Learning PSD-valued functions using kernel sums-of-squares

Boris Muzellec, Francis Bach, Alessandro Rudi

Learning PSD-Valued Functions Using Kernel Sums-of-Squares

Boris Muzellec BORIS.MUZELLEC@INRIA.FR
Francis Bach FRANCIS.BACH@INRIA.FR
Alessandro Rudi ALESSANDRO.RUDI@INRIA.FR
INRIA Paris, 2 rue Simone Iff, 75012, Paris, France
ENS - Département d'Informatique de l'École Normale Supérieure,
PSL Research University, 2 rue Simone Iff, 75012, Paris, France

Abstract

Shape constraints such as positive semi-definiteness (PSD) for matrices or convexity for functions play a central role in many applications in machine learning and sciences, including metric learning, optimal transport, and economics. Yet, very few function models exist that enforce PSD-ness or convexity with good empirical performance and theoretical guarantees. In this paper, we introduce a kernel sum-of-squares model for functions that take values in the PSD cone, which extends kernel sums-of-squares models that were recently proposed to encode non-negative scalar functions. We provide a representer theorem for this class of PSD functions, show that it constitutes a universal approximator of PSD functions, and derive eigenvalue bounds in the case of subsampled equality constraints. We then apply our results to modeling convex functions, by enforcing a kernel sum-of-squares representation of their Hessian, and show that any smooth and strongly convex function may be thus represented. Finally, we illustrate our methods on a PSD matrix-valued regression task, and on scalar-valued convex regression.

Keywords: Positive-definite matrices, kernels, sums of squares, convex functions

1. Introduction

Linear models, and kernel methods in particular, were proved over the past few decades to be of great effectiveness in machine learning, both in supervised and unsupervised learning (see, e.g., Schölkopf et al., 2002). Indeed, they offer an attractive trade-off between representation power, algorithmic simplicity, and theoretical guarantees. Moreover, they constitute a flexible toolbox that is not restricted to vector inputs or scalar outputs: years of kernel design have resulted in kernels that may be used to learn on a vast diversity of data types, or that may go beyond scalar functions and have vector outputs (Carmeli et al., 2010).

Yet, linear models have had limited applications to tasks in which functions must satisfy a given shape constraint, such as monotonicity or convexity, on their whole domain, and not only on the training points. A notable exception is non-negativity: recently, Marteau-Ferey et al. (2020) proposed a kernel sum-of-squares (SoS) model for smooth non-negative scalar functions with universal approximation properties, which was then applied to non-convex optimization by Rudi et al. (2020), to the estimation of optimal transport distances by Vacher et al. (2021), and to optimal control (Berthier et al., 2021). Further, Marteau-Ferey et al. (2020) propose extensions of their model to polyhedral cone constraints. However, some important classes of shape constraints may not be encoded as polyhedral cones: in particular,

neither the cone of positive semi-definite (PSD) matrices nor the cone of convex functions are polyhedral. Yet, they are at the heart of numerous applications in machine learning and sciences: PSD and convexity constraints play a central role in optimal transport (Brenier, 1991), monopolistic games in economics (Rochet and Choné, 1998; Mirebeau, 2016) and Newton's least resistance problem in physics (Lachand-Robert and Oudet, 2005).

Recently, approaches to handle convexity constraints have been proposed within the deep learning community (Amos et al., 2017), and were applied to optimal control (Chen et al., 2018) and optimal transport (Makkuva et al., 2020), among others. However, due to lack of results regarding their approximation properties, those model allow very limited – if any – theoretical guarantees when applied to those problems. Alternatively, Aubin-Frankowski and Szabo (2021) proposed a framework to enforce shape constraints in RKHSs, which comes with performance guarantees. Their method relies on a tightening of cone constraints through compact coverings in vector-valued RKHSs (vRKHSs), i.e. they apply a finite number of SDP inequalities involving the Hessian $\mathbf{H}_f$ of the learnt function f, of the type

$$\eta_m \|f\|_{\mathcal{H}} \mathbf{I}_P \preceq \operatorname{diag}(\mathbf{b}) + \mathbf{H}_f(\tilde{x}_m), \quad m \in [M],$$

with suitable $\eta_m > 0$ and $\tilde{x}_m \in \mathcal{X}$. This allows applying various shape constraints for scalar or vector-valued functions, such as convexity or monotonicity.

In comparison, we propose in this paper a linear model for matrix-valued functions that is guaranteed to take PSD values. This model may be seen as an extension of the sums-of-squares model of Marteau-Ferey et al. (2020) to a "sums-of-outer-products" model that allows representing smooth functions taking values in the (non-polyhedral) PSD cone. Further, our model inherits from the interpolation inequalities on constraints violations which where shown in the scalar case by Rudi et al. (2020). Finally, while modeling PSD-valued functions may be of interest *per se*, our model may also be used to apply convexity (or concavity) constraints on scalar-valued functions by imposing *equality* constraints on the Hessian (compared to the inequality constraints proposed by Aubin-Frankowski and Szabo (2021)), which again allows leveraging interpolation inequalities.

Contributions. In the first part of this work, we propose a linear model for smooth PSD-valued functions that extends the SoS model of Marteau-Ferey et al. (2020), further studied by Rudi et al. (2020). More precisely:

1. We introduce a linear sum-of-squares (SoS) model for functions whose values are PSD matrices (Proposition 1). This model is linearly parameterized by a PSD operator in a Hilbert space, and is defined in terms of an arbitrary vector-valued feature map. In the main body of this work, this feature map is based on the canonical feature map of a scalar-valued RKHS ; hence our results do not require prior knowledge of vector-valued RKHSs. Then, for the sake of completeness, we introduce vector-valued RKHSs (vRKHSs) in Appendix A, and provide proofs and statements of our results in the more general setting of vRKHSs in Appendices B and C.

2. We provide a representer theorem (Theorem 2), that generalizes that of Marteau-Ferey et al. (2020) along with a finite-dimensional formulation for our model that only depends on training points, and derive the dual form of generic convex PSD-valued function learning problems (Theorem 3).

3. We show in Theorem 4 that the class of function we introduce is a universal approximator for PSD-valued functions, which justifies its use in a wide range of problems involving PSD constraints.

4. Extending the results of Rudi et al. (2020) to the PSD setting, we provide scattered data inequalities for the eigenvalues of the PSD matrices output by PSD-valued kernel SoS functions in Theorem 5. These bounds may be used to quantify constraint violations in the case where our model is used to subsample PSD constraints.

In the second part of this paper, we apply our model to the problem of fitting a function under convexity constraints:

5. In Theorem 6, we show that the Hessian of any strongly convex function may be represented as a kernel SoS, which allows handling learning problems with convexity constraints using SoS equality constraints.

6. Leveraging Theorem 5, we bound in Corollary 7 the strong convexity constant of functions learned with subsampled SoS constraints, and bound the error induced by subsampling PSD constraints when optimizing a Lipschitz functional over smooth convex functions.

7. Finally, we illustrate the SoS model on a convex regression task.

2. Background and Setting

We start by stating notation and recalling background material on reproducing kernel Hilbert spaces and positive-semidefinite operators.

2.1 Notation

For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. For two Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$ $\mathcal{L}(\mathcal{H}, \mathcal{G})$ denotes the set of bounded linear maps from $\mathcal{H}$ to $\mathcal{G}$, and we write $\mathcal{L}(\mathcal{H}) = \mathcal{L}(\mathcal{H}, \mathcal{H})$ for short. $\mathcal{C}(\mathcal{X}, \mathcal{Y})$ denotes the set of continuous maps from $\mathcal{X}$ to $\mathcal{Y}$, and we write $\mathcal{C}(\mathcal{X}) = \mathcal{C}(\mathcal{X}, \mathbb{R})$ for short. Likewise, for $m \in \mathbb{N}_+$, $\mathcal{C}^m(\mathcal{X}, \mathcal{Y})$ (resp. $\mathcal{C}^m(\mathcal{X})$) denotes the set of m times continuously differentiable functions from $\mathcal{X}$ to $\mathcal{Y}$ (resp. $\mathcal{X}$ to $\mathbb{R}$). If $\mathbf{A}$ is a symmetric matrix, $\lambda_{\max}(\mathbf{A})$ (resp. $\lambda_{\min}(\mathbf{A})$) denotes its largest (resp. smallest) eigenvalue. For a function f of p variables and a multi-index $\alpha = (\alpha_1, ..., \alpha_p) \in \mathbb{N}^p$, we denote the corresponding partial derivative of f (when it exists)

$$\partial^\alpha f(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_p^{\alpha_p}} f(x),$$

where $|\alpha| \stackrel{\text{def}}{=} \sum_{i=1}^p \alpha_i$. Following Rudi et al. (2020), for a function u that is m times differentiable on $\mathcal{X}$, we define the following semi-norm:

$$|u|_{\mathcal{X},m} \stackrel{\text{def}}{=} \max_{|\alpha|=m} \sup_{x \in \mathcal{X}} |\partial^\alpha u(x)|. \tag{1}$$

2.2 Positive-Semidefinite Matrices and Operators

Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. For a linear operator $A : \mathcal{H} \to \mathcal{H}$, we denote A^* its adjoint, $\mathrm{Tr}(A)$ its trace and $\|A\|_{\mathrm{HS}} = \sqrt{\mathrm{Tr}(A^*A)}$ its Hilbert-Schmidt norm. When $\mathcal{H}$ is finite-dimensional, A can be identified with its matrix representation, in which case the Hilbert-Schmidt norm is equal to the usual Frobenius norm $\|\cdot\|_F$. We denote the set of bounded self-adjoint operators on $\mathcal{H}$ by $\mathrm{Sym}(\mathcal{H})$, in finite dimension this corresponds to the set of symmetric matrices. We say that A positive semi-definite (PSD) and write $A \succeq 0$ if A is bounded, self-adjoint and if for all $x \in \mathcal{H}$, it holds $\langle x, Ax \rangle \geq 0$. If further it holds $\langle x, Ax \rangle = 0 \implies x = 0$, we say that A is positive-definite (PD) and write $A \succ 0$. We respectively denote $\mathbb{S}_+(\mathcal{H})$ and $\mathbb{S}_{++}(\mathcal{H})$ the set of PSD and PD operators on $\mathcal{H}$. For a symmetric matrix with eigenvalue decomposition $\mathbf{M} = \mathbf{U}\Sigma\mathbf{U}^T$, we define its positive and negative parts as $[\mathbf{M}]_+ = \mathbf{U}\max(0, \Sigma)\mathbf{U}^T$ and $[\mathbf{M}]_- = \max(0, -\Sigma)\mathbf{U}^T$ (where the max is applied elementwise).

2.3 Kernels and Reproducing Kernel Hilbert Spaces (RKHS).

We refer to Steinwart and Christmann (2008) and Paulsen and Raghupathi (2016) for a more complete covering of the subject. Let $\mathcal{X}$ be a set and $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. k is a positive-definite kernel if and only if for any set of points $x_1, ..., x_n \in \mathcal{X}$, the matrix of pairwise evaluations $K_{ij} = k(x_i, x_j), i, j \in [n]$ is positive semi-definite. Given a kernel k, there exists a unique associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, that is, a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ satisfying the two following properties:

- For all $x \in \mathcal{X}$, $k_x \overset{\text{def}}{=} k(x, \cdot) \in \mathcal{H}$;
- For all $f \in \mathcal{H}$ and $x \in \mathcal{X}$, $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$. In particular, for all $x, x' \in \mathcal{X}$ it holds $\langle k_x, k_{x'} \rangle_{\mathcal{H}} = k(x, x')$.

A feature map is a bounded map $\phi : \mathcal{X} \mapsto \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}, \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = k(x, x')$. A particular instance is $x \mapsto k_x$, which is referred to as the canonical feature map.

