mapsto (f_{V_n\tilde{A}V_n^*}(x_i))_{1\leqslant i\leqslant n}$ is linear hence continuous, the mapping $A \mapsto J(V_n\tilde{A}V_n^*)$ is lower semi-continuous. Hence, it reaches its minimum on any non empty compact set. Since $\mathcal{H}_n$ is finite dimensional, the set $\left\{\tilde{A} \in \mathcal{S}(\mathcal{H}_n)_+ \ :\ \|\tilde{A}\|_F \leqslant R_0\right\}$ is compact (closed and bounded) and non empty since it contains $\tilde{A}_0$, and hence there exists $\tilde{A}_* \in \mathcal{S}(\mathcal{H}_n)_+$ such that $J(V_n\tilde{A}_*V_n^*) = \inf_{\tilde{A}\in\mathcal{S}(\mathcal{H}_n)_+,\ \|\tilde{A}\|_F\leqslant R_0} J(V_n\tilde{A}V_n^*)$. Going back up the previous equalities, this shows that $A_* = V_n\tilde{A}_*V_n^* \in \mathcal{S}_n(\mathcal{H})_+$ and $J(A_*) = \inf_{A\succeq 0} J(A)$.

B.3.4 Proof of Thm. 1

We will prove the following Thm. 7 whose statement is that of Thm. 1 with more general assumptions.

Theorem 7. *Let L be lower semi-continuous and bounded below, and Ω satisfying Assumption 1. Then Eq. (5) has a solution A_* which can be written in the form*

$$\sum_{i,j=1}^{n} \mathbf{B}_{ij}\phi(x_i)\phi(x_j)^\top, \qquad \text{for some matrix } \mathbf{B} \in \mathbb{R}^{n\times n},\ \mathbf{B} \succeq 0.$$

Moreover, if L is convex, and Ω is of the form Eq. (6) with $\lambda_2 > 0$, this solution is unique. By Eq. (4), A_* corresponds to a function of the form

$$f_*(x) = \sum_{i,j=1}^{n} \mathbf{B}_{ij}k(x, x_i)k(x, x_j).$$

Lemma 3. *The set $\mathcal{S}_n(\mathcal{H})_+$ can be represented in the following way*

$$\mathcal{S}_n(\mathcal{H})_+ = \left\{ \sum_{1\leqslant i,j\leqslant n} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top,\ :\ \mathbf{B} \in \mathbb{R}^{n\times n},\ \mathbf{B} \succeq 0 \right\}.$$

In particular, for any $A \in \mathcal{S}_n(\mathcal{H})_+$, there exists a matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \succeq 0$ such that

$$A = \sum_{1\leqslant i,j\leqslant n} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top \implies \forall x \in \mathcal{X},\ f_A(x) = \sum_{1\leqslant i,j\leqslant n} \mathbf{B}_{i,j}k(x_i, x)k(x_j, x).$$

Proof. Define $S_n : \mathcal{H} \to \mathbb{R}^n$ to be the operator such that

$$\forall h,\ S_n(h) = \left(h^\top\phi(x_i)\right)_{1\leqslant i\leqslant n},$$

with adjoint $S_n^* : \mathbb{R}^n \to \mathcal{H}$ such that

$$\forall \alpha \in \mathbb{R}^n,\ S_n^*\alpha = \sum_{i=1}^{n} \alpha_i\phi(x_i).$$

Note that for any $\mathbf{B} \in \mathbb{R}^{n\times n}$, $S_n^*\mathbf{B}S_n = \sum_{i,j} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top$.

1. Proving $\mathcal{S}_n(\mathcal{H})_+ \subset \left\{\sum_{1\leqslant i,j\leqslant n} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top,\ :\ \mathbf{B} \in \mathbb{R}^{n\times n},\ \mathbf{B} \succeq 0\right\}$. Let $\Pi_n A\Pi_n$ be in $\mathcal{S}_n(\mathcal{H})_+$. Using the previous equality, we want to show there exists $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \succeq 0$ such that $\Pi_n A\Pi_n = S_n^*\mathbf{B}S_n$. Using Lemma 4, we see that Π_n can be written in the form $S_n^*T_n$ where $T_n : \mathcal{H} \to \mathbb{R}^n$ (write $\Pi_n = O_nO_n^*$ and note that O_n is of the form $S_n^*\tilde{O}_n$). Hence, defining $\mathbf{B}$ to be the matrix associated to the operator $T_nAT_n^* : \mathbb{R}^n \to \mathbb{R}^n$, it holds $\Pi_n A\Pi_n = S_n^*\mathbf{B}S_n$. Moreover, $A \succeq 0$ implies $\mathbf{B} = T_nAT_n^* \succeq 0$.

2. Proving $\left\{\sum_{1\leqslant i,j\leqslant n} \mathbf{B}_{i,j}\phi(x_i)\phi(x_j)^\top,\ :\ \mathbf{B} \in \mathbb{R}^{n\times n},\ \mathbf{B} \succeq 0\right\} \subset \mathcal{S}_n(\mathcal{H})_+$. Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ and assume $\mathbf{B} \succeq 0$. Since $\mathbf{B} \succeq 0$, $A := S_n^*\mathbf{B}S_n \succeq 0$. Since S_n^* has its range included in $\mathcal{H}_n$, $\Pi_n S_n^* = S_n^*$. Thus, $\Pi_n A\Pi_n = A$ and hence $A \in \mathcal{S}_n(\mathcal{H})_+$.

The second statement comes from the definition of $f_A(x)$. Indeed assume $A \in \mathcal{S}_n(\mathcal{H})_+$. By definition, $f_A(x) = \phi(x)^\top A \phi(x)$. Moreover, by the previous point, there exists $\mathbf{B} \in \mathbb{R}^{n \times n},\ \mathbf{B} \succeq 0$ such that $A = \sum_{1 \leqslant i,j \leqslant n} \mathbf{B}_{i,j} \phi(x_i)\phi(x_j)^\top$. Combining these two facts yields:

$$\forall x \in \mathcal{X},\ f_A(x) = \sum_{1 \leqslant i,j \leqslant n} \mathbf{B}_{i,j} \phi(x)^\top \phi(x_i)\ \phi(x_j)^\top \phi(x) = \sum_{1 \leqslant i,j \leqslant n} \mathbf{B}_{i,j} k(x, x_i)\ k(x, x_j).$$

The last equality comes from the definition $k(x, \tilde{x}) = \phi(x)^\top \phi(\tilde{x})$.

Proof of Thm. 7. Under the assumptions of Thm. 7, one satisfies the assumptions of Proposition 7. Thus, Eq. (5) has a solution A_* which is in $\mathcal{S}_n(\mathcal{H})_+$. Now applying Lemma 3, A_* can be written in the form $A_* = \sum_{i,j} \mathbf{B}_{i,j} \phi(x_i)\phi(x_j)^\top$ for $\mathbf{B} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \succeq 0$, and hence

$$\forall x \in \mathcal{X},\ f_{A_*}(x) = \sum_{i,j} \mathbf{B}_{i,j} k(x, x_i) k(x, x_j).$$

Uniqueness in the case where Ω is of the form Eq. (6) with $\lambda_2 > 0$ comes from the fact that the loss function is strongly convex in this case, and thus the minimizer is unique.

B.4 Proof of Proposition 3

Recall the definitions of $S_n : \mathcal{H} \to \mathbb{R}^n$ and its adjoint $S_n^* : \mathbb{R}^n \to \mathcal{H}$:

$$\forall h,\ S_n(h) = \left(h^\top \phi(x_i)\right)_{1 \leqslant i \leqslant n},\ \forall \alpha \in \mathbb{R}^n,\ S_n^* \alpha = \sum_{i=1}^n \alpha_i \phi(x_i).$$

Note that the kernel matrix $\mathbf{K} = (k(x_i, x_j))_{1 \leqslant i,j \leqslant n}$ can also be written as $\mathbf{K} = S_n S_n^*$.
Let r be the rank of $\mathbf{K}$ and $\mathbf{V} \in \mathbb{R}^{r \times n}$ be a matrix such that

$$\mathbf{V}^\top \mathbf{V} = \mathbf{K}.$$

Note that $\mathbf{V}$ is of rank r and hence $\mathbf{V}\mathbf{V}^\top$ is invertible, making the following definition of O_n : $\mathbb{R}^r \to \mathcal{H}$ valid:

$$O_n = S_n^* \mathbf{V}^\top (\mathbf{V}\mathbf{V}^\top)^{-1}.$$

The following result holds :

Lemma 4. *$O_n O_n^* = \Pi_n$ and $O_n^* O_n = I_r$.*

Proof. Using the fact that $\mathbf{V}^\top \mathbf{V} = \mathbf{K} = S_n S_n^*$, we have

$$O_n^* O_n = (\mathbf{V}\mathbf{V}^\top)^{-1} \mathbf{V} S_n S_n^* \mathbf{V}^\top (\mathbf{V}\mathbf{V}^\top)^{-1} = (\mathbf{V}\mathbf{V}^\top)^{-1} \mathbf{V}\mathbf{V}^\top \mathbf{V}\mathbf{V}^\top (\mathbf{V}\mathbf{V}^\top)^{-1} = I_r.$$

Now let us show that $O_n O_n^* = \Pi_n$. First of all, $\tilde{\Pi}_n := O_n O_n^*$ is self adjoint and is a projection operator since $\tilde{\Pi}_n^2 = O_n (O_n^* O_n) O_n^* = O_n O_n^* = \tilde{\Pi}_n$ by the previous point. Moreover, its range is included in $\mathrm{span}(\phi(x_i))_{1 \leqslant i \leqslant n}$ since $O_n = S_n^* \tilde{O}_n$ for a certain $\tilde{O}_n$ and the range of S_n^* is $\mathrm{span}(\phi(x_i))_{1 \leqslant i \leqslant n}$. Finally since the rank of S_n^* is also the rank of $S_n S_n^*$ which is r, we deduce that the range of $\mathrm{span}(\phi(x_i))_{1 \leqslant i \leqslant n}$ is of dimension r and hence, since $O_n^* O_n = I_r$ implies that $O_n O_n^*$ is of rank r, putting things together, $\tilde{\Pi}_n = \Pi_n$.

Remark 5 (Constructing $\mathbf{V}$)**.** *In the case where the kernel matrix $\mathbf{K}$ is full rank, $\mathbf{V} \in \mathbb{R}^{n \times n}$ and is invertible, and O_n can be simply written $S_n^* \mathbf{V}^{-1}$.*
In the case where the kernel matrix $\mathbf{K}$ is not full-rank, we build $\mathbf{V}$ as $\mathbf{V} = \mathbf{\Sigma}^{1/2} \mathbf{U}^\top$, where $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is diagonal and $\mathbf{U} \in \mathbb{R}^{n \times r}$ is unitary and correspond to the economy eigendecomposition *of $\mathbf{K}$ where r is the rank of $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$.*

Consider the following generalization of the finite dimensional model proposed in Eq. (8) in the case where $\mathbf{K}$ is not necessarily full rank :

$$\tilde{f}_{\mathbf{A}}(x) = \Phi(x)^\top \mathbf{A} \Phi(x), \qquad \mathbf{A} \in \mathbb{R}^{r \times r},\ \mathbf{A} \succeq 0, \tag{8}$$

where $\Phi : \mathcal{X} \mapsto \mathbb{R}^r$ is defined as $\Phi(x) = O_n^* \phi(x) = (\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V} v(x)$, where $v(x) = (k(x_i, x))_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$.

We are now ready to prove Proposition 3.

Proof of Proposition 3. Recall

$$\min_{\mathbf{A} \succeq 0} L(\tilde{f}_{\mathbf{A}}(x_1), \dots, \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}). \tag{9}$$

The fact that Eq. (9) has a solution, and that this solution is unique if $\lambda_2 > 0$ and L is convex can be seen as a simple consequence of Thm. 7 in the case where the model considered is the finite dimensional model defined in Eq. (8). Let us now prove the other part of the proposition.