A key advantage of RKHSs is that they may be used to translate function-fitting problems into finite-dimensional problems. Indeed, consider a minimization problem of the form

$$\min_{f \in \mathcal{H}} \; L(f(x_1), ..., f(x_n)) + \Omega(\|f\|_{\mathcal{H}}), \tag{2}$$

where $\Omega : \mathbb{R}_+ \mapsto \mathbb{R}$ is a strictly increasing function. In that case, the representer theorem (see, e.g., Steinwart and Christmann, 2008; Paulsen and Raghupathi, 2016, and references therein) shows that solutions of (2) admit the following finite representation: $f = \sum_{i=1}^n \alpha_i k_{x_i}$ for some $\alpha \in \mathbb{R}^n$.

3. Learning Positive-Operator-Valued Functions with Kernel Sums-of-Squares

In this section, we introduce a non-parametric model for functions that take values in $\mathbb{S}_+(\mathbb{R}^d)$, the space of $d \times d$ positive semi-definite matrices.[1] This model has a simple

1. In particular, the dimension p of the inputs need not match that of the output matrices, d.

expression in terms of the feature map ϕ of a RKHS of scalar-valued functions $\mathcal{H}$. Let $\mathcal{H}^d = \{(f_1, \dots, f_d) : f_i \in \mathcal{H}, i \in [d]\}$, and let $\mathbb{S}_+(\mathcal{H}^d)$ denote the space of positive semi-definite self-adjoint operators from $\mathcal{H}^d$ to $\mathcal{H}^d$. Given a feature map $\phi : \mathcal{X} \to \mathcal{H}$, we define our model as

$$F_A(x) = \Phi(x)^* A \Phi(x), \quad \text{with} \quad A \in \mathbb{S}_+(\mathcal{H}^d), \tag{3}$$

where $\Phi(x) \in \mathcal{L}(\mathbb{R}^d, \mathcal{H}^d)$ and the positive semidefinite operator $A \in \mathbb{S}_+(\mathcal{H}^d)$ correspond to

$$\Phi(x) = \underbrace{\begin{pmatrix} \phi(x) & 0_{\mathcal{H}} & \cdots & 0_{\mathcal{H}} \\ 0_{\mathcal{H}} & \phi(x) & \cdots & 0_{\mathcal{H}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathcal{H}} & 0_{\mathcal{H}} & \cdots & \phi(x) \end{pmatrix}}_{d}, \qquad A = \begin{pmatrix} A_{11} & \dots & A_{1d} \\ \vdots & \ddots & \vdots \\ A_{d1} & \dots & A_{dd} \end{pmatrix}, \quad A \succeq 0, \tag{4}$$

with $A_{ij} \in \mathcal{L}(\mathcal{H})$, $i, j \in [d]$. Formally, for $x \in \mathcal{X}$ and $v \in \mathbb{R}^d$, Φ is defined by $\Phi(x)v = (v_1\phi(x), \dots v_d\phi(x))$. Note that eq. (3) can be written equivalently as

$$F_A(x) = \begin{pmatrix} \langle \phi(x),\, A_{11}\phi(x) \rangle & \dots & \langle \phi(x),\, A_{1d}\phi(x) \rangle \\ \vdots & \ddots & \vdots \\ \langle \phi(x),\, A_{d1}\phi(x) \rangle & \dots & \langle \phi(x),\, A_{dd}\phi(x) \rangle \end{pmatrix} \in \mathbb{R}^{d \times d}, \quad \text{with} \quad A \succeq 0.$$

for any $x \in \mathcal{X}$. One can easily check that that the model in eq. (3) is a generalization of Marteau-Ferey et al.'s model for non-negative scalar functions to PSD-valued functions. As such, we will show that they enjoy many similar properties, starting with linearity w.r.t. the parameter A, and the fact that, by construction, the model $F_A(x)$ outputs a positive semidefinite matrix for any $x \in \mathcal{X}$, since A is positive semidefinite. The proof can be found in Appendix B.1, page 24.

Proposition 1 (Linearity and pointwise positivity)**.** *Given $A, B \in \mathcal{L}(\mathcal{H}^d)$ and $\alpha, \beta \in \mathbb{R}$ it holds $F_{\alpha A + \beta B}(x) = \alpha F_A(x) + \beta F_B(x)$. Moreover, if $A \succeq 0$, $F_A(x) \in \mathbb{S}_+(\mathbb{R}^d), \forall x \in \mathcal{X}$.*

Remark 1 (Definition as a sum-of-squares)**.** *Following the nomenclature of Rudi et al. (2020), we refer to the model in eq.* (3) *as a kernel sum-of-squares (SoS). This denomination is motivated by the following observation: let $A = \sum_{k \geq 0} \sigma_k u_k \otimes u_k$ denote the eigenvector decomposition of $A \in \mathbb{S}_+(\mathcal{H}^d)$, where $\sigma_k \geq 0$ and $u_k \in \mathcal{H}^d, k \in \mathbb{N}$. Then, by the reproducing property, we have $F_A(x) = \sum_{k \geq 0} \sigma_k u_k(x) u_k(x)^T, x \in \mathcal{X}$. Hence, $F_A(x)$ is an infinite sum of "squares" of the form $h_k(x)h_k(x)^T$, with functions h_k belonging to the product of RKHS $\mathcal{H}^d$.*

Remark 2 (Matrix-valued kernels)**.** *The construction of the semi-definite model* (3) *holds for more general choices of output spaces and of feature maps Φ. In particular, we can model a function whose values are operators on a separable Hilbert space $\mathcal{Y}$ beyond $\mathbb{R}^d$, by using any operator-valued feature map $\Phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{H} \otimes \mathcal{Y})$ and an operator $A \in \mathbb{S}_+(\mathcal{H} \otimes \mathcal{Y})$, where $\otimes$ is the tensor product. Examples of such feature maps include feature maps associated to vector-valued reproducing kernel Hilbert spaces (vRKHS), which generalize RKHSs to vector-valued functions (Carmeli et al., 2010; Micchelli et al., 2006, see). Like RKHSs, vRKHSs also have an associated kernel function K, which takes values in the space of operators*

$\mathcal{L}(\mathcal{Y}, \mathcal{Y})$, and satisfies $K(x, x') = \Phi(x)^* \Phi(x')$. As an example, the feature map introduced in eq. (4) leads to the simple matrix-valued kernel $K(x, x') = k(x, x')\mathbf{I}_d$. In the remainder of the paper, we will only consider this particular setting, but we extend our results to the general matrix-valued kernel setting in the appendix. We introduce vRKHSs in more detail in Appendix A.

3.1 Representer Theorem, Finite-Dimensional Representation, and Duality

In the previous section, we introduced a model for matrix-valued functions (3) that takes values on the PSD cone (Proposition 1) and is linear with respect to its parameter. However, this parameter is described as an infinite-dimensional positive operator, which is difficult to handle in practice. In this section, we show how to obtain finite-dimensional equivalents of (3) when working on problems for problems of the form (5) that involve a finite number of evaluations of F_A at some given points $\{x_1, ..., x_n\} \subset \mathcal{X}$. More precisely, we start by proving a representer theorem that allows expressing A using a finite-dimensional matrix (Theorem 2), which generalizes Theorem 1 of Marteau-Ferey et al. (2020). Then, as an alternative to the infinite-dimensional features $\Phi(x_1), ..., \Phi(x_n)$, we derive finite-dimensional features $\Psi_1, ..., \Psi_n$ (Section 3.1.2). Finally, we provide a dual formulation for problems of the form (5) that reduces the number of parameters to optimize by a factor n (Theorem 3).

3.1.1 Representer theorem.

Following Marteau-Ferey et al. (2020), for regularized problems of the form

$$\min_{A \in \mathbb{S}_+(\mathcal{H}^d)} \; L(F_A(x_1), \dots, F_A(x_n)) + \Omega(A), \tag{5}$$

with

$$\Omega(A) = \lambda_1 \mathrm{Tr} A + \frac{\lambda_2}{2} \|A\|_{HS}^2, \quad \lambda_1, \lambda_2 \geq 0, \tag{6}$$

we expect that optimal solutions admit a finite-dimensional representation. This is shown in the following theorem, proven in Appendix B.2, page 24. Note that Theorem 2 may be extended to a more general class of regularizers based on spectral functions.

Theorem 2 (Representer theorem)**.** *Let $L : \mathbb{S}_+(\mathbb{R}^d)^n \to \mathbb{R} \cup \{+\infty\}$ be lower semi-continuous and bounded below. Let further Ω be as in eq. (6) with $\lambda_1 > 0$ or $\lambda_2 > 0$. Then eq. (5) has a solution $A_\star$ which may be written*

$$A_\star = \sum_{i,j=1}^{n} \Phi(x_i) \mathbf{C}_{ij} \Phi(x_j)^*,$$

with $\mathbf{C}_{ij} \in \mathbb{R}^{d \times d}, i, j \in [n]$, and $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \dots & \mathbf{C}_{1n} \\ \vdots & & \vdots \\ \mathbf{C}_{n1} & \dots & \mathbf{C}_{nn} \end{pmatrix} \in \mathbb{S}_+(\mathbb{R}^{nd})$. Then, $F_{A_\star}$ can be equivalently written in a finite form in terms of $\mathbf{C}_{ij}$, as follows,

$$F_C(x) = \sum_{i,j=1}^{n} k(x_i, x) k(x_j, x) \mathbf{C}_{ij}, \quad x \in \mathcal{X}. \tag{7}$$

If further L is convex and $\lambda_2 > 0$, $A_\star$ is unique.

3.1.2 Finite-dimensional representation.

The representation of Theorem 2 can be further simplified. Indeed, let $\mathbf{K} \in \mathbb{R}^{n\times n}, K_{ij} = k(x_i, x_j), i, j \in [n]$ and $\tilde{\mathbf{K}} = \mathbf{K} \otimes \mathbf{I}_d \in \mathbb{R}^{nd\times nd}$, where $\otimes$ denotes the Kronecker product. By cyclical invariance of the trace and using the fact that $\Phi(x_j)^\star\Phi(x_i) = k(x_i, x_j)\mathbf{I}_d$, we have

$$\mathrm{Tr}(A_\star) = \sum_{i,j=1}^{n} \mathrm{Tr}(\mathbf{C}_{ij}\Phi(x_j)^*\Phi(x_i)) = \sum_{i,j=1}^{n} k(x_i, x_j)\mathrm{Tr}\mathbf{C}_{ij} = \mathrm{Tr}(\tilde{\mathbf{K}}\mathbf{C}),$$

and for $\ell \in [n]$,

$$F_{A_\star}(x_\ell) = \sum_{i,j=1}^{n} k(x_i, x_\ell)k(x_j, x_\ell)\mathbf{C}_{ij} = (\tilde{\mathbf{K}}\mathbf{C}\tilde{\mathbf{K}})_{\ell,\ell},$$

where $(\tilde{\mathbf{K}}\mathbf{C}\tilde{\mathbf{K}})_{\ell,\ell} \in \mathbb{R}^{d\times d}$ denotes the ℓ-th $d \times d$ block matrix on the diagonal of $\tilde{\mathbf{K}}\mathbf{C}\tilde{\mathbf{K}}$. In particular, if $\mathbf{R} \in \mathbb{R}^{n\times n}$ denotes the upper-triangular Cholesky factor of $\mathbf{K}$ (i.e. $\mathbf{R}$ is the upper triangular matrix satisfying $\mathbf{R}^T\mathbf{R} = \mathbf{K}$), then $\tilde{\mathbf{R}} = \mathbf{R} \otimes \mathbf{I}_d$ is the upper-triangular factor of $\tilde{\mathbf{K}}$. Letting $\mathbf{B} = \tilde{\mathbf{R}}\mathbf{C}\tilde{\mathbf{R}}^T$, we then have

$$\mathrm{Tr}A_\star = \mathrm{Tr}\mathbf{B}, \qquad F_{A_\star}(x_\ell) = \Psi_\ell^T\mathbf{B}\Psi_\ell, \quad \ell \in [n],$$

where $\Psi_\ell = \mathbf{R}_{:,\ell} \otimes \mathbf{I}_d \in \mathbb{R}^{nd\times d}$ denotes the ℓ-th $nd \times d$ block column of $\tilde{\mathbf{K}}$ and $\mathbf{R}_{:,\ell}$ is the ℓ-th column of $\mathbf{R}$. Further, the model $F_{A_\star}$ has the following finite dimensional representation

$$F_B(x) = \Psi(x)^T\mathbf{B}\Psi(x), \tag{8}$$

with $\Psi(x) = (\mathbf{R}^{-T}v(x)) \otimes \mathbf{I}_d \in \mathbb{R}^{nd\times d}$, and $v(x) = (k(x, x_i))_{i=1}^n \in \mathbb{R}^n$. To conclude, using Theorem 2 and the change of variables above, problems of the form of eq. (5) admit a fully finite-dimensional characterization in terms of the following optimization problem over a positive semi-definite matrix of size $nd \times nd$:

$$\min_{\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})} L(F_B(x_1), ..., F_B(x_n)) + \Omega(\mathbf{B}), \tag{9}$$

with $F_B(x_i) = \Psi_i^T\mathbf{B}\Psi_i,\ i \in [n]$.

3.1.3 Dual formulation.

When the loss function L in eq. (5) is convex, we may obtain an equivalent dual formulation that involves n matrices of size $d \times d$, involving the Fenchel conjugate of L, defined as $L^*(Y) \stackrel{\text{def}}{=} \sup_{X\in\mathrm{Sym}(\mathbb{R}^d)^n} \{\sum_{i=1}^n \langle \mathbf{Y}^{(i)}, \mathbf{X}^{(i)}\rangle - L(X)\}, \quad Y \in \mathrm{Sym}(\mathbb{R}^d)^n$.

Theorem 3. *Let $L : \mathrm{Sym}(\mathbb{R}^d)^n \to \mathbb{R}$ be convex, l.s.c. and bounded from below. Let Ω be as in eq. (6). Assume that eq. (9) admits a feasible point, and denote L^* the Fenchel conjugate of L. Then,*

(i) If $\lambda_2 > 0$ and $\lambda_1 \geq 0$, eq. (9) has the following dual formulation:

$$\sup_{\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n} -L^\star(\Gamma) - \frac{1}{2\lambda_2} \Big\| \Big[\sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T + \lambda_1 \mathbf{I}_{nd} \Big]_- \Big\|_F^2, \tag{10}$$

where $\Gamma^{(i)} \in \mathrm{Sym}(\mathbb{R}^d), i \in [n]$ denotes the i-th matrix element of $\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n$, and $[\mathbf{M}]_-$ denotes the negative part of $\mathbf{M}$. This supremum is attained. Further, if $\Gamma_\star$ is a optimal solution of eq. (10), an optimal $\mathbf{B}^\star$ for eq. (9) is

$$\mathbf{B}^\star = \frac{1}{\lambda_2} \Psi^{-1} \Big[\sum_{i=1}^n \Psi_i \Gamma_\star^{(i)} \Psi_i^T + \lambda_1 \mathbf{I}_{nd} \Big]_- \Psi^{-T}.$$

(ii) If $\lambda_2 = 0$ and $\lambda_1 > 0$, eq. (9) has the following dual formulation:

$$\sup_{\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n} -L^\star(\Gamma) \quad s.t. \quad \sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T + \lambda \mathbf{I}_{nd} \succeq 0.$$

Proof in Appendix B.3, page 28.

3.2 Approximation Properties

In the previous sections, we have introduced a model for smooth PSD-valued functions that relies on infinite-dimensional operators and features, and then showed that for finite optimization problems of the form (5) such problems admit a finite representation. We now study the approximation properties of eq. (3). We start by showing that, under mild assumptions on the RKHS $\mathcal{H}$, our model (3) is a universal approximator for PSD-valued functions (Theorem 4).

We adapt the notion of *universality* (Micchelli et al., 2006) to PSD-valued functions: we say that a set of PSD-valued functions $\mathcal{F}$ is universal if for any compact set $\mathcal{Z} \subset \mathcal{X}$, any $\varepsilon > 0$ and any function $g \in \mathcal{C}(\mathcal{Z}, \mathbb{S}_+(\mathbb{R}^d))$, there exists $f \in \mathcal{F}$ such that $\|f_{|\mathcal{Z}} - g\|_{\mathcal{Z}} \leq \varepsilon$, where $\|g\|_{\mathcal{Z}} \overset{\text{def}}{=} \sup_{x \in \mathcal{Z}} \|g(x)\|_\infty$ denotes the max norm over $\mathcal{Z}$. In the following theorem, proved in Appendix B.4, page 30, we show that our model for PSD-valued functions is universal.

Theorem 4 (Universal approximation)**.** *Let $d \in \mathbb{N}, d \geq 1$, $\mathcal{H}$ be a separable Hilbert space and $\phi : \mathcal{X} \to \mathcal{H}$ a universal map (see Micchelli et al., 2006). Then*

$$\mathcal{F} \overset{\text{def}}{=} \{F_A : A \in \mathbb{S}_+(\mathcal{H}^d), \mathrm{Tr} A < \infty\}$$

is a universal approximator of $d \times d$ PSD-valued functions over $\mathcal{X}$.

As examples, universal kernels include the Gaussian kernel $\exp(-\|x - x'\|^2/\sigma^2)$ or the exponential kernel $\exp(-\|x - x'\|/\sigma)$. Informally, Theorem 4 shows that the SoS model (3) may be used to approximate any PSD-valued function arbitrarily well. As a consequence, this model is well-suited to PSD function learning problems, or to model problems with PSD constraints.

3.3 Numerical Illustration: PSD Least-Squares Regression

Let $x_1, ..., x_n \in \mathbb{R}^p$ and $\mathbf{M}_1, ..., \mathbf{M}_n \in \mathbb{S}_+(\mathbb{R}^d)$. We wish to solve

$$\mathcal{P} \stackrel{\text{def}}{=} \min_{\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})} \frac{1}{2n} \sum_{i=1^n} \|F_{\mathbf{B}}(x_i) - \mathbf{M}_i\|_F^2 + \lambda_1 \mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2, \tag{11}$$

with $\lambda_2 > 0$ and $\lambda_1 \geq 0$. Following Theorem 3, we have the following dual formulation:

$$\mathcal{P} = -\frac{1}{2} \min_{\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n} n \sum_{i=1}^n \|\Gamma^{(i)}\|_F^2 + \sum_{i=1}^n \langle \Gamma^{(i)}, \mathbf{M}_i \rangle_F + \frac{1}{\lambda_2} \left\| \Big[\sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T + \lambda_1 \mathbf{I}_{nd} \Big]_- \right\|_F^2, \tag{12}$$

with the primal-dual optimality relation $\mathbf{B}^\star = \frac{1}{\lambda_2} \Psi^{-1} [\sum_{i=1}^n \Psi_i \Gamma_\star^{(i)} \Psi_i^T + \lambda_1 \mathbf{I}_{nd}]_- \Psi^{-T}$. As can be seen in this example, the dual formulation (12) has several advantages over the primal (11). First, the dual problem has only $O(d^2 n)$ parameters (and thus $O(d^2 n)$ memory footprint), compared to the primal which has $O(d^2 n^2)$ parameters. Second, and more importantly, the dual problem no longer has PSD constraints, but only symmetry constraints, which makes it more amenable to (accelerated) first-order methods — at the price of a negative part term that must be computed using SVD. Since (12) is a smooth and strongly convex minimization problem (with convexity parameter $\mu = n$ and smoothness $L = n + \frac{1}{\lambda_2} \lambda_{\max}(\mathbf{K} \circ \mathbf{K})$ with $\mathbf{K} \in \mathbb{R}^{n \times n}, K_{ij} = k(x_i, x_j), i, j \in [n]$), we may compute a solution using accelerated gradient descent (Nesterov, 2003). In particular, since the gradient of the objective function in (12) is a vector of symmetric matrices whenever evaluated at symmetric arguments, we may solve (12) without need for projections if the $\Gamma^{(i)} \in \mathbb{R}^{d \times d}, i \in [n]$ are initialized as symmetric matrices.

We illustrate our model on a geodesic interpolation task. Here, the matrices $\mathbf{M}_1, ..., \mathbf{M}_n \in \mathbb{S}_+(\mathbb{R}^2)$ are sampled from a Bures geodesic (Bhatia et al., 2019) joining $\mathbf{M}_1$ to $\mathbf{M}_n$, at evenly sampled times $x_1 = 0 \leq x_2 \leq ... \leq x_n = 1$. We fit a PSD kernel SoS model by solving (12) using the exponential kernel, and selecting the kernel bandwidth and regularization parameters λ_1 and λ_2 using leave-one-out cross-validation. The results are illustrated in Figure 1, and in Figure 6 in Appendix D on a rank one geodesic. As can be seen in Figure 1, the PSD model is able to faithfully learn the geodesic. Further, Figure 6 shows that our model is able to handle singular inputs, in which case it yields matrices which have a larger conditioning, although we can observe a smoothing effect which yields rank 2 matrices instead of rank 1 in the true geodesic.

As an alternative to our model, one could consider learning PSD matrices under the form $L(x)L(x)^T$ where L is an unconstrained matrix-valued function, or under the form $e^{L(x)}$. However, in the case of the least-squares objective (11), both would result in non-convex problems, for which obtaining a global minimum is generally NP-hard.

Finally, further applications involving PSD-valued functions could be treated in a similar way, such as estimating the conditional covariance of an heteroscedastic Gaussian process, or learning the metric tensor of a Riemannian manifold.

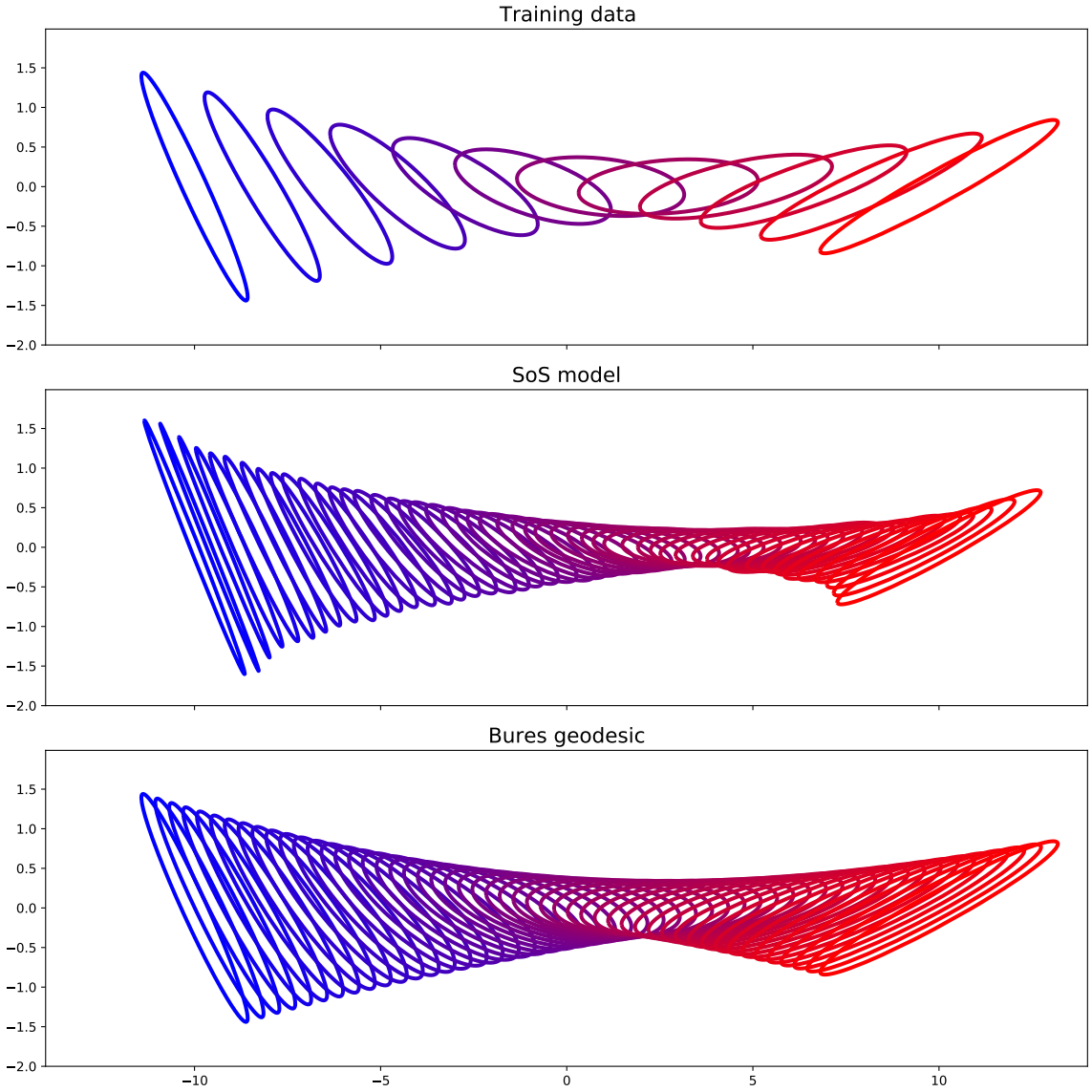


Figure 1: Interpolation of a Bures geodesic from 12 data points (top). The matrices are represented as the level sets of Gaussian distributions: $\{x : \mathcal{N}(x; 0, \Sigma) \leq r\}$ for $r = 0.1$. We use the exponential kernel and select all hyperparameters using cross-validation. The learned model is represented in the middle figure, and the full geodesic is plotted in the bottom figure for comparison.