Start by noting that with our definition of O_n, for all $\mathbf{A} \in \mathbb{R}^{r \times r},\ \mathbf{A} \succeq 0$,

$$f_{O_n \mathbf{A} O_n^*} = \tilde{f}_{\mathbf{A}}. \tag{a}$$

Moreover,

$$\left\{ O_n \mathbf{A} O_n^* \ :\ \mathbf{A} \in \mathbb{R}^{r \times r},\ \mathbf{A} \succeq 0 \right\} = \mathcal{S}_n(\mathcal{H})_+. \tag{b}$$

Finally, since O_n is an isometry which implies $\Omega(O_n \mathbf{A} O_n^*) = \Omega(\mathbf{A})$ and by Eq. (a), for any $\mathbf{A} \in \mathcal{S}(\mathbb{R}^n)_+$, it holds :

$$L(f_{O_n \mathbf{A} O_n^*}(x_1), ..., f_{O_n \mathbf{A} O_n^*}(x_n)) + \Omega(O_n \mathbf{A} O_n^*) = L(\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n)) + \Omega(\mathbf{A}). \tag{c}$$

Now combining Eq. (c) and Eq. (b), any solution $\mathbf{A}_*$ to Eq. (9) corresponds to a solution $A_* \in \operatorname{argmin}_{A \in \mathcal{S}_n(\mathcal{H})_+} L(f_A(x_1), ..., f_A(x_n)) + \Omega(A)$, where $A_* = O_n \mathbf{A}_* O_n^*$. Now using Eq. (15) in the proof of Proposition 7, we see that A_* is also a minimizer of Eq. (5) hence the result.

Note that the fact that the condition number of the problem, if it exists, is preserved because O_n is an isometry.

B.5 Proof of Thm. 2 and algorithmic consequence.

In this section, we prove Thm. 2 and explain how to derive an efficient algorithm to solve it in certain cases.

Let us start by proving the following lemma.

Lemma 5. *Let $\lambda_1, \lambda_2 \geqslant 0$ and assume $\lambda_2 > 0$. Let Ω_+ be defined on $\mathcal{S}(\mathbb{R}^r)$ as follows :*

$$\Omega_+(A) = \begin{cases} \lambda_1 \|A\|_\star + \frac{\lambda_2}{2}\|A\|_F^2 & \text{if } A \succeq 0; \\ +\infty & \text{otherwise .} \end{cases}$$

Then Ω_+ is a closed convex function, and its Fenchel conjugate is given for any $B \in \mathcal{S}(\mathbb{R}^r)$ by the formula:

$$\Omega_+^*(B) = \frac{1}{2\lambda_2} \left\| [B - \lambda_1 I]_+ \right\|_F^2 .$$

Moreover, Ω_+ is differentiable at every point, and is $1/\lambda_2$ smooth. Its gradient is given by:

$$\nabla \Omega_+^*(B) = \frac{1}{\lambda_2} [B - \lambda_1 I]_+ .$$

Proof. Write

$$\Omega_+(A) = \iota_{\mathcal{S}(\mathbb{R}^r)_+} + \lambda_1\|A\|_\star + \frac{\lambda_2}{2}\|A\|_F^2.$$

Here, ι_C stands for the characteristic function of the convex set C, i.e. $\iota_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise. Since $\|\cdot\|_F^2$ and $\|\cdot\|_\star$ are both convex, continuous, and real valued, and since $\iota_{\mathcal{S}(\mathbb{R}^r)_+}$ is closed since $\mathcal{S}(\mathbb{R}^r)_+$ is a closed non-empty convex subset of $\mathcal{S}(\mathbb{R}^r)$, this shows that Ω_+ is indeed convex and closed. Note that it is continuous on its domain $\mathcal{S}(\mathbb{R}^r)_+$. Moreover, it is strongly convex since $\lambda_2 > 0$. Fix $B \in \mathcal{S}(\mathbb{R}^r)$ and consider the problem

$$\sup_{A\in\mathcal{S}(\mathbb{R}^r)} \operatorname{Tr}(AB) - \Omega_+(A) = \sup_{A\succeq 0} \operatorname{Tr}(A(B-\lambda_1 I)) - \frac{\lambda_2}{2}\|A\|_F^2$$

Since Ω_+ is strongly convex, we know there exists a unique solution to this problem.

Note that $A_* = \operatorname{argmax} \operatorname{Tr}(AB) - \Omega_+(A)$ if and only if

$$A_* = \operatorname{argmin}_{A\in\mathcal{S}(\mathbb{R}^r)_+} \frac{1}{2}\left\|\left(A - \frac{1}{\lambda_2}(B-\lambda_1 I)\right)\right\|^2.$$

That is A_* is the orthogonal projection of $\frac{B-\lambda_1 I}{\lambda_2}$ on $\mathcal{S}(\mathbb{R}^r)_+$ for the Frobenius scalar product. Hence, $A_* = \left[\frac{B-\lambda_1 I}{\lambda_2}\right]_+$.

Here, for any symetric matrix C, we denote with $[C]_+$ resp $[C]_-$ its positive resp negative part. Given an eigendecomposition $C = U\Sigma U^T$ with Σ diagonal, they are defined by $[C]_+ = U\max(0,\Sigma)U^T$ and $[C]_- = U\max(0,-\Sigma)U^T$. Hence, the Fenchel conjugate of Ω_+ is given by

$$\Omega_+^*(B) = \frac{1}{2\lambda_2}\left\|[B-\lambda_1 I]_+\right\|_F^2.$$

Consider $\omega_+^* : \sigma \in \mathbb{R} \mapsto \max(0,\sigma^2) \in \mathbb{R}$. ω_+^* is 1-smooth and differentiable, and $(\omega_+^*)'(\sigma) = \max(0,\sigma)$. Hence, the function

$$B \mapsto \operatorname{Tr}(\omega_+^*(B)) = \|[B]_+\|_F^2$$

is differentiable and 1-smooth, with differential given by the spectral function $(\omega_+^*)'(B) = [B]_+$. Hence, Ω_+ is differentiable and $\nabla\Omega_+^*(B) = \frac{1}{\lambda_2}[B-\lambda_1 I]_+$, and is $1/\lambda_2$ smooth.

Theorem 8 (Convex dual problem)**.** *Let $L : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be convex closed function and L^* be the Fenchel conjugate of L (see [BV04] for the definition of closed and of the dual conjugate). Assume Ω is of the form Eq. (6). Assume there exists $\mathbf{A} \in \mathbb{R}^{r\times r}$, $\mathbf{A} \succeq \mathbf{0}$ such that L is continuous in $(\tilde{f}_{\mathbf{A}}(x_i))_{1\leqslant i\leqslant n}$.*

Then the problem in Eq. (9) has the following dual formulation,

$$\sup_{\alpha\in\mathbb{R}^n} -L^*(\alpha) - \tfrac{1}{2\lambda_2}\|[\mathbf{V}\operatorname{Diag}(\alpha)\mathbf{V}^\top + \lambda_1\mathbf{I}]_-\|_F^2, \tag{10}$$

and this supremum is atteined. Let $\alpha^* \in \mathbb{R}^n$ be a solution of (10). Then, the solution of (5) is obtained via (7), with $\mathbf{B} \in \mathbb{R}^{r\times r}, \mathbf{B} \succeq 0$ as

$$\mathbf{B} = \mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}}\left(\frac{1}{\lambda_2}\left[\mathbf{V}\operatorname{Diag}(\alpha_*)\mathbf{V}^\top + \lambda_1 I\right]_-\right)(\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}}\mathbf{V}. \tag{11}$$

Proof of Thm. 8.

We apply theorem 3.3.1 of [BL10] with the following parameters (on le left, the ones in theorem 3.3.1 of [BL10] and on the right the ones by which we replace them).

$\mathbf{E}$	$\mathcal{S}(\mathbb{R}^r)$
$\mathbf{Y}$	$\mathbb{R}^n$
$A : \mathbf{E} \to \mathbf{Y}$	$R : \mathbf{A} \in \mathcal{S}(\mathbb{R}^r) \mapsto (\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n)) \in \mathbb{R}^n$
$f : \mathbf{E} \to]-\infty, +\infty]$	$\Omega_+ : \mathcal{S}(\mathbb{R}^r) \to]-\infty, +\infty]$
$g : \mathbf{Y} \to]-\infty, +\infty]$	$L : \mathbb{R}^n \to]-\infty, +\infty]$
$p = \inf_{x \in \mathbf{E}} g(Ax) + f(x)$	$p = \inf_{\mathbf{A} \in \mathcal{S}(\mathbb{R}^r)} L(\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n)) + \Omega_+(\mathbf{A})$
$d = \sup_{\phi \in \mathbf{Y}} -g^*(\phi) - f^*(-A^*\phi)$	$d = \sup_{\alpha \in \mathbb{R}^n} -L^*(\alpha) - \Omega_+^*(-R^*(\alpha))$

Indeed, for all $1 \leqslant i \leqslant n$, if Φ is defined in Eq. (8), $\Phi(x_i) = \mathbf{V}e_i$ and thus $\tilde{f}_{\mathbf{A}}(x_i) = \Phi(x_i)^\top \mathbf{A}\Phi(x_i) = e_i^\top(\mathbf{V}^\top\mathbf{A}\mathbf{V})e_i$. Thus, for any $\mathbf{A} \in \mathcal{S}(\mathbb{R}^r)$, $R(\mathbf{A}) := \left(\tilde{f}_{\mathbf{A}}(x_i)\right)_{1 \leqslant i \leqslant n} = \mathrm{Diag}(\mathbf{V}^\top\mathbf{A}\mathbf{V})$. The following properties are satisfied :

- L is lower semi-continuous, convex and bounded below hence closed (see [BL10]);
- similarly, Ω_+ is a non negative closed convex function, with dual Ω_+^* given in Lemma 5 which is differentiable and smooth;
- $\mathrm{dom}(\Omega_+) = \mathcal{S}(\mathbb{R}^n)_+$;
- R is linear, and for any $\alpha \in \mathbb{R}^n$, it holds $R^*\alpha = \mathbf{V}\,\mathrm{Diag}(\alpha)\mathbf{V}^\top$;
- The dual d can therefore be re-expressed as Eq. (10), using the expressions for Ω_+^* and R^* :

$$\sup_{\alpha \in \mathbb{R}^n} -L^*(\alpha) - \frac{1}{2\lambda_2} \left\| \left[\mathbf{V}\,\mathrm{Diag}(\alpha)\mathbf{V}^\top + \lambda_1 I\right]_- \right\|_F^2 \tag{10}$$

- Assume there exists $\mathbf{A} \in \mathbb{R}^{r \times r}$, $\mathbf{A} \succeq \mathbf{0}$ such that L is continuous in $(\tilde{f}_{\mathbf{A}}(x_i))_{1 \leqslant i \leqslant n}$. Then there exists a point of continuity of g such which is also in $R\,\mathrm{dom}\, f$, hence the assumption of theorem 3.3.1 of [BL10] is satisfied.

Applying theorem 3.3.1 of [BL10], the following properties hold:

- $d = p$,
- d is atteined for a certain $\alpha_* \in \mathbb{R}^n$. Indeed, there exists $\mathbf{A} \in \mathrm{dom}\,\Omega_+$ such that $R(\mathbf{A}) \in \mathrm{dom}(L)$. Thus , $L(R(\mathbf{A})) + \Omega_+(\mathbf{A}) < +\infty$ and hence $d < +\infty$. Moreover, since L and Ω_+ are lower bounded, this shows that d is lower bounded and hence $d > -\infty$. Hence d is finite and thus is atteined by theorem 3.3.1.