4. Using the Model to Approximate PSD Contraints in Optimization

Rudi et al. (2020) propose a general technique to deal with constrained optimization problems of the form $\min_\theta L(\theta)$ subject to $g(\theta, x) \geq 0,\ \forall x \in \mathcal{X}$, for an objective function L and constraint function g. It consists in two main steps:

1. First, express the dense set of inequality constraints as an equality constraint in terms of the PSD model for non-negative functions (Marteau-Ferey et al., 2020);
2. Then, apply the equality constraints only on a suitably chosen set of points (e.g., if $\mathcal{X}$ is compact, sampled uniformly at random).

Given $\hat{X} = \{x_1, \ldots, x_n\} \in \mathcal{X}$, Rudi et al. study the effect of approximating the problem above with the problem

$$\min_{\theta, \mathbf{B} \in \mathbb{S}_+(\mathbb{R}^n)} L(\theta) + \Omega(\mathbf{B}) \ \text{ subject to } \ g(\theta, x_i) = \psi_i^\top \mathbf{B} \psi_i, \quad i = 1, \ldots, n,$$

for suitable vectors $\psi_i \in \mathbb{R}^n$, and Ω a regularization for B. They show that, this way, (a) it is possible to use results from approximation theory and leverage the degree of differentiability of the constraint function g, dramatically improving the convergence rate of the approximation for highly differentiable constraints; (b) if L is convex and g is linear in θ, the resulting problem is a finite-dimensional convex problem.

Here we propose to use a similar technique for optimization problems subject to positive semidefinite constraints, using the model introduced in eq. (3), and in particular its finite-dimensional characterization eq. (8). In particular, given a problem of the form

$$\min_{\theta \in \Theta} L(\theta) \quad \text{subject to} \quad g(\theta, x) \succeq 0, \quad \forall x \in \mathcal{X}, \tag{13}$$

for $g : \Theta \times \mathcal{X} \to \mathrm{Sym}(\mathbb{R}^d)$, we suggest to apply the same steps above on the positive semidefinite constraints using the model in eq. (3). The first step corresponds to

$$\min_{\theta \in \Theta, A \succeq 0} L(\theta) + \Omega(A) \quad \text{subject to} \quad g(\theta, x) = F_A(x), \quad \forall x \in \mathcal{X},$$

where the regularizer Ω is defined in eq. (6) and F_A is the model in eq. (3). The second step leads to the following problem:

$$\min_{\theta \in \Theta, \mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})} L(\theta) + \Omega(\mathbf{B}) \quad \text{subject to} \quad g(\theta, x_i) = \Psi_i^\top \mathbf{B} \Psi_i, \quad i = 1, \ldots, n, \tag{14}$$

where $\Psi_i \in \mathbb{R}^{nd \times d}$ (defined in section 3.1.2) are efficiently computable. Note that the latter problem has a finite number of constraints, is finite dimensional and, when L is convex and g is linear in θ, can be solved with standard techniques of convex optimization. In the next theorem we quantify the effect of approximating the constraint set and show that we can leverage the degree of differentiability of the constraint function. Given θ and defining $F : \mathcal{X} \to \mathrm{Sym}(\mathbb{R}^d)$ as $F(x) = g(\theta, x)$, we study the error of approximating the constraint $F(x) \succeq 0, \forall x \in \mathcal{X}$, with the constraint $F(x_i) = \Psi_i^T \mathbf{B} \Psi_i, i \in [n]$ for some $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})$. We quantify the violation of the constraint F outside of $\hat{X}$, i.e., $F(x) \succeq -\varepsilon I$ for any $x \in \mathcal{X}$

and for a ε that depends on the smoothness of the kernel, the smoothness of F and the *fill distance* of $\hat{X}$ w.r.t. $\mathcal{X}$:

$$h_{\hat{X},\mathcal{X}} \overset{\text{def}}{=} \sup_{x\in\mathcal{X}} \min_{i\in[n]} \|x - x_i\|.$$

In particular, when F is m-times differentiable, Theorem 5 shows that if $F(x_i) = \Psi_i^T \mathbf{B}\Psi_i, i \in [n]$ then $F(x) \succeq -\varepsilon I$ over the whole $\mathcal{X}$, where $\varepsilon = O(h^m_{\hat{X},\mathcal{X}})$, leveraging the degree of differentiability of the function F. This result is will be used later in Section 5 to study the optimization problem associated to approximating a convex function. Like Rudi et al. (2020), we require the following assumptions on the domain $\mathcal{X}$ and the RKHS $\mathcal{H}$.

Assumption 1 (Geometric properties of $\mathcal{X}$)**.** *There exists $r > 0$ and a bounded set $S \subset \mathbb{R}^p$ such that $\mathcal{X} = \bigcup_{x\in S} B_r(x)$, where $B_r(x)$ is the open ball centered in x of radius r.*

Assumption 2 (Properties of the RKHS $\mathcal{H}$)**.** *Given a bounded open set $\mathcal{X} \subset \mathbb{R}^p$, let $\mathcal{H}$ be a RKHS of scalar functions on $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{H}}$ kernel k satisfying the following conditions:*

a) There exists $M \geq 1$ such that $\|u \cdot v\|_{\mathcal{H}} \leq M\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}, u, v \in \mathcal{H}$;

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

c) For some $m \in \mathbb{N}_+$ and $D_m \geq 1$, the kernel k satisfies

$$\max_{|\alpha|=m} \sup_{x,y\in\mathcal{X}} |\partial_x^\alpha \partial_y^\alpha k(x, y)| \leq D_m^2 < \infty;$$

Like Rudi et al. (2020), let us remark that that Assumption 2c requires that $\mathcal{H}$ is a RKHS with a m times differentiable kernel, and implies that $\mathcal{H} \subset \mathcal{C}^m(\mathcal{X})$ and $|u|_{\mathcal{X},m} \leq D_m\|u\|_{\mathcal{H}}$. While the SoS model, introduced later in (3), defines a valid PSD-valued function for any feature map Φ, Assumption 1 and Assumption 2 are required to obtain the sampled inequality bounds in Theorem 5 and Corollary 7 and the PSD representation result for Hessians of smooth and strongly convex functions in Theorem 6.

Example 1 (Sobolev kernel (Wendland, 2004))**.** *Let $p \in \mathbb{N}_+$ and $\mathcal{X} \subset \mathbb{R}^p$ be a bounded open set. Let $s > p/2$, and define*

$$k_s(x, x') = c_s\|x - x'\|^{s-p/2}\mathcal{K}_{s-p/2}(\|x - x'\|), \quad \forall x, x' \in \mathcal{X}, \tag{15}$$

where $\mathcal{K}_{s-p/2} : \mathbb{R} \to \mathbb{R}$ is the Bessel function of the second kind with parameter $s - p/2$ (see e.g. Wendland, 2004, Definition 5.10), and $c_s = \frac{2^{1+p/2-s}}{\Gamma(s-p/2)}$ is chosen so that $k_s(x, x) = 1$ for all $x \in \mathcal{X}$. Then, the function k_s is a kernel. When $\mathcal{X}$ has locally Lipschitz boundary (and in particular, when Assumption 1 is satisfied), its corresponding RKHS $\mathcal{H}$ is $W_2^S(\Omega)$, the Sobolev space of functions whose weak-derivatives are square-integrable up to order s (Adams and Fournier, 2003). Then eq. (15) satisfies Assumption 2 for any $m \in \mathbb{N}_+$ s.t. $m < s - p/2$ (see Rudi et al., 2020, Proposition 1) with the following constants:

$$M = (2\pi)^{p/2}2^{s+1/2}, \quad D_m = (2\pi)^{p/2}\sqrt{\frac{\Gamma(m + p/2)\Gamma(s - p/2 - m)}{\Gamma(s - p/2)\Gamma(p/2)}}.$$

Finally, note that in the particular case $s = p/2 + 1/2$, we have $k_s(x, x') = \exp(-\|x - x'\|)$ and that a scale factor can be added as $k_s(x, x') = \exp(-\|x - x'\|/\sigma)$, and similarly in eq. (15) by replacing $\|x - x'\|$ with $\|x - x'\|/\sigma$.

Theorem 5. *Let $\mathcal{X} \subset \mathbb{R}^p$ satisfy Assumption 1 for some $r > 0$. Let k satisfy Assumption 2a and c for some $m \in \mathbb{N}_+$. Let $\hat{X} = \{x_1, ..., x_n\} \subset \mathcal{X}$ such that $h_{\hat{X},\mathcal{X}} \leq r \min(1, \frac{1}{18(m-1)^2})$, $h_{\hat{X},\mathcal{X}} \overset{\text{def}}{=} \sup_{x \in \mathcal{X}} \min_{i \in [n]} \|x - x_i\|$. Let $F \in C^m(\mathcal{X}, \mathrm{Sym}(\mathbb{R}^d))$ and assume there exists $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})$ such that*

$$F(x_i) = \Psi_i^T \mathbf{B} \Psi_i, \quad i \in [n],$$

where the Ψ_i's are defined in Section 3.1. Then, it holds

$$\forall x \in \mathcal{X}, \quad \lambda_{\min}(F(x)) \geq -\varepsilon, \quad \textit{where} \quad \varepsilon = C h_{\hat{X},\mathcal{X}}^m,$$

and $C = C_0(\sum_{i=1}^d |F_{ii}|_{\mathcal{X},m} + M D_m \mathrm{Tr}\mathbf{B})$ with $C_0 = 3\frac{p^m}{m!} \max(1, 18(m-1)^2)^m$, and where F_{ij} is the the scalar-valued function corresponding to the element of F with index $i, j \in [d]$.

As a remark, let us notice that the bounds in Theorem 5 involve both the dimension p of $\mathcal{X}$ and the size d of the matrices output by F, which need not be equal. Theorem 5 is proved in Appendix B.5, page 31. We sketch here the main arguments of the proof. The followed strategy consists in using scattered data inequality for scalar functions given by Theorem 13 of Rudi et al. (2020) (itself adapted from Wendland (2004)). To do so, we start by considering a fixed direction in the unit sphere $u \in S^{d-1}$, and apply the theorem to the smooth scalar function $g_u(.) \overset{\text{def}}{=} u^T(F(.) - F_{\mathbf{B}}(.))u$, to obtain lower bounds on $u^T F(.)u$. When then derive global bounds (independent of u) over S^{d-1}. Finally, since the lowest eigenvalue of a symmetric matrix $\mathbf{M} \in \mathrm{Sym}(\mathbb{R}^d)$ is $\min_{u \in S^{d-1}} u^T \mathbf{M} u$, we obtain a global lower bound on $\lambda_{\min}(F(x))$ for $x \in \mathcal{X}$. Using Theorem 5, we may turn F into a function that is PSD everywhere by adding $\varepsilon \mathbf{I}_d$. Whenever the constraint function $g(\theta, x)$ allows relating the addition of the constant term $\varepsilon \mathbf{I}_d$ to a change in θ (e.g. if $g(\theta, x) = A(x)[\theta] + B(x)$ and $A[x]$ linear), we may use Theorem 5 to bound $L(\theta_\star) - L(\hat{\theta})$, where $\theta_\star$ is the solution of eq. (13) and $\hat{\theta}$ the solution of eq. (14), as in Theorem 8 in Section 5 for optimization under convexity constraints.

5. Modeling Convex Functions with Sum-of-Squares Hessians

In Section 3, we introduced a model for functions that take PSD matrices as values. While some problems such as those mentioned in the same section involve learning PSD-valued functions directly, positive semidefiniteness most commonly appears as a constraint in optimization problems. In particular, when the constraint is smooth enough, the prevalent problem of optimizing a scalar-valued function under convexity constraints can be reformulated using PSD constraints on the Hessian, which allows leveraging the smoothness of the constraint, as shown in Theorem 5. That is, problems of the form

$$\min J(f) \quad \text{s.t.} \quad f \in \mathcal{C}^2(\mathcal{X}) \text{ is convex}, \tag{16}$$

may be written as follows

$$\min_{f \in \mathcal{C}^2(\mathcal{X})} J(f) \quad \text{s.t.} \quad H_f(x) \succeq 0, \quad \forall x \in \mathcal{X}. \tag{17}$$

Problems of the form (16) are notoriously difficult to handle and as a consequence most works on this topic have focused on piecewise-affine functions (Seijo and Sen, 2011), or functions defined on low-dimensional (usually 2D) domains (Mérigot and Oudet, 2014). In Section 4, we extended the approach introduced by Rudi et al. (2020) for non-negative constraints, to approximate optimization problems subject to a dense set of positive semidefinite constraints, as the problem in eq. (17). In the present section, we use the proposed extension to approximate efficiently eq. (17). In particular, in the next two subsections we will transform eq. (17) as described in Section 4, and we will derive quantitative bounds on the approximation error and computational complexity of the finite dimensional problem in eq. (19).

5.1 From Positive Semidefinite to Equality Constraints

The first step of the approach described in section 4 consists in transforming the positive semidefinite constraints into equality constraints with respect to the PSD sum-of-squares model. In our case, this corresponds to leveraging the PSD sum-of-squares model introduced in Section 3 to enforce positiveness constraints on the Hessian of the variable f. After adding a regularizer, which is required to obtain the finite-dimensional representation of Theorem 2 and to control the eigenvalues of H_f (Theorem 5), the resulting problem has the following form:

$$\min_{f\in\mathcal{H}, A\in\mathbb{S}_+(\mathcal{H}^d)} J(f) + \Omega(A) \quad \text{s.t.} \quad H_f(x) = F_A(x), \quad \forall x \in \mathcal{X}, \tag{18}$$

where Ω is defined in eq. (6). In the following theorem, we show that encoding the Hessian as a PSD SoS allows recovering any sufficiently smooth strongly convex function.

Theorem 6 (PSD sum-of-squares representation for convex functions)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{H}$ be a RKHS of functions on $\mathcal{X}$ satisfying Assumption 2a-c, and such that $\mathcal{C}^s(\mathcal{X}) \subset \mathcal{H}$, $s \in \mathbb{N}$. Let $f \in \mathcal{C}^{s+2}(\mathcal{X})$ be strongly convex. Then, there exists $A \in \mathbb{S}_+(\mathcal{H}^d)$ such that $H_f(x) = F_A(x), \forall x \in \mathcal{X}$.*

Proof in Appendix C.1, page 33. Let us mention that the smoothness constraint $\mathcal{C}^s \subset \mathcal{H}$ implies $s > d/2$, but that the strong convexity is merely a sufficient condition. As an example, the function $f : (x, y) \in \mathbb{R}^2 \to \frac{1}{2}(x - y)^2$ admits the constant Hessian $H_f(x, y) = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, which is positive semi-definite, but not positive definite. However, its Hessian may be encoded as a PSD SoS for any RKHS containing constant functions. Hence, we conjecture that finer existence conditions could be proven.

5.2 Finite-Dimensional Representation

In this section, we show how enforcing convexity via a SoS model for the Hessian, $H_f(x) = F_A(x),\ \forall x \in \mathcal{X}$ with $A \succeq 0$, leads to tractable optimization.

5.2.1 Subsampling the Constraints.

The generic problem (18) has an infinite number of constraints. Hence, it is not amenable to computation. As described in Section 4, we subsample its constraints on a set $\{x_1, ..., x_n\} \subset \mathcal{X}$.

Following Theorem 2, we then obtain a finite-dimensional version of problem (18):

$$\min_{f\in\mathcal{H},\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})} J(f)+\Omega(\mathbf{B}) \;\text{ s.t. }\; H_f(x_j)=\Psi_j^T\mathbf{B}\Psi_j,\quad j\in[n]. \tag{19}$$

Subsampling the constraints guarantees that solutions of eq. (19) are convex in the neighborhood of $\{x_1,...,x_n\}$, but does not guarantee global convexity. However, in the next result we leverage Theorem 5 to obtain global bounds on the convexity deficit of f, i.e., the magnitude of the quadratic function that we have to add to f to make it convex, as a function of the fill distance $h_{\hat{X},\mathcal{X}} \overset{\text{def}}{=} \sup_{x\in\mathcal{X}} \min_{i\in[n]} \|x-x_i\|$.

Corollary 7 (Theorem 5)**.** *Let $\mathcal{X}$ satisfy Assumption 1 for some $r>0$. Let k satisfy Assumption 2a and Assumption 2c for some $m>d/2$. Let $\hat{X}=\{x_1,...,x_n\}\subset\mathcal{X}$ such that $h_{\hat{X},\mathcal{X}}\leq r\min(1,\frac{1}{18(m-1)^2})$. Let $f\in\mathcal{C}^{m+2}(\mathcal{X})$ and $\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{dn})$ be such that*

$$H_f(x_j)=\Psi_j^T\mathbf{B}\Psi_j,\quad j\in[n].$$

Then,

$$f(x)+\tfrac{\eta}{2}\|x\|^2 \;\; \textit{is convex} \;\; \textit{for} \;\; \eta\geq Ch_{\hat{X},\mathcal{X}}^m,$$

where $C=C_0(\sum_{i=1}^d|(H_f)_{ii}|_{\mathcal{X},m}+MD_m\mathrm{Tr}\mathbf{B})$ and C_0 is as in Theorem 5, and where $(H_f)_{ij}$ is the the scalar-valued function corresponding to the element of the Hessian H_f with index $i,j\in[d]$.

Proof in Appendix C.2, page 33. Corollary 7 guarantees that the strong convexity deficit of f goes to zero as the the number of subsampled inequalities increases, provided they cover the domain $\mathcal{X}$. It is important to note that the strong convexity deficit goes to zero with a rate that is exponentially decreasing in the degree of differentiability of f. Let us remark that another option to obtain convex function beyond adding $\frac{\eta}{2}\|x\|^2$ correspond to enforcing the constraints $H_f(x_j)=\Psi_j^T\mathbf{B}\Psi_j+\eta\mathbf{I}_d,\;\; j\in[n]$ in eq. (19). However, the constant η given by Corollary 7 depends on f and $\mathbf{B}$ and is therefore not known in advance. Under further assumptions on the functional J, η can be used to quantify the error induced by solving (19) instead of (17), as shown in the following theorem.

Theorem 8. *Let $\mathcal{X}$ satisfy Assumption 1 for some $r>0$. Let $\hat{X}=\{x_1,...,x_n\}\subset\mathcal{X}$ such that $h_{\hat{X},\mathcal{X}}\leq r\min(1,\frac{1}{18(m-1)^2})$. Let $\mathcal{H},k$ satisfy Assumption 2a and Assumption 2c for some $m>d/2$, and let $\mathcal{F}$ satisfy Assumption 2c for $m'=m+2$. Let $(f_\star,A_\star)$ be a solution of*

$$\min_{f\in\mathcal{F},\; A\in\mathbb{S}_+(\mathcal{H}^d)} J(f) \;\; s.t. \;\; H_f(x)=F_A(x),\;\; x\in\mathcal{X},$$

and $(\hat{f},\hat{\mathbf{B}})$ be the solution of

$$\min_{f\in\mathcal{F},\; \mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})} J(f)+\rho\|f\|_{\mathcal{F}}^2+\lambda\mathrm{Tr}\mathbf{B} \;\; s.t. \;\; H_f(x_j)=\Psi_j^T\mathbf{B}\Psi_j,\;\; j\in[n],$$

and let η be as in Corollary 7. Assume that J is L-Lipschitz for some seminorm $N(\cdot)$. Then, $\tilde{f}=x\mapsto\hat{f}(x)+\frac{\eta}{2}\|x\|^2$ is a convex function on $\mathcal{X}$ and it holds

$$J(\tilde{f})-J(f_\star)\leq R^2(1+\tfrac{MC_1}{\lambda}h_{\hat{X},\mathcal{X}}^m)\,\rho \;+\; C_1dR\; h_{\hat{X},\mathcal{X}}^m. \tag{20}$$

where $R^2 = \frac{\lambda}{\rho}\mathrm{Tr}A_\star + \|f_\star\|^2_{\mathcal{F}}$ *and* $C_1 = \frac{1}{2}LC_NC_0\max(D_m, D_{m+2})$, $C_N = N(s)$ *and* s *is the function* $s(x) = \|x\|^2$ *for* $x \in \mathbb{R}^d$. *In particular, when* $\rho = \lambda = MC_1h^m_{\hat{X},\mathcal{X}}$, *then*

$$J(\tilde{f}) - J(f_\star) \;\leq\; Q\; h^m_{\hat{X},\mathcal{X}},$$

with $Q = 4MC_1(\mathrm{Tr}A_\star + \|f_\star\|^2_{\mathcal{F}} + \frac{d^2}{M^2})$.

Proof in Appendix C.3, page 34. Note that if we choose the discretization points uniformly at random in $\mathcal{X}$ and $\mathcal{X}$ is a convex set, then $h_{\hat{X},\mathcal{X}} \leq (C\frac{1}{n}\log(1/\delta))^{1/d}$, with probability $1-\delta$. Then, the algorithm above leads to an error in the order of

$$J(\tilde{f}) - J(f_\star) \leq\; Q'\;(\mathrm{Tr}A_\star + \|f_\star\|^2_{\mathcal{F}} + \frac{d^2}{M^2})\; n^{-m/d},$$

where $Q' = Q(\log\frac{1}{\delta})^{m/d}$, while Q depends only on m, d and is exponential in $\max(m, d)$.

5.2.2 Encoding the constraints.

To enforce the constraints $H_f(x_j) = \Psi_j^T\mathbf{B}\Psi_j,\ j \in [n]$ we must choose a representation of the Hessian of f at points $\{x_1, ..., x_n\}$. In this section, we propose two such representations. The first (i) leverages the reproducing property on derivatives (Zhou, 2008), to obtain an exact representation, whereas the second (ii) is an approximate representation, that is only asymptotically accurate but is more amenable to computation.

(i) Reproducing property for derivatives. If the kernel k is regular enough, reproducing properties hold for function derivatives (Zhou, 2008). More precisely, if $k \in \mathcal{C}^{2s}(\mathcal{X}\times\mathcal{X})$ for some $s \in \mathbb{N}_+$, then it holds:

- $\forall x \in \mathcal{X}$, $\forall \alpha \in \mathbb{N}^p$ s.t. $|\alpha| \leq s$, $\partial^\alpha k_x \overset{\text{def}}{=} \frac{\partial^{|\alpha|}k(s,\cdot)}{\partial s_1^{\alpha_1},...,\partial s_p^{\alpha_p}}\Big|_{s=x} \in \mathcal{H}$;
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \forall \alpha \in \mathbb{N}^p$ s.t. $|\alpha| \leq s$, $\partial^\alpha f(x) = \langle f,\, \partial^\alpha k_x\rangle_{\mathcal{H}}$. In particular, for all $x, x' \in \mathcal{X}$ it holds $\langle \partial^\alpha k_x,\, \partial^\alpha k_{x'}\rangle_{\mathcal{H}} = \frac{\partial^{2|\alpha|}k(s,t)}{\partial s_1^{\alpha_1},...,\partial s_p^{\alpha_p}\partial t_1^{\alpha_1},...,\partial t_p^{\alpha_p}}\Big|_{\substack{s=x\\t=x'}}$.