Now using Exercise 4.2.17 of [BL10] since L and Ω_+ are closed convex and since Ω_+^* is differentiable, we see that the optimal solution of the primal problem $\mathbf{A}_*$ is given by the following formula:

$$\mathbf{A}_* = \nabla\Omega_+^*(-R^*\alpha^*) = \frac{1}{\lambda_2}\left[\mathbf{V}\,\mathrm{Diag}(\alpha_*)\mathbf{V}^\top + \lambda_1 I\right]_- .$$

Thus, for any $x \in \mathcal{X}$, using the definition of $\Phi(x)$, it holds

$$\begin{aligned}\tilde{f}_{\mathbf{A}}(x) &= \Phi(x)^\top \mathbf{A}_* \Phi(x) \\ &= v(x)^\top \mathbf{V}^\top (\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}} \left(\frac{1}{\lambda_2}\left[\mathbf{V}\,\mathrm{Diag}(\alpha_*)\mathbf{V}^\top + \lambda_1 I\right]_-\right)(\mathbf{V}\mathbf{V}^\top)^{-\mathbf{1}}\mathbf{V}v(x).\end{aligned}$$

Thus, setting

$$\mathbf{B} = \mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-1}\left(\frac{1}{\lambda_2}\left[\mathbf{V}\,\mathrm{Diag}(\alpha_*)\mathbf{V}^\top + \lambda_1 I\right]_-\right)(\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V},$$

it holds $\tilde{f}_{\mathbf{A}}(x) = v(x)^\top \mathbf{B} v(x)$. Since $v(x) = (k(x, x_i))_{1\leqslant i\leqslant n} \in \mathbb{R}^n$, this shows the result. In particular, note that when $\mathbf{V}$ is invertible (i.e. when $\mathbf{K}$ is full rank) then the equation above is exactly Eq. (11), since $\mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-1} = \mathbf{V}^{-1}$.

Proof of Thm. 2. It is a direct consequence of the previous theorem.

Note that the conditions of theorem Thm. 2 are satisfied in many interesting cases, such as the ones described in the following proposition.

Proposition 8. *Assume one of the following conditions is satisfied :*

(i) $\mathrm{dom}(L) = \mathbb{R}^n$;

(ii) $\mathbb{R}^n_{++} \subset \mathrm{dom}(L)$ *and* $k(x_i, x_i) > 0$ *for all* $1 \leqslant i \leqslant n$

(iii) $\mathbf{K}$ *is full rank and there exists a continuity point* α_0 *of* L *such that* $\alpha_0 \in \mathbb{R}^n_+$.

Then there exists $\mathbf{A} \in \mathcal{S}(\mathbb{R}^{\mathbf{n}})_+$ *such that* L *is continuous in* $(\tilde{f}_{\mathbf{A}}(x_1), ..., \tilde{f}_{\mathbf{A}}(x_n))$.

Proof. Let us prove these points.

- if $\mathrm{dom}(L) = \mathbb{R}^n$, since L is convex, L is continuous everywhere. Taking $\mathbf{A} = \mathbf{0}$, the result holds.
- if $k(x_i, x_i) > 0$ for all $i > 0$, then taking $\mathbf{A} = I_r$, we have $(\tilde{f}_{\mathbf{A}}(x_i))_{1\leqslant i\leqslant n} = (k(x_i, x_i))_{1\leqslant i\leqslant n}$ which is in $\mathbb{R}^n_{++}$. Since $\mathbb{R}^n_{++} \subset \mathrm{dom}(L)$ and $\mathbb{R}^n_{++}$ is open, L is continuous on $\mathbb{R}^n_{++}$ and hence, $\mathbf{A}$ satisfies the desired property.
- Let α_0 be a continuity point of L in $\mathbb{R}^n_+$. If we assume $\mathbf{K}$ is full rank, then in particular, $\mathbf{V} \in \mathbb{R}^{n\times n}$ is of rank n and invertible. Thus, there exists $\mathbf{A} \in \mathcal{S}(\mathbb{R}^r)_+$ such that

$$\mathbf{V}^\top\mathbf{A}\mathbf{V} = \mathrm{Diag}(\alpha_0) \implies (\tilde{f}_{\mathbf{A}}(x_i))_{1\leqslant i\leqslant n} = \alpha_0.$$

Discussion on how to solve Eq. (10) Proximal splitting methods can be applied to solve Eq. (10) such as FISTA [BT09], provided the proximal operator of L^* can be computed (see [PB14] for the definition of the proximal operator). Indeed, Eq. (10) can be written as

$$\min_{\alpha\in\mathbb{R}^n} F(\alpha) = f(\alpha) + g(\alpha), \qquad f(\alpha) = \Omega^*_+(-\mathbf{V}\,\mathrm{Diag}(\alpha)\mathbf{V}^\top),\ g(\alpha) = L^*(\alpha).$$

where Ω^*_+ has been defined in Lemma 5 and has been shown to be smooth and differentiable. Thus, since $\alpha \mapsto \mathbf{V}\,\mathrm{Diag}(\alpha)\mathbf{V}^\top$ is linear, f is smooth and differentiable. Moreover, one can have access to the gradient of f by performing an eigenvalue decomposition of $\mathbf{V}\,\mathrm{Diag}(\alpha)\mathbf{V}^\top$ whose complexity is bounded above by $\mathcal{O}(r^3)$. Thus, one can apply one of the algorithms in section 4 of [BT09] in order to compute an optimal solution to Eq. (10). Moreover, a bound on the performance of the algorithm is given in theorem 4.4 of this same work. Note that if L is of the form $L(\alpha) = \sum_{i=1}^n \ell_i(\alpha_i)$, it suffices to be able to compute the proximal operator of the ℓ_i to get a proximal operator for L^* (see [PB14]).

B.6 Proof and additional discussion of Thm. 3

We recall the notion of universality [MXZ06], in particular *cc-universality* [SFL11], here explicited in the context of non-negative functions. A set $\mathcal{F}$ is a *universal approximator* for non-negative functions on $\mathcal{X}$ if, for any compact subset $\mathcal{Z}$ of $\mathcal{X}$, we have that the set $\mathcal{F}|_{\mathcal{Z}}$ of restrictions on $\mathcal{Z}$, defined as $\mathcal{F}|_{\mathcal{Z}} = \{f|_{\mathcal{Z}} \mid f \in \mathcal{F}\}$, is dense in the set $C^+(\mathcal{Z})$ of non-negative continuous functions over $\mathcal{Z}$ in the maximum norm. In the following theorem we prove the cc-universality of the proposed model

Theorem 9. *Let $\mathcal{X}$ be a locally compact Hausdorff space, $\mathcal{H}$ a separable Hilbert space and $\phi : \mathcal{X} \to \mathcal{H}$ a cc-universal feature map. Let $\|\cdot\|_\circ$ be a norm for $\mathcal{S}(\mathcal{H})$ such that $\|\cdot\|_\star \geq \|\cdot\|_\circ$. Then $\mathcal{F}_\phi^\circ$ is a cc-universal approximator for the non-negative functions on $\mathcal{X}$.*

Proof. Proving that the proposed model is a cc-universal approximator for non-negative functions, is equivalent to require that given a compact set $\mathcal{Z} \subseteq \mathcal{X}$, a non-negative function $g : \mathcal{Z} \to \mathbb{R}_+$ and $\epsilon > 0$, there exists $f_{A_{g,\mathcal{Z},\epsilon}} \in \mathcal{F}_\phi^\circ$ such that $\|g - f_{A_{g,\mathcal{Z},\epsilon}}\|_{C(Z)} \leqslant \epsilon$. In particular, let $Q = 2\|g\|_{C(\mathcal{Z})}^{1/2} + \epsilon^{1/2}$, since ϕ is *cc*-universal, given $\mathcal{Z}, g, \epsilon$, there exists $w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}$ such that $\|\sqrt{g} - \phi(\cdot)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}\|_{C(Z)} \leqslant \frac{\epsilon}{Q}$. Define $A_{g,\mathcal{Z},\epsilon} = w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \otimes w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}$. Note that for any $x \in \mathcal{X}$,

$$f_{A_{g,\mathcal{Z},\epsilon}}(x) = \phi(x)^\top A_{g,\mathcal{Z},\epsilon}\phi(x) = \phi(x)^\top \left(w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \otimes w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \right) \phi(x) = (\phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}})^2. \tag{16}$$

Then, by denoting with $h(x) = \sqrt{g(x)} - \phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}}$, we have

$$\begin{aligned}
\|g - f_{A_{g,\mathcal{Z},\epsilon}}\|_{C(Z)} &= \sup_{x\in\mathcal{Z}} |g(x) - (\phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}})^2| && (17)\\
&= \sup_{x\in\mathcal{Z}} \left| \left(\sqrt{g(x)} - \phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \right) \left(\sqrt{g(x)} + \phi(x)^\top w_{\sqrt{g},\mathcal{Z},\frac{\epsilon}{Q}} \right) \right| && (18)\\
&= \sup_{x\in\mathcal{Z}} |h(x)(2\sqrt{g(x)} - h(x))| && (19)\\
&\leqslant \|h\|_{C(\mathcal{Z})}(2\|\sqrt{g}\|_{C(\mathcal{Z})} + \|h\|_{C(\mathcal{Z})}) && (20)\\
&\leqslant \frac{\epsilon}{Q}\left(2\|g\|_{C(\mathcal{Z})}^{1/2} + \frac{\epsilon}{Q}\right) \leqslant \epsilon. && (21)
\end{aligned}$$

The last step is due to the fact that $\epsilon/Q \leqslant \sqrt{\epsilon}$, then $2\|g\|_{C(\mathcal{Z})}^{1/2} + \frac{\epsilon}{Q} \leqslant Q$.

B.7 Proof and additional discussion of Thm. 4

In Thm. 10, stated below, we prove that $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi$ under the very general assumption that $\mathcal{G}_\phi$ is a multiplication algebra, i.e.. if $\mathcal{G}_\phi$ is closed under pointwise product of the functions. In Thm. 11 we specify this result when $\mathcal{G}_\phi$ is a Sobolev space, proving that $\mathcal{E}_\phi \subsetneq \mathcal{F}_\phi^\circ$. Thm. 4 is a direct consequence of the latter theorem.

General result when $\mathcal{G}_\phi$ is a multiplication algebra. First we endow $\mathcal{G}_\phi$ with a Hilbertian norm. Define $\|\cdot\|_{\mathcal{G}_\phi}$ as $\|f_w\|_{\mathcal{G}_\phi} = \|w\|_{\mathcal{H}}$, for any $w \in \mathcal{H}$.

Definition 3. *$\mathcal{G}_\phi$ is a multiplication algebra, when there exists a constant C such that the unit function $u : \mathcal{X} \to \mathbb{R}$ that maps $x \mapsto 1$ for any $x \in \mathcal{X}$ is in $\mathcal{G}_\phi$ and*

$$\|f \cdot g\|_{\mathcal{G}_\phi} \leqslant C\|f\|_{\mathcal{G}_\phi}\|g\|_{\mathcal{G}_\phi}, \qquad \forall\, f, g \in \mathcal{G}_\phi, \tag{22}$$

where we denote by $f \cdot g$ the pointwise multiplication, i.e., $(f \cdot g)(x) = f(x)g(x)$ for all $x \in \mathcal{X}$.

Remark 6 (Renormalizing the constant)**.** *Note that when $\mathcal{G}_\phi$ is a multiplication algebra for a constant C, it is always possible to define an equivalent norm $\|\cdot\|'_{\mathcal{G}_\phi}$ as $\|\cdot\|'_{\mathcal{G}_\phi} = C\|\cdot\|_{\mathcal{G}_\phi}$ for which $\mathcal{G}_\phi$ is a multiplication algebra with constant* 1.