In this work, we will mostly consider derivatives up to order 2. Hence, we will use the following simplified notations:

$$\partial_i k_x \overset{\text{def}}{=} \frac{\partial k(s,\cdot)}{\partial s_i}\Big|_{s=x} \quad\text{and}\quad \partial_{ij}k_x \overset{\text{def}}{=} \frac{\partial^2 k(s,\cdot)}{\partial s_i \partial s_j}\Big|_{s=x},\quad x \in \mathcal{X}, i, j \in [p].$$

Let $\{e_p, p \in [d]\}$ denote the canonical basis of $\mathbb{R}^d$. Then, for $v \in \mathcal{X}$ we have $H_f(v) = \sum_{p,q=1}^d \frac{\partial^2 f(v)}{\partial v_p v_q} e_p e_q^T$. Hence, we may rewrite the constraints as

$$\sum_{p,q=1}^d \langle f,\, \partial_{pq}k_{x_j}\rangle e_p e_q^T = \Psi_j^T\mathbf{B}\Psi_j,\quad j \in [n]. \tag{21}$$

This yields the following problem:

$$\min_{f\in\mathcal{H},\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})} J(f) + \Omega(\mathbf{B})\;\text{ s.t. }\;\sum_{p,q=1}^d \langle f,\, \partial_{pq}k_{x_j}\rangle e_p e_q^T = \Psi_j^T\mathbf{B}\Psi_j,\quad j \in [n]. \tag{22}$$

We may then turn eq. (22) into a finite-dimensional optimization problem by considering its dual formulation. This is illustrated on a convex regression problem in Appendix C.5.

(ii) Approximate representation. While the representation in eq. (21) allows us to model the Hessian exactly, it may be cumbersome to use in practice. As an example, if the objective functional J contains quadratic terms, solving eq. (22) requires evaluating $\langle \partial_{pq} k_{x_i}, \partial_{rs} k_{x_j} \rangle$ for $p, q, r, s \in [d], i, j \in [n]$, which quickly becomes prohibitive as the dimension d increases. As an alternative, one may expand f along some features $\{k_{z_i}, i \in [\ell]\}$ (which may be given by the problem, e.g. in regression problems), and optimize the weights $\alpha \in \mathbb{R}^\ell$ such that $f(x) = \sum_{i=1}^\ell \alpha_i k(x, z_i)$. The constraints are then applied on α by deriving the Hessian of f:

$$\sum_{p,q=1}^{d} \sum_{i=1}^{\ell} \alpha_i \frac{\partial^2 k(x_j, z_i)}{\partial x_j^p x_j^q} e_p e_q^T = \Psi_j^T \mathbf{B} \Psi_j, \quad j \in [n]. \tag{23}$$

Compared to the exact representation eq. (21), this representation only involves second order derivatives of the kernel, compared to fourth order. Further, it allows controlling the size of the expansion more easily: compared to (ii), using (i) implies using an additional coefficient for each $\partial_{pq} k_{x_j}$, of which there are $nd(d+1)/2$ (taking symmetry into account). However, while the pair $(0_\ell, 0_{\ell d \times \ell d})$ always satisfies eq. (23), compared to eq. (21), some hypothesis is required to ensure that set of pairs $(\alpha, \mathbf{B}) \in \mathbb{R}^\ell \times \mathbb{S}_+(\mathbb{R}^{\ell d})$ satisfying eq. (23) has non-empty interior, so that it can be used in practice. Hence, we require the following hypothesis:

(H_1) There exists $\alpha \in \mathbb{R}^\ell$ such that $f(\cdot) \stackrel{\text{def}}{=} \sum_{i=1}^\ell \alpha_i k(\cdot, z_i)$ is strictly convex in $x_j, j \in [n]$.

As examples, for fixed points $(x_1, ..., x_n) \in \mathbb{R}^d$ and with the Gaussian or the exponential kernel, there exists an α satisfying (H_1) if $z_i = x_i, i \in [n]$ and the bandwidth is small enough, or if the kernel is universal and the number of distinct sample points ℓ is larger than nd^2. We leave to finding finer conditions satisfying (H_1) for future work.

5.2.3 Choosing a representation for smooth convex functions.

We end this section with a discussion on the choice of PSD SoS to enforce convexity for smooth functions. Compared to eq. (18), several alternative approaches based on scalar kernel SoS that enforce convexity may be considered, which we now briefly examine. Let us motivate our choice based on a few criteria that should be taken into account:

1. *Representation power*: the class of functions that may be encoded;
2. *Sampling*: the amount of data that must be sampled to learn such a representation;
3. *Scattered data inequalities*: the convexity guarantees one can obtain from subsampled constraints;
4. *Computational cost*: the cost of enforcing those constraints in convex variational problems.

In light of those criteria, let us consider two alternative approaches. The first consists in enforcing that the function f lies above its tangents:

$$\forall x, y \in \mathcal{X}, f(x) - f(y) - \langle \nabla f(y),\, x - y \rangle = \langle \phi(x, u),\, A\phi(x, u) \rangle_{\mathcal{H}}, \quad A \in \mathbb{S}_+(\mathcal{H}(\mathcal{X} \times \mathcal{X})), \tag{24}$$

where $\mathcal{H}(\mathcal{X} \times \mathcal{X})$ denotes a RKHS of functions over $\mathcal{X} \times \mathcal{X}$. The second consists in enforcing that the Hessian of f should have non-negative eigenvalues by representing bilinear products as kernel SoS:

$$\forall x \in \mathcal{X}, \forall u \in S^{d-1}, u^T H[f](x) u = \langle \phi(x,u),\, A\phi(x,u) \rangle_{\mathcal{H}}, \quad A \in \mathbb{S}_+(\mathcal{H}(\mathcal{X} \times S^{d-1})), \quad (25)$$

where $S^{d-1} \subset \mathbb{R}^d$ is the unit sphere. Following the same proof techniques as in Theorem 6 and Theorem 5 of Vacher et al. (2021), one may show that (25) and eq. (18) have the same representation power, whereas (24) may represent function that are 1 order less regular (i.e. $m > 1 + d/2$ instead of $m > 2 + d/2$). In terms of sampling, (24) requires covering $\mathcal{X}^2$, (25) covering $\mathcal{X} \times S^{d-1}$, while eq. (18) only $\mathcal{X}$, which is more efficient and in particular allows obtaining smaller discrete problems. Next, extending the scattered data inequalities from Rudi et al. (2020) (Theorem 13), one may show that (24) yields $\forall x, y \in \mathcal{X}, f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq -\eta$ with $\eta > 0$, which does not translate to (strong) convexity guarantees, while (25) and eq. (18) (from Corollary 7) imply that f is $-\eta$-strongly convex for some $\eta > 0$. Finally, while (24) and (25) rely on scalar-valued kernels compared to matrix-valued kernels, the fact that they require sampling two variables leads to a number of dual variables of the order $O(n^2)$ (if we sample n variables from each domain), compared to $O(nd^2)$ for matrix models. Since in most applications, d is negligible compared to n, the PSD matrix model is also more interesting computationally.

6. Application to Convex Regression

In this section, we illustrate the model introduced in Section 5 on a convex regression task. In convex regression, we are given a training set $(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, over which we aim to fit a function $f : \mathbb{R}^d \to \mathbb{R}$ according to a least squares loss, under the constraint that f is convex:

$$\min_{f:\mathbb{R}^d \to \mathbb{R}} \ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \ \text{ s.t. } \ f \ \text{ convex.} \quad (26)$$

This problem was first considered in the 1950's (Hildreth, 1954), motivated by applications to economics. The prevalent approach is to solve (26) on the set of piecewise-affine functions (Seijo and Sen, 2011), which amounts to solving the following linearly constrained quadratic program:

$$\min_{\substack{\theta \in \mathbb{R}^n \\ \zeta_1, ... \zeta_n \in \mathbb{R}^d}} \ \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta_i)^2 \ \text{ s.t. } \ \theta_i + \zeta_i^T (x_i - x_j) \leq \theta_j, \quad i, j \in [n]. \quad (27)$$

In eq. (27), θ_i represents the value of f at x_i, and ζ_i a subgradient of f at x_i. Hence, the constraints in eq. (27) correspond to eq. (24) for piecewise-linear functions. From there, the estimator may be computed at a new point x as $\hat{f}(x) = \max_{i=1,..,n} \{\hat{\theta}_i + \hat{\zeta}_i^T (x - x_i)\}$. While this method yields a convex function, it may not leverage the smoothness of the data, e.g. when the samples are distributed as $Y = f_\star(X) + \varepsilon$ for some smooth function $f_\star$ and noise ε.

Using the kernel SoS model introduced in Section 5, we may perform linear regression with smooth functions, under approximate convexity guarantees given by Corollary 7. We

focus here on the approximate representation (ii) of Section 5, and refer to Appendix C.5 for the exact representation (i) using reproducing properties of derivatives. We consider the particular case where we enforce convexity at the sample points $x_1, ..., x_n$. This can be straightforwardly generalized to the case where convexity is enforced on a different set $\{v_1, ..., v_\ell\} \subset \mathcal{X}$. Plugging $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ in eq. (26) with the constraints eq. (23) and with an additional ridge regularization term $\|f\|_{\mathcal{H}}^2 = \alpha^T \mathbf{K} \alpha$, we obtain the following problem.

$$\begin{aligned} \min_{\substack{\alpha \in \mathbb{R}^n \\ \mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})}} \quad & \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^n \alpha_i k(x_i, x_j))^2 + \lambda_1 \mathrm{Tr} \mathbf{B} + \frac{\lambda_2}{2} \|\mathbf{B}\|_F^2 + \rho \alpha^T \mathbf{K} \alpha \\ \text{s.t.} \quad & \forall j \in [n], \forall p, q \in [d], \quad \sum_{i=1}^n \alpha_i \frac{\partial^2 k(x_i, x_j)}{\partial x_j^p \partial x_j^q} = e_p^T (\Psi_j^T \mathbf{B} \Psi_j) e_q. \end{aligned} \tag{28}$$

Proposition 9. *Assuming* (H_1)*, problem* (28) *admits the following dual formulation:*

$$\min_{\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n} Z(\Gamma)^T (\frac{1}{n} K^2 + \rho K)^{-1} Z(\Gamma) + \frac{1}{2\lambda_2} \|[\sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T + \lambda_1 \mathbf{I}_{nd}]_-\|_F^2, \tag{29}$$

with $Z(\Gamma) \overset{\text{def}}{=} \frac{1}{2} \sum_{i=1}^n \sum_{p,q=1}^d \Gamma_{p,q}^{(i)} \frac{\partial^2 K(X, x_j)}{\partial x_j^p \partial x_j^q} + \frac{1}{n} \mathbf{K} y \in \mathbb{R}^n$*,* $\mathbf{K} \overset{\text{def}}{=} [k(x_i, x_j)]_{i,j \in [n]} \in \mathbb{R}^{n \times n}$ *and*

$$\forall p, q \in [d], j \in [n], \frac{\partial^2 K(X, x_j)}{\partial x_j^p \partial x_j^q} \overset{\text{def}}{=} \left[\frac{\partial^2 K(x_i, x_j)}{\partial x_j^p \partial x_j^q} \right]_{i \in [n]} \in \mathbb{R}^n.$$

Proof in Appendix C.4, page 35. Problem (29) is smooth and convex, hence it is amenable to accelerated gradient descent. However, its Hessian may have arbitrarily small eigenvalues, therefore it is generally not strongly convex. The computational bottleneck in solving eq. (29) is the computation of the negative part, which takes $O(n^2 d^3)$ time to form the matrix $\sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T + \lambda_1 \mathbf{I}_{nd}$ and $O(n^3 d^3)$ time to compute its SVD. Following Rudi and Ciliberto (2021) and Muzellec et al. (2021), we may reduce the computational load by using a Nystrom approximation of $\mathbf{K}$ (see e.g. Williams and Seeger, 2001) to lower the size of the features $\Psi_i, i \in [n]$ from $nd \times d$ to $rd \times d$ with $r < n$. This brings down the cost of forming the matrix to $O(nrd^3)$ and the cost of the SVD to $O(r^3 d^3)$, hence to a total cost of $O(nrd^3 + r^3 d^3)$ operations per iteration.