Theorem 10 (General version when $\mathcal{G}_\phi$ is an algebra)**.** *Let $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Let $\mathcal{X}$ be a compact space and ϕ be a bounded continuous map such that $\mathcal{G}_\phi$ is a multiplication algebra, then $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi^\circ$.*

Proof. Let $g \in \mathcal{E}_\phi$ and take $f \in \mathcal{G}_\phi$ such that $g(x) = e^{f(x)}$ for all $x \in \mathcal{X}$. First we prove that $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi^\circ$. With this goal, first we prove that $\sqrt{g} \in \mathcal{G}_\phi$ and then we construct a rank one positive operator such that $f_{A_g}(x) = g(x)$ for every $x \in \mathcal{X}$. We start noting that, given $f \in \mathcal{G}_\phi$ and $t \in \mathbb{N}$, f^t defined by $f \cdot f^{t-1}$ for $t \in \mathbb{N}$ satisfies $f^t \in \mathcal{G}_\phi$, with $\|f^t\|_{\mathcal{G}_\phi} \leqslant C^t\|f\|^t_{\mathcal{G}_\phi}$, by repeated application of the Eq. (22). Moreover note that the function $s = \sum_{t\in\mathbb{N}} \frac{1}{2^t t!} f^t$, satisfies $s \in \mathcal{G}_\phi$, indeed

$$\|s\|_{\mathcal{G}_\phi} \leqslant \sum_{t\in\mathbb{N}} \frac{1}{2^t t!}\|f^t\|_{\mathcal{G}_\phi} \leqslant \sum_{t\in\mathbb{N}} \frac{1}{2^t t!} C^t\|f\|^t_{\mathcal{G}_\phi} \leqslant e^{C\|f\|_{\mathcal{G}_\phi}/2}.$$

Moreover s satisfies $s(x) = \sqrt{g(x)}$ for all $x \in \mathcal{X}$, indeed for $x \in \mathcal{X}$ we have

$$s(x) = \phi(x)^\top s = \sum_{t\in\mathbb{N}} \frac{1}{2^t t!}\phi(x)^\top f^t = \sum_{t\in\mathbb{N}} \frac{1}{2^t t!} f^t(x) = e^{f(x)/2} = \sqrt{g(x)}.$$

Now let $A_g = s \otimes s$, we have that $\|A_g\|_\circ \leqslant \|A_g\|_\star$ by assumption, and $\|A_g\|_\star = \|s\|^2_{\mathcal{G}_\phi} < \infty$, so the function $f_{A_g} \in \mathcal{F}_\phi^\circ$ and for any $x \in \mathcal{X}$

$$f_{A_g}(x) = \phi(x)^\top A_g \phi(x) = \phi(x)^\top (s \otimes s)\phi(x) = (\phi(x)^\top s)^2 = g(x).$$

Since for any $g \in \mathcal{E}_\phi$ there exists $f_{A_g} \in \mathcal{F}_\phi^\circ$ that is equal to g on their domain of definition, we have that $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi^\circ$.

Now we are going to specialize the result above for Sobolev spaces.

Result for Sobolev spaces The result below is based on the general result in Thm. 10, however it is possible to do a proof based only on norm inequalities for compositions of functions in Sobolev space (see for example [BM01]). While more technical, this second approach would allow to derive also a more quantitative analysis on the norms of the functions in $\mathcal{G}_\phi$ and $\mathcal{F}_\phi^\circ$. We will leave this for a longer version of this work.

Theorem 11. *Let $\|\cdot\|_\star \trianglerighteq \|\cdot\|_\circ$. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{X}$ compact with locally Lipschitz boundary and let $\mathcal{G}_\phi = W_2^m(\mathcal{X})$. Let $x_0 \in \mathcal{X}$. Then the following holds:*

(a) $\mathcal{E}_\phi \subsetneq \mathcal{F}_\phi^\circ$. (b) The function $f_{x_0}(x) = e^{-\|x-x_0\|^{-2}} \in C^\infty(\mathcal{X})$ satisfies $f_{x_0} \in \mathcal{F}_\phi^\circ$ and $f_{x_0} \notin \mathcal{E}_\phi$.

Proof. First we prove that $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi^\circ$, via Thm. 10, then we. To apply this result we need first to prove that $\mathcal{G}_\phi = W_2^m(\mathcal{X})$ is a multiplication algebra when $W_2^m(\mathcal{X})$ is a RKHS as in our case.
Step 1, $m > d/2$**.** First note that $\mathcal{G}_\phi$ satisfies $m > d/2$ since $W_2^m(\mathcal{X})$ admits a representation in terms of a separable Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$, i.e., it is a *reproducing Kernel Hilbert space* and for the same reason $\|\cdot\|_{\mathcal{G}_\phi}$ is equivalent to $\|\cdot\|_{W_2^m(\mathcal{X})}$ [Wen04].

Step 2. $\mathcal{G}_\phi$ is a multiplication algebra. Applying Thm. 10. Since $\mathcal{G}_\phi = W_2^m(\mathcal{X})$ with $m > d/2$, then it is a multiplication algebra. This result is standard (e.g. see pag. 106 of [AF03] for $m \in \mathbb{N}$ and $\mathcal{X} = \mathbb{R}^d$) and we report it in Lemma 8 in Appendix C. Then we apply Thm. 10 obtaining $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi^\circ$.

Step 3. Proving that $f_{x_0} \in \mathcal{F}_\phi^\circ$ and not in $\mathcal{E}_\phi$. By construction the function $v(x) = e^{-1/(2\|x-x_0\|^2)}$ is in $C^\infty(\mathcal{X})$ and so in $W_2^m(\mathcal{X})$ for any $m \geqslant 0$. Since $\mathcal{G}_\phi = W_2^m(\mathcal{X})$, then $v \in \mathcal{G}_\phi$, i.e., there exists $w \in \mathcal{H}$ such that $w^\top \phi(\cdot) = v(\cdot)$. Define $A_v = w \otimes w$, then

$$f_{A_v}(x) = \phi(x)^\top A_v \phi(x) = (w^\top \phi(x))^2 = v^2(x) = f_{x_0}(x), \quad \forall x \in \mathcal{X}.$$

Then $f_{x_0} = f_{A_v}$ on $\mathcal{X}$, i.e., $f_{x_0} \in \mathcal{F}_\phi^\circ$. To conclude note that, f_{x_0} does not belong to $\mathcal{E}_\phi$, since $x_0 \in \mathcal{X}$ and $f_{x_0}(x_0) = 0$, while for any $g \in \mathcal{E}_\phi$ we have $\inf_{x \in X} g(x) > 0$. Indeed, we have that for any $f \in \mathcal{G}_\phi$, $\|f\|_{C(\mathcal{X})} = \sup_{x \in \mathcal{X}} |f(x)| < \infty$, since $\mathcal{G}_\phi = W_2^m(\mathcal{X}) \subset C(\mathcal{X})$. Moreover, given $g \in \mathcal{G}_\phi$, and denoting by $f \in \mathcal{G}_\phi$ the function such that $g = e^f$, we have that $\inf_{x \in \mathcal{X}} g(x) \geqslant e^{-\|f\|_{C(\mathcal{X})}} > 0$. Finally, since $\mathcal{E}_\phi \subseteq \mathcal{F}_\phi^\circ$, but there exists $f_{x_0} \in \mathcal{F}_\phi^\circ$ and not in $\mathcal{E}_\phi$, then $\mathcal{E}_\phi \subsetneq \mathcal{F}_\phi^\circ$.

Proof of Thm. 4. This result is a direct application of Thm. 11, since $\mathcal{X} = [-R, R]^d$, with $R \in (0, \infty)$ is a compact set with Lipschitz boundary.

B.8 Proof of Thm. 5

We recall here the Rademacher complexity and prove Thm. 5. This latter theorem is obtained from the following Thm. 12 that bounds the *empirical Rademacher complexity* introduced below. First we recall that the function class $\mathcal{F}_{\phi,L}^\circ$ is defined as

$$\mathcal{F}_{\phi,L}^\circ = \{f_A \mid A \succeq 0, \|A\|_\circ \leqslant L\},$$

for a given norm $\|\cdot\|_\circ$ on operators, a feature map $\phi : \mathcal{X} \to \mathcal{H}$ and $L > 0$. Now we define the empirical Rademacher complexity and the Rademacher complexity [BM02]. Given $x_1, \ldots, x_n \in \mathcal{X}$, the empirical Rademacher complexity for a class $\mathcal{F}$ of functions mapping $\mathcal{X}$ to $\mathbb{R}$, is defined as

$$\widehat{R}_n(\mathcal{F}) = 2\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right|,$$

where σ_i independent Rademacher random variables, i.e., $\sigma_i = -1$ with probability $1/2$ and $+1$ with probability $1/2$ and the expectation is on $\sigma_1, \ldots, \sigma_n$. Let ρ be a probability distribution on $\mathcal{X}$ and $x_1, \ldots, x_n$ sampled independently according to ρ. The *Rademacher complexity* $R_n(\mathcal{F})$ is defined as

$$R_n(\mathcal{F}) = \mathbb{E}\widehat{R}_n(\mathcal{F}),$$

where the last expectation is on $x_1, \ldots, x_n$. In the following theorem we bound $\widehat{R}_n$.

Theorem 12. *Let $\|\cdot\|_\circ \trianglerighteq \|\cdot\|_F$. Let $x_1, \ldots, x_n \in \mathcal{X}$, $L \geqslant 0$.*

$$\widehat{R}_n(\mathcal{F}_{\phi,L}^\circ) \leqslant \frac{2L}{n} \sqrt{\sum_{i=1}^n \|\phi(x_i)\|^4}.$$

Proof. Given $f_A \in \mathcal{F}^{\circ}_{\phi,L}$, since $\|\cdot\|_{\circ}$ is stronger or equivalent to Hilbert-Schmidt norm, we have that $\|A\|_F \leqslant \|A\|_{\circ} \leqslant L$. Since A is bounded and $\phi(\cdot) \in \mathcal{H}$, by linearity of the trace we have $f_A(x) = \phi(x)^\top A\phi(x) = \mathrm{Tr}(A\,\phi(x) \otimes \phi(x))$ for any $x \in \mathcal{X}$. Then, by linearity of the trace

$$\hat{R}_n(\mathcal{F}^{\circ}_{\phi,L}) = 2\mathbb{E} \sup_{f\in\mathcal{F}^{\circ}_{\phi,L}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right| = 2\mathbb{E} \sup_{A\succeq 0, \|A\|_{\circ}\leqslant L} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \phi(x_i)^\top A\phi(x_i)\right| \tag{23}$$

$$= 2\mathbb{E} \sup_{A\succeq 0, \|A\|_{\circ}\leqslant L} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \,\mathrm{Tr}(A\,(\phi(x_i)\otimes\phi(x_i)))\right| \tag{24}$$

$$= 2\mathbb{E} \sup_{A\succeq 0, \|A\|_{\circ}\leqslant L} \left|\mathrm{Tr}\left(A\,\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\phi(x_i)\otimes\phi(x_i)\right)\right)\right| \tag{25}$$

Now since $\|\cdot\|_{\circ}$ is stronger or equivalent to $\|\cdot\|_F$ this means that $\{A \in \mathcal{S}(\mathcal{H}) \mid \|A\|_{\circ} \leqslant L\} \subseteq \{A \in \mathcal{S}(\mathcal{H}) \mid \|A\|_F \leqslant L\}$, then

$$2\mathbb{E} \sup_{A\succeq 0, \|A\|_{\circ}\leqslant L} \left|\mathrm{Tr}\left(A\,\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\phi(x_i)\otimes\phi(x_i)\right)\right)\right| \tag{26}$$

$$\leqslant 2\mathbb{E} \sup_{A\succeq 0, \|A\|_{F}\leqslant L} \left|\mathrm{Tr}\left(A\,\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\phi(x_i)\otimes\phi(x_i)\right)\right)\right| \tag{27}$$

$$\leqslant 2\mathbb{E} \sup_{A\succeq 0, \|A\|_{F}\leqslant L} \|A\|_F \left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi(x_i)\otimes\phi(x_i)\right\|_F \tag{28}$$

$$\leqslant 2L\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi(x_i)\otimes\phi(x_i)\right\|_F. \tag{29}$$