6.1 Numerical Experiments

We illustrate our method on a convex regression task where the data is sampled from a convex function, whose Hessian is PSD but not positive definite, with the addition of noise. More precisely, let $a > 0$ and define $f_a(x) = (\cos(ax) - 1)/a^2 + x^2/2$. We have $f_a''(x) = 1 - \cos(ax) \geq 0$. Hence, f is convex, but not strictly so: its second derivative has countably many zeroes. In our experiments, we sample features $X_1, ..., X_n$ uniformly in a hyper cube $[-b, b]^d$, and generate outputs

$$Y_i = f_a(\|X_i\|) + \eta \cdot \varepsilon_i, \ i \in [n], \text{ with } \eta > 0 \text{ and } \varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, 1), \ i \in [n]. \tag{30}$$

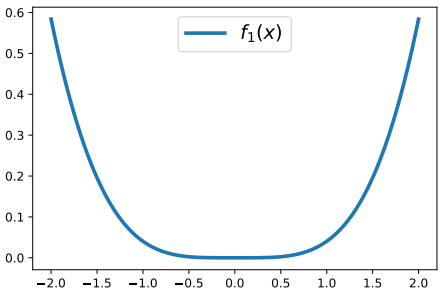
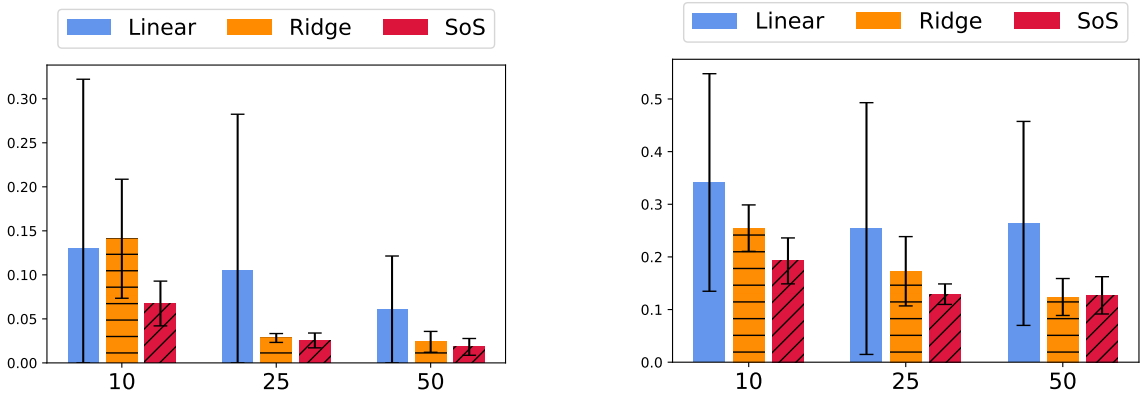


Figure 2: Ground truth (1D)

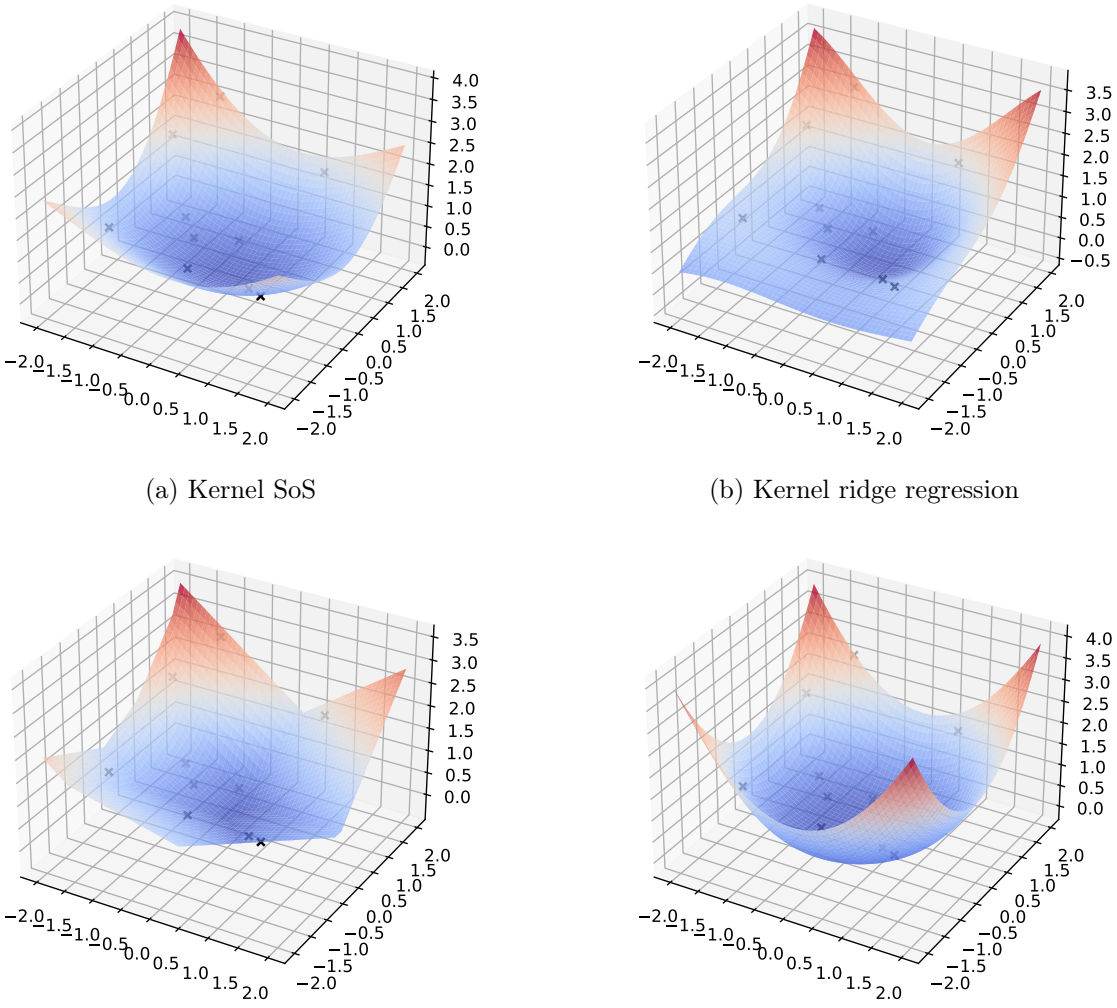
We then fit a regression function by solving (29), using the Gaussian kernel and selecting the hyperparameters with 5-fold cross-validation. We compare kernel SoS models to the piecewise linear convex regression as in eq. (27), and to unconstrained kernel ridge regression fitted using cross-validatoin (which may output a non-convex function, as in Figure 4). A sample result is illustrated in Figure 4. In Figure 3, we report the mean square error of all 3 methods (evaluated by sampling 10000 additional points according to eq. (30)), as a function of the number of training samples and of the noise level η. In Figure 3, the fact that kernel



(a) $\eta = 0.1$ (b) $\eta = 0.3$

Figure 3: Mean square error (MSE) of functions fitted using convex piecewise linear regression, kernel ridge regression and kernel convex sum-of-squares regression on 2D data generated according to eq. (30) with $a = 1$, as a function of noise (η) and number of samples (x-axis). Full bars represent averaged scores over 10 runs, error bars correspond to plus/minus standard deviation.

ridge regression performs overall better that piecewise linear convex regression shows that taking smoothness into account allows an improvement of the mean square error. The kernel SoS formulation (28) allows enforcing both smoothness and convexity priors, which yields an additional improvement over kernel ridge regression, as shown in Figure 3. This is also illustrated in Figure 4: in this example, piecewise linear convex regression yields a convex but non-smooth function, kernel ridge regression yields a smooth but non-convex function,



(a) Kernel SoS

(b) Kernel ridge regression

(c) Piecewise linear convex regression

(d) Ground truth

Figure 4: Outcome of kernel convex SoS regression, kernel ridge regression and piecewise linear convex regression fitted on 10 points compared to ground truth f_1. The hyperparameters of kernel SoS and ridge regression are selected using 5-fold cross-validation.

while kernel SoS regression yields a smooth and convex function that generalizes better w.r.t. the "true" function.

Conclusion and Future Work

In this paper, we have introduced an extension of the kernel SoS models of Marteau-Ferey et al. (2020) to PSD matrix-valued functions. While we chose to present our model using a particular feature map that relies on a scalar-valued kernel, it straightforwardly extends to more general matrix-valued kernels, as shown in Appendix A, which may be of use to leverage the properties of given kernels, such as rotational-free or divergence-free kernels (Macêdo and Castro, 2008). We then showed how PSD-valued sums-of-squares could be used in variational problems with convexity constraints, and illustrated our method on a convex regression task. Possible further applications of our method include economics and monopolistic games in particular (Choné and Le Meur, 2001; Mirebeau, 2016), shape optimization problems such as Newton's least resistance problem (Lachand-Robert and Oudet, 2005; Mérigot and Oudet, 2014) or optimal transport with a quadratic cost (Brenier, 1991; Makkuva et al., 2020).

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 from the European Research Council (grants SEQUOIA 724063 and REAL 947908), and support by grants from Région Ile-de-France.

Appendix A. Vector-Valued Functions and Matrix-Valued Kernels

A.1 Vector-Valued RKHSs

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y}$ a Hilbert space. A map $K : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{Y} \times \mathcal{Y}$ is called a $\mathcal{Y}$-reproducing kernel (Carmeli et al., 2010) if $\sum_{i,j=1}^{N} \langle K(x_i, x_j) y_i, y_j \rangle \geq 0$ for any $x_1, ..., x_N \in \mathcal{X}$ and $y_1, ..., y_N \in \mathcal{Y}$. Given a $\mathcal{Y}$-reproducing kernel K, there is a unique Hilbert space $\mathcal{H}_K \subset \mathcal{Y}^{\mathcal{X}}$ satisfying

(i) $\forall x \in \mathcal{X}, K_x \in \mathcal{L}(\mathcal{Y}, \mathcal{H}_K)$ where for $y \in \mathcal{Y}$, $K_x y$ is defined by $(K_x y)(t) = K(t, x)y, t \in \mathcal{X}$,

(ii) $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_K, f(x) = K_x^* f$ where $K_x^* \in \mathcal{L}(\mathcal{H}_K, \mathcal{Y})$ denotes the adjoint of K_x.

In particular, we have $K(x, y) = K_x^* K_y, x, y \in \mathcal{X}$. When $\mathcal{Y} \subset \mathbb{R}^d$, $K(x, y)$ can be identified with a $d \times d$ matrix. Further, under the normality assumption (Micchelli et al., 2006), $K(x, x)$ is positive definite for all $x \in \mathcal{X}$.

Example 2 (Matrix-valued kernels)**.**

1. *Separable kernels:* $K(x, x') = k(x, x')\mathbf{M}$ *where* $\mathbf{M} \in \mathbb{S}_+(\mathbb{R}^d)$ *encodes the dependence between the outputs. The setting presented in Section 3 corresponds to choosing* $\mathbf{M} = \mathbf{I}_d$ *and* $k(\cdot, \cdot)$ *the kernel associated to the sobolev space* $H^s(\mathcal{X}, \mathbb{R})$*, which already allows encoding functions in* $C^m(\mathcal{X}, \mathbb{R}^d)$ *for* s *large enough.*

2. *Non-separable kernels: e.g. divergence-free and curl-free kernels (Macêdo and Castro, 2008).*

Finally, in the proof of Theorem 15, which extends Theorem 5 to more general matrix-valued kernels, we require the following assumption on $\mathcal{H}_K$, which relates the norms of a vector-valued function $f \in \mathcal{H}_K$ to its component functions.