To conclude denote by ζ_i the random variable $\sigma_i\phi(x_i)\otimes\phi(x_i)$. Then

$$\mathbb{E}\left\|\tfrac{1}{n}\sum_{i=1}^n \sigma_i\phi(x_i)\otimes\phi(x_i)\right\|_F^2 = \mathbb{E}\,\|\tfrac{1}{n}\sum_{i=1}^n \zeta_i\|_F$$

$$= \mathbb{E}\sqrt{\mathrm{Tr}\left(\left(\tfrac{1}{n}\sum_{i=1}^n \zeta_i\right)^*\left(\tfrac{1}{n}\sum_{i=1}^n \zeta_i\right)\right)} = \mathbb{E}\sqrt{\mathrm{Tr}\left(\tfrac{1}{n^2}\sum_{i,j=1}^n \zeta_i\zeta_j\right)}.$$

By Jensen inequality, the concavity of the square root, and the linearity of the trace

$$\mathbb{E}\sqrt{\mathrm{Tr}\left(\tfrac{1}{n^2}\sum_{i,j=1}^n \zeta_i\zeta_j\right)} \leqslant \sqrt{\mathbb{E}\;\mathrm{Tr}\left(\tfrac{1}{n^2}\sum_{i,j=1}^n \zeta_i\zeta_j\right)} = \sqrt{\tfrac{1}{n^2}\sum_{i,j=1}^n \mathrm{Tr}(\mathbb{E}\zeta_i\zeta_j)}.$$

Now note that for $i \in \{1,\ldots,n\}$, we have $\mathbb{E}_{\sigma_i}\zeta_i = 0$, moreover $\mathbb{E}\sigma_i^2 = \|\phi(x_i)\|^2\phi(x_i)\otimes\phi(x_i)$. Finally, given $x_1,\ldots,x_n$, we have that ζ_i is independent from ζ_j, when $i \neq j$. Then when $i \neq j$ we have $\mathrm{Tr}(\mathbb{E}\zeta_i\zeta_j) = \mathrm{Tr}((\mathbb{E}_{\sigma_i}\zeta_i)(\mathbb{E}_{\sigma_j}\zeta_j)) = 0$. When $i = j$ we have $\mathrm{Tr}(\mathbb{E}\zeta_i\zeta_j) = \mathrm{Tr}(\mathbb{E}\zeta_i^2) = \|\phi(x_i)\|^4$. So

$$\frac{1}{n^2}\sum_{i,j=1}^n \mathrm{Tr}(\mathbb{E}\zeta_i\zeta_j) = \frac{1}{n^2}\sum_{i=1}^n \|\phi(x_i)\|^4.$$

From which we obtain the desired result.

Now we are ready to bound R_n as follows

Proof Thm. 5. The proof is obtained by applying Thm. 12 and considering that $\|\phi(x)\|$ is uniformly bounded by c on $\mathcal{X}$. □

B.9 Proof of Proposition 4

See Appendix A for the basic technical assumptions on $\mathcal{X}$, $\mathcal{H}$ and ϕ. In particular $\mathcal{X}$ is Polish and ϕ is continuous and uniformly bounded by a constant c.

Proof of Proposition 4. In the following we will consider integrability and measurability with respect to a measure dx on $\mathcal{X}$. In particular $p : \mathcal{X} \to \mathbb{R}$ is an integrable function on $\mathcal{X}$ with respect to the measure dx. Now define $\Psi(x) = p(x)\phi(x)\phi(x)^\top$. We have that Ψ is measurable, since ϕ and p are measurable. Since p is integrable, p is finite almost everywhere, and hence $\Psi(x) = p(x)\phi(x)\phi(x)^\top$ is defined and trace class almost everywhere, and satisfies

$$\|\Psi(x)\|_\star = |p(x)| \, \|\phi(x)\|^2_{\mathcal{H}} \leqslant |p(x)|c^2 \text{ almost everywhere.}$$

Since the space of trace class operators is separable, this shows that Ψ is Bochner integrable and thus that the operator $W_p = \int_{x\in\mathcal{X}} \phi(x)\phi(x)^\top p(x)dx$ is well defined and trace class, with trace norm bounded by $\kappa^2\|p\|_{L^1(\mathcal{X})}$. Moreover, by linearity of the integral, for any $A \in \mathcal{S}(\mathcal{H})$,

$$\mathrm{Tr}(AW_p) = \int_{\mathcal{X}} \mathrm{Tr}(A\phi(x)\phi(x)^\top)p(x)dx = \int_{\mathcal{X}} f_A(x)p(x)dx,$$

where the last equality follows from the definition of f_A and the fact that

$$\mathrm{Tr}(A\phi(x)\phi(x)^\top) = \mathrm{Tr}(\phi(x)^\top A\phi(x)) = \phi(x)^\top A\phi(x) = f_A(x).$$

Remark 7 (Extension to more general linear functionals.)**.** *Note that the linearity of the model in A allows to generalize very easily the construction above to any linear functional that we want to apply to the model. This is especially true when the model has a finite dimensional representation as Eq. (7), i.e. $f_{\mathbf{B}} = \sum_{ij=1}^n \mathbf{B}_{i,j}k(x,x_i)k(x,x_j)$ with $\mathbf{B} \succeq 0$. In this case, given a linear functional $\mathcal{L} : C(\mathcal{X}) \to \mathbb{R}$, we have*

$$\mathcal{L}(f_{\mathbf{B}}) = \sum_{i,j=1}^n \mathbf{B}_{i,j}\mathcal{L}(k(x,x_i)k(x,x_j)) = \mathrm{Tr}(\mathbf{B}\mathbf{W}_{\mathcal{L}}),$$

where $(\mathbf{W}_{\mathcal{L}})_{i,j} = \mathcal{L}(k(x,x_i)k(x,x_j))$ for $i,j = 1,\ldots,n$.

B.10 Proof of Proposition 5

In Appendix B.10 and Appendix B.11, we will use the following notations.

Let $h, p \in \mathbb{N}$ and $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2$ be separable Hilbert spaces.

- $A = (A_s)_{1\leqslant s\leqslant p} \in \mathcal{S}(\mathcal{H})^p$ will denote a family of self-adjoint operators;
- Given a feature map $\phi : \mathcal{X} \to \mathcal{H}$ and $A = (A_s)_{1\leqslant s\leqslant p} \in \mathcal{S}(\mathcal{H})^p$ we will define the function f_A as follows

$$\forall x \in \mathcal{X},\ f_A(x) = (f_{A_s}(x))_{1\leqslant s\leqslant p} = \left(\phi(x)^\top A_s\phi(x)\right)_{1\leqslant s\leqslant p} \in \mathbb{R}^p, \qquad f_A : \mathcal{X} \to \mathbb{R}^p$$

- Given a matrix $C \in \mathbb{R}^{p\times h}$ which corresponds to a list of column vectors $(c^t)_{1\leqslant t\leqslant h} \in (\mathbb{R}^p)^h$, we define

$$K^C(\mathcal{H}) := \left\{ A = (A_s)_{1\leqslant s\leqslant p} \in \mathcal{S}(\mathcal{H})^p \ : \ \sum_{s=1}^p c_s^t A_s \succeq 0,\ 1 \leqslant t \leqslant h \right\}$$

- For any $A = (A_s)_{1 \leqslant s \leqslant p} \in \mathcal{S}(\mathcal{H}_1)^p$ and any bounded linear operator $L : \mathcal{H}_1 \to \mathcal{H}_2$, LAL^* will be a slight abuse of notation to denote the family $(LA_sL^*)_{1 \leqslant s \leqslant p} \in \mathcal{S}(\mathcal{H}_2)^p$.

Proof of Proposition 5. Let $p, h \in \mathbb{N}$ and let $C \in \mathbb{R}^{p \times h}$ be a matrix representing the column vectors $c^1 ... c^h$.
Let $\mathcal{Y}$ be the polyhedral cone defined by C, i.e. $\mathcal{Y} = \left\{ y \in \mathbb{R}^p \ : \ C^\top y \geqslant 0 \right\}$.
Let $\mathcal{H}$ be a separable Hilbert space and $\phi : \mathcal{X} \to \mathcal{H}$ be a fixed feature map.
With our previous notations, our goal is to prove that for any $A = (A_s)_{1 \leqslant s \leqslant p} \in \mathcal{S}(\mathcal{H})^p$,

$$A \in K^C(\mathcal{H}) \implies \forall x \in \mathcal{X},\ f_A(x) \in \mathcal{Y}.$$

Assume $A \in K^C(\mathcal{H})$ and let $x \in \mathcal{X}$. By definition, $f_A(x) = (\phi(x)^\top A_s \phi(x))_{1 \leqslant s \leqslant p} \in \mathbb{R}^p$. Hence,

$$C^\top f_A(x) = \left(\sum_{s=1}^p c_s^t \phi(x)^\top A_s \phi(x) \right)_{1 \leqslant t \leqslant h} = \left(\phi(x)^\top \left(\sum_{s=1}^p c_s^t A_s \right) \phi(x) \right)_{1 \leqslant t \leqslant h}.$$

Since $A \in K^C(\mathcal{H})$, for all $1 \leqslant t \leqslant h$, it holds $\sum_{s=1}^p c_s^t A_s \succeq 0$. In particular, this implies $\phi(x)^\top \sum_{s=1}^p c_s^t A_s \phi(x) \geqslant 0$ for all $1 \leqslant t \leqslant h$. Hence

$$C^\top f_A(x) \geqslant 0 \implies f_A(x) \in \mathcal{Y}.$$

B.11 Proof of Thm. 6

Using the notations of the previous section, the goal of this section is to solve a problem of the form

$$\inf_{A \in K^C(\mathcal{H})} L(f_A(x_1), ..., f_A(x_n)) + \Omega(A), \tag{14}$$

for given $p, h \in \mathbb{N}$, $C \in \mathbb{R}^{p \times h}$, separable Hilbert space $\mathcal{H}$, feature map $\phi : \mathcal{X} \to \mathcal{H}$, regularizer Ω, loss function $L : \mathbb{R}^n \to \mathbb{R} \cup +\infty$ and $x_1, ..., x_n \in \mathcal{X}$.

We start by stating the form of the regularizers we will be using.

Assumption 2. *Let $p \in \mathbb{N}$. For any separable Hilbert space $\mathcal{H}$ and any $A = (A_s)_{1 \leqslant s \leqslant p} \in \mathcal{S}(\mathcal{H})^p$, Ω is of the form*

$$\Omega(A) = \sum_{s=1}^p \Omega_s(A_s), \qquad \Omega_s(A_s) = \lambda_{s,1} \|A_s\|_\star + \frac{\lambda_{s,2}}{2} \|A_s\|_F^2,$$

where $\lambda_{s,1}, \lambda_{s,2} \geqslant 0$ and $\lambda_{s,1} + \lambda_{s,2} > 0$.

Lemma 6 (Properties of Ω)**.** *Let Ω be a regularizer such that Ω satisfies Assumption 2. Then Ω satisfies the following properties.*

(i) For any separable Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and any linear isometry $O : \mathcal{H}_1 \to \mathcal{H}_2$, i.e., such that $O^*O = I_{\mathcal{H}_1}$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H}_1)^p,\ \Omega(OAO^*) = \Omega(A).$$

(ii) For any separable Hilbert space $\mathcal{H}$ and any orthogonal projection $\Pi \in \mathcal{S}(\mathcal{H}_1)$, i.e. satisfying $\Pi = \Pi^*$, $\Pi^2 = \Pi$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H})^p,\ \Omega(\Pi A \Pi) \leqslant \Omega(A).$$

(iii) For any finite dimensional Hilbert space $\mathcal{H}_n$, taking $||A_s||_{op}$ to be the operator norm on $\mathcal{H}_n$,

$$\Omega \text{ is continuous}, \qquad \Omega(A) \underset{\sup_s ||A_s||_{op} \to +\infty}{\longrightarrow} +\infty$$

Proof. Note that since

$$\Omega(A) = \sum_{s=1}^{p} \Omega_s(A_s), \qquad \Omega_s(A_s) = \lambda_{s,1}\|A_s\|_\star + \frac{\lambda_{s,2}}{2}\|A_s\|_F^2,$$

where $\lambda_{s,1}, \lambda_{s,2} \geqslant 0$ and $\lambda_{s,1} + \lambda_{s,2} > 0$, it is actually sufficient to prove the following result. Let $\lambda_1, \lambda_2 \geqslant 0$ and assume $\lambda_1 + \lambda_2 > 0$. Let for any $A \in \mathcal{S}(\mathcal{H})$, $\Omega(A) = \lambda_1\|A\|_\star + \frac{\lambda_2}{2}\|A\|_F^2$. Then the following hold:

(i) For any separable Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and any linear isometry $O : \mathcal{H}_1 \to \mathcal{H}_2$, i.e., such that $O^*O = I_{\mathcal{H}_1}$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H}_1)^p,\ \Omega(OAO^*) = \Omega(A).$$

(ii) For any separable Hilbert space $\mathcal{H}$ and any orthogonal projection $\Pi \in \mathcal{S}(\mathcal{H}_1)$, i.e. satisfying $\Pi = \Pi^*,\ \Pi^2 = \Pi$, it holds

$$\forall A \in \mathcal{S}(\mathcal{H})^p,\ \Omega(\Pi A \Pi) \leqslant \Omega(A).$$

(iii) For any finite dimensional Hilbert space $\mathcal{H}_n$,

$$\Omega \text{ is continuous}, \qquad \Omega(A) \underset{\|A\|_{op} \to +\infty}{\longrightarrow} +\infty,$$

where we denote by $\|\cdot\|_{op}$ the operatorial norm.

1. (i) has already been proven in Lemma 2.

2. Let us prove (ii). Let $\mathcal{H}$ be a separable Hilbert space, Π an orthogonal projection on $\mathcal{H}$ and $A \in \mathcal{S}(\mathcal{H})$.

Using the fact that $\|B\|_\star = \sup_{\|C\|_{op} \leqslant 1} \mathrm{Tr}(BC)$, where $\|C\|_{op}$ denotes the operator norm on $\mathcal{S}(\mathcal{H})$, we have by property of the trace

$$\|\Pi A \Pi\|_\star = \sup_{\|C\|_{op} \leqslant 1} \mathrm{Tr}(\Pi A \Pi C) = \sup_{\|C\|_{op} \leqslant 1} \mathrm{Tr}(A(\Pi C \Pi)).$$

Now since $\|\Pi C \Pi\|_{op} \leqslant \|C\|_{op} \leqslant 1$, it holds $\sup_{\|C\|_{op} \leqslant 1} \mathrm{Tr}(A(\Pi C \Pi)) \leqslant \sup_{\|C\|_{op} \leqslant 1} \mathrm{Tr}(AC) = \|A\|_\star$. Thus:

$$\|\Pi A \Pi\|_\star \leqslant \|A\|_\star.$$

Moreover, since $\Pi \preceq I$, it holds $\Pi A \Pi A \Pi \preceq \Pi A^2 \Pi$. Hence,

$$\|\Pi A \Pi\|_F^2 = \mathrm{Tr}(\Pi A \Pi \Pi A \Pi) \leqslant \mathrm{Tr}(\Pi A^2 \Pi)$$

Now using the fact that $\mathrm{Tr}(\Pi A^2 \Pi) = \mathrm{Tr}(A \Pi A)$, we can once again use the fact that $\Pi \preceq I$ to show that $A \Pi A \preceq A^2$ and hence $\mathrm{Tr}(A \Pi A) \leqslant \mathrm{Tr}(A^2)$. Putting things together, we have shown

$$\mathrm{Tr}(\Pi A \Pi \Pi A \Pi) \leqslant \mathrm{Tr}(A^2) \implies \|\Pi A \Pi\|_F^2 \leqslant \|A\|_F^2.$$

Thus, by summing the inequalities, $\Omega(\Pi A \Pi) \leqslant \Omega(A)$.

3. The proof of (iii) is straightforward. The continuity of Ω comes from the fact that it is a norm on any finite dimensional Hilbert space. Moreover, since $\lambda_1 > 0$ or $\lambda_2 > 0$, Ω goes to infinity.

Remark 8. *As in the previous sections, the fact that Ω satisfies these three properties is actually sufficient to complete the proof.*

Recall that $\mathcal{H}_n$ is the finite dimensional subset of $\mathcal{H}$ spanned by the $\phi(x_i)$. Recall that Π_n is the orthogonal projection on $\mathcal{H}_n$, i.e.

$$\Pi_n \in \mathcal{S}(\mathcal{H}),\ \Pi_n^2 = \Pi_n,\ \mathrm{range}(\Pi_n) = \mathcal{H}_n.$$

Define $K_n^C(\mathcal{H})$ to be the following subspace of $K^C(\mathcal{H})$:

$$K_n^C(\mathcal{H}) := \left\{\Pi_n A \Pi_n \ :\ A \in K^C(\mathcal{H})\right\}.$$

It is straightforward to show that $K_n^C(\mathcal{H}) \subset K^C(\mathcal{H})$ since projecting left and right preserves the linear inequalities.

Proposition 9. *Let L be a lower semi-continuous function which is bounded below, and assume Ω satisfies Assumption 2. Then Eq. (14) has a solution A^* which is in $K_n^C(\mathcal{H})$.*

Proof. In this proof, denote with J the function defined by

$$\forall A \in \mathcal{S}(\mathcal{H})^p,\ J(A) := L(f_A(x_1), ..., f_A(x_n)) + \Omega(A).$$

Our goal is to prove that the problem $\inf_{A \in K^C(\mathcal{H})} J(A)$ has a solution which is in $K_n^C(\mathcal{H})$, i.e. of the form $\Pi_n A \Pi_n$ for some $A \in K_n^C(\mathcal{H})$.

1. Let us start by fixing $A \in K^C(\mathcal{H})$.
First note that since Π_n is the orthogonal projection on $\mathrm{span}(\phi(x_i))_{1 \leqslant i \leqslant n}$, in particular $\Pi_n \phi(x_i) = \phi(x_i)$ for all $1 \leqslant i \leqslant n$. Thus, for any $1 \leqslant i \leqslant n$,

$$f_A(x_i) = (\phi(x_i)^\top A_s \phi(x_i))_{1 \leqslant s \leqslant p} = (\phi(x_i)^\top \Pi_n A_s \Pi_n \phi(x_i))_{1 \leqslant s \leqslant p} = f_{\Pi_n A \Pi_n}(x_i).$$

Here, the first and last equalities come from the definition of f_A and $f_{\Pi_n A \Pi_n}$. Thus,

$$J(A) = L(f_{\Pi_n A \Pi_n}(x_1), ..., f_{\Pi_n A \Pi_n}(x_n)) + \Omega(A).$$

Now since Ω satisfies Assumption 2, by the second point of Lemma 6, it holds $\Omega(\Pi_n A \Pi_n) \leqslant \Omega(A)$, hence

$$J(\Pi_n A \Pi_n) \leqslant J(A).$$

This last inequality combined with the fact that $K_n^C(\mathcal{H}) = \left\{\Pi_n A \Pi_n \ :\ A \in K^C(\mathcal{H})\right\} \subset K^C(\mathcal{H})$ show that

$$\inf_{A \in K_n^C(\mathcal{H})} J(A) = \inf_{K^C(\mathcal{H})} J(A).$$

2. Let us now show that $\inf_{A\in K_n^C(\mathcal{H})} J(A)$ has a solution. Let us exclude the case where $J=+\infty$, in which case $A=0$ can be taken to be a solution.

Let V_n be the injection $V_n : \mathcal{H}_n \hookrightarrow \mathcal{H}$. Note that $V_nV_n^* = \Pi_n$ and $V_n^*V_n = I_{\mathcal{H}_n}$. These simple facts easily show that

$$K_n^C(\mathcal{H}) = V_nK^C(\mathcal{H}_n)V_n^* = \left\{V_n\tilde{A}V_n^* \ : \ \tilde{A}\in K_n^C(\mathcal{H}_n)\right\}.$$

Thus, our goal is to show that $\inf_{\tilde{A}\in K_n^C(\mathcal{H}_n)} J(V_nAV_n^*)$ has a solution.

By the first point of Lemma 6, since $V_n^*V_n = I_{\mathcal{H}_n}$, it holds

$$\forall \tilde{A}\in\mathcal{S}(\mathcal{H}_n),\ \Omega(V_n\tilde{A}V_n^*)=\Omega(\tilde{A}) \implies J(V_n\tilde{A}V_n^*) = L(f_{V_n\tilde{A}V_n^*}(x_1),...,f_{V_n\tilde{A}V_n^*}(x_n))+\Omega(\tilde{A}).$$

Let $\tilde{A}_0\in K^C(\mathcal{H}_n)$ be a point such that $J_0 := J(V_n\tilde{A}_0V_n^*)<\infty$. Let c_0 be a lower bound for L. By the third point of Lemma 6, there exists a radius R_0 such that for all $\tilde{A}\in\mathcal{S}(\mathcal{H}_n)$,

$$\|\tilde{A}\|_F > R_0 \implies \Omega(\tilde{A}) > J_0 - c_0.$$

Since c_0 is a lower bound for L, this implies

$$\inf_{\tilde{A}\in K^C(\mathcal{H}_n)} J(V_n\tilde{A}V_n^*) = \inf_{\tilde{A}\in K^C(\mathcal{H}_n),\ \|\tilde{A}\|_F\leqslant R_0} J(V_n\tilde{A}V_n^*).$$

Now since L is lower semi-continuous, Ω is continuous by the last point of Lemma 6, and $\tilde{A}\mapsto (f_{V_n\tilde{A}V_n^*}(x_i))_{1\leqslant i\leqslant n}$ is linear hence continuous, the mapping $A\mapsto J(V_n\tilde{A}V_n^*)$ is lower semi-continuous. Hence, it reaches its minimum on any non empty compact set. Since $\mathcal{H}_n$ is finite dimensional, the set $\left\{\tilde{A}\in K^C(\mathcal{H}_n) \ : \ \|\tilde{A}\|_F\leqslant R_0\right\}$ is compact (closed and bounded) and non empty (it contains $\tilde{A}_0$), and hence there exists $\tilde{A}_*\in K_n^C(\mathcal{H})$ such that $J(V_n\tilde{A}_*V_n^*) = \inf_{\tilde{A}\in K^C(\mathcal{H}_n),\ \|\tilde{A}\|_F\leqslant R_0} J(V_n\tilde{A}V_n^*)$. Going back up the previous equalities, this shows that $A_* := V_n\tilde{A}_*V_n^*\in K_n^C(\mathcal{H})$ and $J(A_*)=\inf_{A\succeq 0} J(A)$.

Lemma 7. *The set $K_n^C(\mathcal{H})$ can be represented in the following way*

$$K_n^C(\mathcal{H}) = \left\{(S_n^*\mathbf{B}_sS_n)_{1\leqslant s\leqslant p}\in\mathcal{S}(\mathcal{H})^p \ : \ \mathbf{B}=(\mathbf{B}_s)_{1\leqslant s\leqslant p}\in K^C(\mathbb{R}^n)\right\}$$

In particular, for any $A\in K_n^C(\mathcal{H})$, there exists p symmetric matrices $\mathbf{B}=(\mathbf{B}_s)_{1\leqslant s\leqslant p}\in K^C(\mathbb{R}^n)$ such that

$$\forall x\in\mathcal{X},\ f_A(x) = \left(\sum_{1\leqslant i,j\leqslant n}[\mathbf{B}_s]_{i,j}k(x_i,x)k(x_j,x)\right)_{1\leqslant s\leqslant p}.$$

Proof. The proof is exactly analoguous to the proof of Lemma 3.

We will prove the following Thm. 13 which statement is that of Thm. 6 with more precise assumptions.

Theorem 13. *Let L be lower semi-continuous and bounded below, and Ω satisfying Assumption 2. Then Eq. (5) has a solution of the form*

$$f_*(x) = \left(\textstyle\sum_{i,j=1}^n[\mathbf{B}_s]_{i,j}k(x,x_i)k(x,x_j)\right)_{1\leqslant s\leqslant p},$$

for some family $\mathbf{B}=(\mathbf{B}_s)_{1\leqslant s\leqslant p}\in K^C(\mathbb{R}^n)$. Moreover, if L is convex, this solution is unique.

Proof of Thm. 13. The proof is completely analoguous to that of Thm. 7, combining Lemma 7 and Proposition 9.

C Additional proofs

Lemma 8. *Let $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a compact set with Lipschitz boundary. Let $m > d/2$. Then $W_2^m(\mathcal{X})$ is a multiplication algebra (see Definition 3).*

Proof. When $m \in \mathbb{N}$ and $m > d/2$, then $W_2^m(\mathbb{R}^d)$ is a multiplication algebra [AF03]. When $m \notin \mathbb{N}$, by Eq. 2.69 pag. 138 of [Tri06] we have that $F_{2,2}^m(\mathbb{R}^d)$ is a multiplication algebra when $m > d/2$, where $F_{2,2}^m$ is the Triebel-Lizorkin space of smoothness m and order $2, 2$ and corresponds to $W_2^m(\mathbb{R}^d)$, i.e., $F_{2,2}^m(\mathbb{R}^d) = W_2^m(\mathbb{R}^d)$ [Tri06].

So far we have that $m > d/2$ implies that $W_2^m(\mathbb{R}^d)$ is a multiplication algebra, now we extend this result to $W_2^m(\mathcal{X})$. Note that since $\mathcal{X}$ is compact and with Lipschitz boundary, for any $f \in W_2^m(\mathcal{X})$ there exists an *extension* $\tilde{f} \in W_2^m(\mathbb{R}^d)$ such that $\tilde{f}|_{\mathcal{X}} = f$ and $\|\tilde{f}\|_{W_2^m(\mathbb{R}^d)} \leqslant C_1 \|f\|_{W_2^m(\mathcal{X})}$ with C_1 depending only on $m, d, \mathcal{X}$ (see Thm. 5.24 pag. 154 for $m \in \mathbb{N}$ and 7.69 when $m \notin \mathbb{N}$ pag 256 [AF03]). Then, since for any $f : \mathbb{R}^d \to \mathbb{R}$, by construction we have $\|f|_X\|_{W^m(\mathcal{X})} \leqslant \|f\|_{W^m(\mathbb{R}^d)}$ [AF03]. Then, for any $f, g \in W_2^m(\mathcal{X})$, denoting by $\tilde{f}, \tilde{g}$ the extensions of f, g, we have

$$\|f \cdot g\|_{W_2^m(\mathcal{X})} = \|\tilde{f}|_{\mathcal{X}} \cdot \tilde{g}|_{\mathcal{X}}\|_{W_2^m(\mathcal{X})} \leqslant \|\tilde{f} \cdot \tilde{g}\|_{W_2^m(\mathbb{R}^d)} \tag{30}$$

$$\leqslant C\|\tilde{f}\|_{W_2^m(\mathbb{R}^d)}\|\tilde{g}\|_{W_2^m(\mathbb{R}^d)} \leqslant CC_1^2\|f\|_{W_2^m(\mathcal{X})}\|g\|_{W_2^m(\mathcal{X})}. \tag{31}$$

To conclude $u : \mathcal{X} \to \mathbb{R}$ that maps $x \mapsto 1$ has bounded norm corresponding to $\|u\|^2_{W_2^m(\mathcal{X})} = \int_{\mathcal{X}} dx$. So $W_2^m(\mathcal{X})$ when $m > d/2$ and $\mathcal{X}$ is compact with Lipschitz boundary is a multiplication algebra.

D Additional details on the other models

Recall that the goal is to solve a problem of the form Eq. (1), i.e.

$$\min_{f \in \mathcal{F}} L(f(x_1), ..., f(x_n)) + \Omega(f).$$

In this section, $\phi : \mathcal{X} \to \mathcal{H}$ will always denote a feature map, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive semi definite kernel on $\mathcal{X}$ ($k(x, x') = \phi(x)^\top \phi(x')$ if k is the positive semi-definite kernel associated to ϕ). Given a kernel k, $\mathbf{K} \in \mathbb{R}^{n \times n}$ will always denote the positive semi-definite kernel matrix with coefficients $\mathbf{K}_{i,j} = k(x_i, x_j),\ 1 \leqslant i, j \leqslant n$.

Generalized linear models (GLM). Consider generalized linear models of the form, $f_w(x) = \psi(w^\top \phi(x))$. Assume the regularizer is of the form $\Omega(f_w) = \frac{\lambda}{2}\|w\|^2$. Using the representer theorem [CL09], any solution to Eq. (1) is of the form $w = \sum_{i=1}^n \alpha_i \phi(x_i)$ and thus Eq. (1) becomes the following finite dimensional problem in α:

$$\min_{\alpha \in \mathbb{R}^n} L(\psi(\mathbf{K}\alpha)) + \frac{\lambda}{2}\alpha^\top \mathbf{K}\alpha. \tag{32}$$

In the case where one wishes to learn a density function with respect to a basis measure ν, a common choice of model is functions of the form

$$p_\alpha(x) = \frac{\exp(g(x))}{\int_{\tilde{x} \in \mathcal{X}} \exp(g(\tilde{x}))d\nu(\tilde{x})}, \qquad g(x) = \sum_{i=1}^n \alpha_i k(x_i, x).$$

where k is a positive semi-definite kernel on $\mathcal{X}$. The prototypical problem one solves to find the best p_α is

$$\min_{\alpha \in \mathbb{R}^n} L(p_\alpha(x_1), ..., p_\alpha(x_n)) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{33}$$

In the specific case where the loss function is the negative log likelihood $L(z_1, ..., z_n) = \frac{1}{n} \sum_{i=1}^n -\log(z_i)$, it can be shown that Eq. (33) is convex in α.

In practice, we solve Eq. (32) by applying standard gradient descent with restarts, as the problem is non convex.

To solve Eq. (33), since the problem is convex, the algorithm is guaranteed to converge. However, since we can only estimate the quantity $\int_{\tilde{x} \in \mathcal{X}} \exp(g(\tilde{x})) d\nu(\tilde{x})$; we do so by taking a measure ν from which we can sample. However, this becomes intractable as the dimension grows, as the experiments on density estimation will put into light.

Non-negative coefficients models (NCM). Recall the definition of an NCM. It represent non-negative functions as $f_\alpha(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$, with $\alpha_1, \dots \alpha_n \geqslant 0$, given a kernel $k(x, x') \geqslant 0$ for any $x, x' \in \mathcal{X}$. In this case, the prototypical problem is of the form :

$$\min_{\alpha \geqslant 0} L(\mathbf{K}\alpha) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{34}$$

If we are performing density estimation with respect to the measure ν, one wishes to impose $\int_{\mathcal{X}} f_\alpha(x) d\nu(x) = 1$, which can be seen as an affine constraint over α, since

$$\int_{\mathcal{X}} f_\alpha(x) d\nu(x) = \mathbf{u}^\top \alpha, \qquad \mathbf{u} = \left(\int_{\mathcal{X}} k(x, x_i) d\nu(x) \right)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n.$$

In this case, the prototypical problem will be of the form

$$\min_{\substack{\alpha \geqslant 0 \\ \mathbf{u}^\top \alpha = 1}} L(\mathbf{K}\alpha) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha. \tag{35}$$

If L is a convex smooth function, both problems Eq. (34) and Eq. (35) can be solved using projected gradient descent, since the projections on the set $\alpha \geqslant 0$ and the simplex

$$\left\{ \alpha \in \mathbb{R}^n \ : \ \alpha \geqslant 0, \ \mathbf{u}^\top \alpha = 1 \right\}$$

can be computed in closed form.

In the main paper, we mention that NCM models do not satisfy **P2** i.e. that they cannot approximate any function arbitrarily well. We implement Example 2 in the following way. Let $g(x) = e^{-\|x\|^2/2}$. Take $k(x, x') = e^{-\|x - x'\|^2}$, n points $(x_1, ..., x_n)$ taken uniformly in the interval $[-5, 5]$. To find the function f_α which best approximates g, we perform least squares regression, i.e. solve the prototypical problem Eq. (34) with the square loss function

$$L(y) = \frac{1}{2n} \sum_{i=1}^n |y_i - g(x_i)|^2.$$

We perform cross validation to select the value of λ for each value of n. In Fig. 2, we show the obtained function f_α for $n = 100, 1000, 10000$. This clearly illustrates that with this model, we cannot approximate g in a good way, no matter how many points n we have.

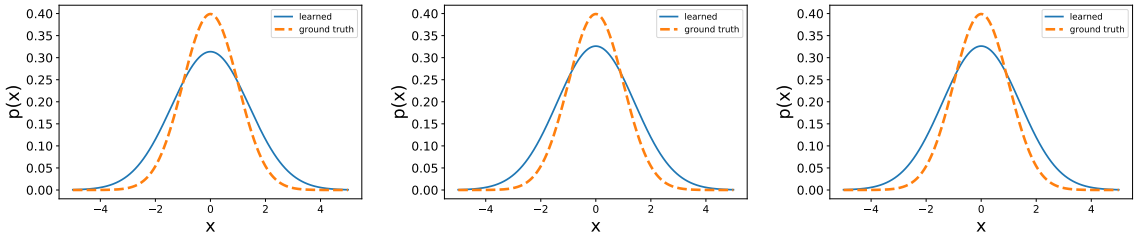


Figure 2: Best approximation of g using NCM with *(left)* $n = 100$ *(center)* $n = 1000$ *(right)* $n = 10000$ points.

Partially non-negative linear models (PNM). Consider partially non negative models of the form $f_w(x) = w^\top \phi(x)$, with $w \in \{w \in \mathcal{H} \mid w^\top \phi(x_1) \geqslant 0, \dots, w^\top \phi(x_n) \geqslant 0\}$ (that is we impose $f_w(x_i) \geqslant 0$). Take Ω to be of the form $\frac{\lambda}{2}\|w\|^2$ in Eq. (1). Using the representer theorem in [CL09], we can show that there is a solution of this problem of the form $f_\alpha = \sum_{i=1}^n \alpha_i k(x, x_i)$, leading to the following optimization problem in α to recover the optimal solution:

$$\min_{\mathbf{K}\alpha \geqslant 0} L(\mathbf{K}\alpha) + \frac{\lambda}{2}\alpha^\top \mathbf{K}\alpha \tag{36}$$

If we want to impose that the resulting f_α sums to one for a given measure ν on $\mathcal{X}$, we proceed as in Eq. (35) and solve

$$\min_{\substack{\mathbf{K}\alpha \geqslant 0 \\ \mathbf{u}^\top \alpha = 1}} L(\mathbf{K}\alpha) + \frac{\lambda}{2}\alpha^\top \mathbf{K}\alpha. \tag{37}$$

However, there is no guarantee that the resulting f_α will be a density, as will be made clear in the next section on density estimation.

In the experiments, we solve Eq. (36) and Eq. (37) in the following way. We first compute a cholesky factor of $\mathbf{K}$: $\mathbf{K} = \mathbf{V}^\top \mathbf{V}$. Changing variables by setting $\mathbf{V}\beta = \alpha$, the objective functions become strongly convex in β. We then compute the dual of these problems and apply a proximal algorithm like FISTA, since the proximal operator of L is always known in our experiments.

E Additional details on the experiments

In this section, we provide additional details on the experiments. The code will be available online. Recall that we consider four different models for functions with non-negative outputs : GLM, PNM, NCM and our model.

Kernels. All the models we consider depend on certain positive semi definite kernels k. In all the experiments, we have taken the kernels to be Gaussian kernels with width σ:

$$\forall x, x' \in \mathbb{R}^d,\ k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right).$$

Regularizers. For GLM, PNM and NCM, the regularizer for the underlying linear models are always of the form $\frac{\lambda}{2}\|w\|^2$ where w is the parameter of the linear model, which translates to $\frac{\lambda}{2}\alpha^\top \mathbf{K}\alpha$ where the α are the coefficients of the finite dimensional representation. For our model, we always take the regularizer to be of the form $\lambda\left(\|A\|_\star + 0.05\|A\|_F^2\right)$.

Parameter selection. In all experiments except for the one on density estimation in the main paper (in which we fix $\sigma = 1$ and select λ), we select the parameters σ of the kernels involved as well as the parameters λ for the regularizers using K fold cross validation with $K = 7$. This means that once the data set has been generated, we randomly divide it into two sets : the training set containing 70% of the data and the test set containing 30% of the data. We then train our model for the given σ, λ and report the performance on the test set. We repeat this operation $K = 7$ times and consider the mean performance on the test set to be a good indicator of the performance of our model for a given set of parameters. We then select the best parameters by doing a grid search. The code for this cross-validation will be available online.

Formulations and algorithms. The formulations of our three problems : density estimation, regression with Gaussian heteroscedastic errors, and multiple quantile regression, have been expressed in the main paper in a generic way involving functions with unconstrained outputs, and functions with outputs constrained to be non negative and sometimes summing to one. We always model functions with unconstrained outputs with a linear model with gaussian kernel, and model the functions with constrained outputs with the four models for non-negative functions we consider: ours, PNM, GLM and NCM.

In practice, we implement the methods PNM, GLM and NCM as explained in Appendix D. In particular, we use FISTA for PNM, and our model, dualizing the equality constraints for density estimation. This relies on the fact that the proximal operators of the log likelihood, the objective function for heteroscedastic regression as well as the pinball loss can be computed in closed form, and that the regularization is smooth in the right coordinates.

Details on the experiments of the main text. Here, we add a few precisions on the toy distributions we have used to sample data and the number of sampled used when not specified in the main text.

- For heteroscedastic regression, the data was generated as the toy data in section 5 of [LSC05], with $n = 80$ points.
- For quantile regression, the data points (x_i, y_i) were generated according to the following distribution for (X, Y) : $X \sim \frac{1}{2}U(0, 1/3) + \frac{1}{2}U(2/3, 1)$ and $Y|x \sim \mathcal{N}(0, \sigma(x))$ where

$$\sigma(x) = \begin{cases} -x + 1/3 & \text{for } 0 \leqslant x \leqslant 1/3 \\ x - 2/3 & \text{for } 2/3 \leqslant x \leqslant 1 \\ 0 & \text{otherwise .} \end{cases}$$

  Here, U stands for the uniform distribution. Moreover, in order to perform the experiments in the main paper, we have used 500 sample points.

Density estimation in dimension 10 with $n = 1000$. In this paragraph, we consider the following experiment. Let $d = 10$, $X \in \mathbb{R}^d$ be a random variable distributed as a mixture of Gaussians :

$$X \sim \frac{1}{2}\mathcal{N}(-2e_1, 1/\sqrt{2\pi}I_d) + \frac{1}{2}\mathcal{N}(2e_1, 1/\sqrt{2\pi}I_d)$$

where e_1 is the first vector of the canonical basis of $\mathbb{R}^d$.

Let $n = 1000$ and let $(x_1, ..., x_n)$ be n iid samples of X. We perform the four different methods, cross validating both the regularization parameter λ and the kernel parameter σ at each time. We learn the density in the form

$$p(x) = f(x)\nu(x), \qquad \nu \text{ is the density associated with } \mathcal{N}(0, 5I_d).$$

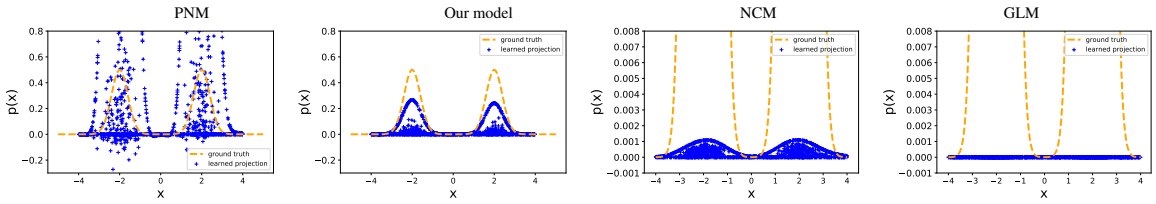


Figure 3: Representation of the densities learned by the different models.

We then use our models for densities to compute the best f in its class using the negative log-likelihood as a loss function. It is crucial that we can sample from ν in order to approximate the integral in the case of GLMs.

In order to visualize the results of the different algorithms in Fig. 3, we compute the learnt distribution p, and then sample randomly $n_0 = 500$ points from a uniform distribution on the box centered at 0 and of width 5 in order to explore regions where the density is close to zero, n_0 points sampled from the true distribution of the data, in order to explore points where the density is representative, and n_0 points on the line $[-4, 4] \times \{0\}^{d-1}$ where the density is at its highest. We then project onto the first coordinate, i.e. given a point $x = (x_i)_{1 \leqslant i \leqslant d}$ and the associated predicted density $p(x)$, we plot the point $(x_1, p(x))$. Note that for readability, we have used the same scale for our model and the PNM, and a smaller scale for the two others since the learnt density is much flatter.

Let us now analyse the results in Fig. 3. Note that in terms of performance, i.e. log likelihood on the test set, the first two models (PNM and our model) are quite close and are better than the two others.

- **PNM**. As in $d = 1$ we see that for $d = 10$ the problems of non-negativity for PNM are exacerbated, making it not suitable to learn a probability distribution. Indeed there are low density regions where the optimization problem pushes the model to be negative. Since by constraint we have $\int f d\nu = 1$, the volume of the negative regions is used to push up the function in the regions with high density. So $\int |f| d\nu \gg 1$, while it should be $\int |f| d\nu = 1$. This is confirmed by the behavior of the cross validation.

- **Our model** Our model seems to perform reasonably well.

- **NCM**. This problem is particularly difficult for NCM. Indeed, as the width of the kernel decreases, the model is unable to learn since it overfits in the direction e_1 and it would require way more points than $n = 1000$. However, as soon as the width of the kernel is good for e_1, the learnt distribution becomes too heavy tailed in the direction orthogonal to e_1.

- **GLM**. It is interesting to note that GLM completely fails, because the measure ν which we take as a reference measure has a support which has only double variance compared to p, but in 10 dimensions it corresponds to a support with way larger volume compared to the one of the target distribution. In particular, the estimation of the integral, which was possible in $d = 1$ with 10000 i.i.d. points from ν, in 10 dimensions becomes almost impossible (it would require way more sampling points). Note that we sample the points from ν to simulate the real-world situation where p is a measure from which it is difficult to sample from, while ν is an simple measure to sample from which contains the support of p. Further experiments show that if one takes the target distribution to sample, one obtains a good model, which reassures us in the fact that this is not a coding error but a real phenomenon.

F Relationship to [BF15]

As mentioned in the main paper, the model in Eq. (4) has already been considered in [BF15] with a similar goal as ours. This paper is a workshop publication that has only be lightly peer-reviewed and contains fundamental flaws. In particular, they provide an incorrect characterization of the solution of Eq. (5), that limits the representation power of the model to the one of non-negative coefficients models, that, as we have seen in Sec. 2.1 and in Example 2, has poor approximation properties and cannot be universal. This severe limitation affects also the optimization framework (which also only relies on general-purpose toolboxes such as CVX (`http://cvxr.com/cvx/`), which are not scalable to large n).

Indeed, in their main result, the representer theorem incorrectly characterizes A^* the solution of Eq. (5) as

$$A^* \in R_n \cap \mathcal{S}(\mathcal{H})_+, \qquad R_n = \left\{ \sum_{i=1}^n \alpha_i \phi(x_i) \otimes \phi(x_i) \mid \alpha \in \mathbb{R}^n \right\},$$

and $\mathcal{S}(\mathcal{H})_+ = \{A \in \mathcal{S}(\mathcal{H}) \mid A \succeq 0\}$. Note, however that $R_n \subseteq \mathcal{S}(\mathcal{H}_n) \subset \mathcal{S}(\mathcal{H})$ by construction, where $\mathcal{H}_n = \text{span}\{\phi(x_1), \dots, \phi(x_n)\}$. So their characterization corresponds to

$$A^* \in \left\{ A = \sum_{i=1}^n \alpha_i \phi(x_i) \otimes \phi(x_i) \mid \alpha \in \mathbb{R}^n, A \succeq 0 \right\}.$$

Now, for simplicity, consider the interesting case where ϕ is universal and $x_1, \dots, x_n$ are distinct points. Then $(\phi(x_i))_{i=1}^n$ forms a basis for $\mathcal{H}_n$ and the only $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ that guarantee $A \succeq 0$ are $\alpha_1 \geqslant 0, \dots, \alpha_n \geqslant 0$, i.e.,

$$R_n = \left\{ A = \sum_{i=1}^n \alpha_i \phi(x_i) \otimes \phi(x_i) \mid \alpha_1 \geqslant 0, \dots, \alpha_n \geqslant 0 \right\}.$$

Note that this class of operators leads only to *non-negative coefficients models*. Indeed, let $A \in R_n$ and denote by $k(x, x')$ the function $k(x, x') = (\phi(x)^\top \phi(x'))^2$, then

$$f_A(x) = \phi(x)^\top A \phi(x) = \sum_{i=1}^n \alpha_i (\phi(x)^\top \phi(x_i))^2 = \sum_{i=1}^n \alpha_i k(x, x_i), \quad \forall\, x \in \mathcal{X}.$$

Since k is a kernel (it is an integer power of $\phi(x)^\top \phi(x')$ that is a kernel [SS02]) and $\alpha_1 \geqslant 0, \dots, \alpha_n \geqslant 0$, then f_A belongs to the non-negative coefficients models.

Instead, we know by our Thm. 1 that $A^* \in \mathcal{S}(\mathcal{H}_n)_+$ and more explicitly, by Thm. 2 that A^*, the solution of Eq. (5) is characterized by the non-positive part operator of a symmetric matrix $[\cdot]_+$. By Thm. 3 we already know that our model is universal while NCM and thus the characterization in [BF15] cannot be universal.

References

[AF03] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Elsevier, 2003.

[BF15] J. Andrew Bagnell and Amir-massoud Farahmand. Learning positive functions in a Hilbert space. In *NIPS Workshop on Optimization (OPT2015)*, pages 3240–3255, 2015.

[BL10] Jonathan M. Borwein and Adrian S. Lewis. *Convex analysis and nonlinear optimization: theory and examples; 2nd ed.* CMS Books in Mathematics. Springer, Dordrecht, 2010.

[BM01] Haïm Brezis and Petru Mironescu. Gagliardo-nirenberg, composition and products in fractional sobolev spaces. 2001.

[BM02] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[BTA11] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

[BV04] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CL09] Elliott Ward Cheney and William Allan Light. *A Course in Approximation Theory*, volume 101. American Mathematical Soc., 2009.

[GGK04] Israel Gohberg, Seymour Goldberg, and Marinus A. Kaashoek. *Basic Classes of Linear Operators*. 2004.

[LSC05] Quoc V. Le, Alex J. Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496, 2005.

[MXZ06] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[PB14] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[SFL11] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[SS02] Bernhard Schölkopf and Alexander Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[Tri06] Hans Triebel. *Theory of Function Spaces III*, volume 100. Birkhäuser Basel, 2006.

[Wen04] Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.