Assumption 3 (Properties of $\mathcal{H}$ and $\mathcal{H}_K$)**.** *$\mathcal{H}$ and $\mathcal{H}_K$ are respectively a scalar-valued and vector-valued RKHS satisfying:*

a) For $f = (f^{(1)}, ..., f^{(d)}) \in \mathcal{H}_K$, the $f^{(i)}$'s are in $\mathcal{H}$;

b) $\mathcal{H}$ satisfies Assumption 2 for some $m \in \mathbb{N}_+$;

c) There exists $B_K > 0$ such that $\sum_{i=1} \|f^{(i)}\|_{\mathcal{H}}^2 \leq B_K \|f\|_{\mathcal{H}_K}^2$.

In particular, Assumption 3c is satisfied when $K(x, x') = k(x, x')\mathbf{I}_d$, in which case $B_K = 1$. More generally, for a separable kernel $K(x, x') = k(x, x')\mathbf{M}$ (as in Example 2), we may derive bounds from the identity $\|f\|_{\mathcal{H}_K}^2 = \sum_{i,j=1}^d \mathbf{M}_{i,j}^\dagger \langle f^{(i)}, f^{(j)} \rangle_{\mathcal{H}}$, where $\mathbf{M}^\dagger$ is the pseudo-inverse of $\mathbf{M}$ (see Álvarez et al., 2012, Section 4.1).

A.2 Vector-Valued Sums-of-Squares

We model positive-operator-valued functions using the following representation:

$$F_A(x) = K_x^\star A K_x, \quad \text{with} \quad A \in \mathbb{S}_+(\mathcal{H}_K). \tag{31}$$

From the definitions above we have that $\forall x, F_A(x) \in \mathcal{L}(\mathcal{Y})$, F_A is self-adjoint and $\forall y \in \mathcal{Y}$, $\langle y, F_A(x) y \rangle = \langle y, K_x^\star A K_x y \rangle = \langle K_x y, A K_x y \rangle \geq 0$ by semi positive-definiteness of A, hence $\forall x \in \mathcal{X}, F_A(x) \in \mathbb{S}_+(\mathcal{Y})$. Equation (31) generalizes eq. (3): indeed, the latter can be recovered by taking $\mathcal{Y} \subset \mathbb{R}^d$, $K(x, x') = k(x, x')\mathbf{I}_d$ and $K_x = \Phi(x)$. When A admits a finite-dimensional representation (as obtained from Theorem 2) of the form

$$A = \sum_{i,j=1}^n K_{x_i} \mathbf{C}_{ij} K_{x_j}^*,$$

with $\mathbf{C}_{ij} \in \mathbb{R}^{d \times d}, i, j \in [n]$, and $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \dots & \mathbf{C}_{1n} \\ \vdots & & \vdots \\ \mathbf{C}_{n1} & \dots & \mathbf{C}_{nn} \end{pmatrix} \in \mathbb{S}_+(\mathbb{R}^{nd})$, then, A corresponds to a function of the form

$$F_C(x) = \sum_{i,j=1}^n K(x_i, x) \mathbf{C}_{ij} K(x_j, x).$$

Note that compared to eq. (7), the kernel terms $K(x_i, x)$ and $K(x_j, x)$ are $d \times d$ symmetric matrices that do not necessarily commute with $\mathbf{C}_{ij}$.

Further, as in Section 3.1, if $\mathbf{K} = \begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) \\ \vdots & & \vdots \\ K(x_n,x_1) & & K(x_n,x_n) \end{pmatrix} \in \mathbb{R}^{nd \times nd}$ and $\mathbf{R}$ denotes the upper-triangular Cholesky factor of $\mathbf{K}$ (verifying $\mathbf{R}^T \mathbf{R} = \mathbf{K}$), then $\tilde{\mathbf{R}} = \mathbf{R} \otimes \mathbf{I}_d$ is the

upper-triangular factor of $\tilde{\mathbf{K}}$. Letting $\mathbf{B} = \tilde{\mathbf{R}}\mathbf{C}\tilde{\mathbf{R}}^T$, we then have

$$\mathrm{Tr}A = \mathrm{Tr}\mathbf{B}$$

and

$$F_A(x_i) = \Psi_i^T \mathbf{B} \Psi_i, \quad i \in [n],$$

where Ψ_i denotes the i-th $nd \times d$ block column of $\tilde{\mathbf{K}}$. Further, the model can be queried at new points using the expression

$$F_{\mathbf{B}}(x) = \Psi(x)^T \mathbf{B} \Psi(x),$$

with $\Psi(x) = \tilde{\mathbf{R}}^{-T} v(x)$, and $v(x) = (K(x, x_i))_{i=1}^n \in \mathbb{R}^{nd \times d}$.

Appendix B. Proofs

For the sake of generality, we write the proofs of the results introduced in Section 3 within the matrix-valued kernel framework introduced in Appendix A. Taking $\mathcal{H}_K = \mathcal{H}^d$, $K_x = \Phi(x)$ (where $\Phi(x)$ is defined in eq. (4)) and $K(x, x') = k(x, x')\mathbf{I}_d$, we recover the setting of Section 3.

B.1 Proof of Proposition 1

Proof Let us prove linearity: let $A, B \in \mathcal{L}(\mathcal{H}_K)$ and $\alpha, \beta \in \mathbb{R}$. Since $\mathcal{L}(\mathcal{H}_K)$ is a vector space, it holds $\alpha A + \beta B \in \mathcal{L}(\mathcal{H}_K)$. Further, we have

$$\begin{aligned} F_{\alpha A + \beta B}(x) &= K_x^\star(\alpha A + \beta B) K_x \\ &= \alpha K_x^\star A K_x + \beta K_x^\star B K_x \\ &= \alpha F_A(x) + \beta F_B(x), \end{aligned}$$

which shows linearity w.r.t. the parameter A. Let us now assume that $A \in \mathbb{S}_+(\mathcal{H}_K)$, let us show that for all $x \in \mathcal{X}$, $F_A(x) \in \mathbb{S}_+(\mathbb{R}^d)$. First, let us observe that since $A \in \mathbb{S}_+(\mathcal{H}_K)$ and $K_x \in \mathcal{L}(\mathbb{R}^d, \mathcal{H}_K)$ (as defined in Appendix A), we have $F_A(x) \in \mathcal{L}(\mathcal{H}_K)$. Further, $F_A(x)$ is self-adjoint:

$$F_A(x)^\star = K_x^\star A^\star K_x = K_x^\star A K_x = F_A(x).$$

Let us finally show that $F_A(x)$ is positive semi-definite: let $v \in \mathbb{R}^d$, we have

$$\langle v,\, F_A(x) \rangle = \langle K_x v,\, A K_x v \rangle_{\mathcal{H}_K} \geq 0,$$

by positive semi-definiteness of A on $\mathcal{H}_K$. ■

B.2 Proof of Theorem 2

Following Marteau-Ferey et al. (2020), we divide the proof of Theorem 2 in two main results: Proposition 11 and Lemma 12. Let us start with a few definitions:

- Let $\mathcal{H}_K^n = \mathrm{Span}\{K_{x_i}\}_{i=1}^n = \{\sum_{i=1}^n K_{x_i} y_i : y_i \in \mathbb{R}^d, i \in [n]\}$;
- Π_n denotes the orthogonal projection on $\mathcal{H}_K^n$;

- Finally, define $S_n(\mathcal{H}_K)_+ = \{\Pi_n A \Pi_n : A \in \mathbb{S}_+(\mathcal{H}_k)\} \subset \mathbb{S}_+(\mathcal{H}_K)$.

Let $\mathcal{S}_n : \mathcal{H}_K \to \mathbb{R}^{n\times d}$ defined as

$$\mathcal{S}_n(h) = \left(K^\star_{x_i} h\right)_{1\le i\le n}, \quad h \in \mathcal{H}_K.$$

Its adjoint is

$$\mathcal{S}_n^*(\alpha) = \sum_{i=1}^n K_{x_i}\alpha_i, \alpha \in \mathbb{R}^{n\times d}.$$

Further, define $\mathbf{K} = \begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) \\ \vdots & & \vdots \\ K(x_n,x_1) & & K(x_n,x_n) \end{pmatrix} \in \mathbb{R}^{nd\times nd}$, with rank r ($r = nd$ in the case of a universal kernel (Micchelli et al., 2006)). Let $\mathbf{V} \in \mathbb{R}^{nd\times nd}$ be a matrix such that $\mathbf{V}^T\mathbf{V} = \mathbf{K}$, and define $O_n : \mathbb{R}^n \to \mathcal{H}_K$ as $O_n = S_n^*\mathbf{V}(\mathbf{V}\mathbf{V}^T)^{-1}$. We start with a technical result that will be useful to prove Lemma 12.

Lemma 10 (Marteau-Ferey et al. (2020), Lemma 4)**.** *It holds (i) $O_nO_n^* = \Pi_n$ and (ii) $O_n^*O_n = \mathrm{Id}_r$.*

Proof The proof is identical to that of Marteau-Ferey et al. (2020), Lemma 4. We restate it, for the sake of self-inclusiveness. Since $\mathbf{V}^T\mathbf{V} = \mathbf{K} = \mathcal{S}_n\mathcal{S}_n^*$, we have

$$O_nO_n^* = (\mathbf{V}\mathbf{V}^T)^{-1}\mathbf{V}S_nS_n^*\mathbf{V}^T(\mathbf{V}\mathbf{V}^T)^{-1} = (\mathbf{V}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{V}^T\mathbf{V}\mathbf{V}^T(\mathbf{V}\mathbf{V}^T)^{-1} = \mathbf{I}_r,$$

which proves (ii). Let us now prove (i). Let $\tilde{\Pi}_n = O_nO_n^*$. $\tilde{\Pi}_n$ is a projection operator, since it is self-adjoint and satisfies $\tilde{\Pi}_n^2 = O_n(O_n^*O_n)O_n^* = O_nO_n^* = \tilde{\Pi}_n$, from (ii). Moreover, it follows from the definition of O_n that the range of Π_n is included in the range of $\mathcal{S}_n^*$, which is $\mathcal{H}_K^n$. Finally, we observe that $\operatorname{rk}\mathcal{S}_n^* = \operatorname{rk}\mathcal{S}_n\mathcal{S}_n^* = r$, which implies that the dimension of $\mathcal{H}_K^n$ is r. Since $\operatorname{rk}O_nO_n^* = \operatorname{rk}O_n^*O_n = r$ from (ii), this implies in turn that $\tilde{\Pi}_n = \Pi_n$. ■

We are now ready to prove the main part of Theorem 2.

Proposition 11 (Marteau-Ferey et al. (2020), Proposition 7)**.** *Let L be a l.s.c. function which is bounded below, and Ω as in eq.* (6)*. Then, eq.* (5) *has a solution $A^\star$ that is in $\mathcal{S}_n(\mathcal{H}_K)_+$.*

Proof

Step 1. For $A \in \mathbb{S}_+(\mathcal{H}_K)$, let us denote $J(A) = L(F_A(x_1), ..., F_A(x_n)) + \Omega(A)$, the objective of (5). We start by showing that

$$\inf_{A\in\mathcal{S}_n(\mathcal{H}_K)_+} J(A) = \inf_{A\in\mathbb{S}_+(\mathcal{H}_K)} J(A).$$

Fix $A \in \mathbb{S}_+(\mathcal{H}_K)$. We will show that $J(\Pi_nA\Pi_n) \le J(A)$. First, observe that since Π_n is the orthogonal projection on $\mathrm{Span}\{K_{x_i}, i \in [n]\}$, it holds $K_{x_i} = \Pi_nK_{x_i}$, and therefore

$$F_A(x_i) = K^\star_{x_i}AK_{x_i} = K^\star_{x_i}\Pi_nAK_{x_i} = F_{\Pi_nA\Pi_n}(x_i), \quad i \in [n].$$

In particular, this implies that $L(F_A(x_1), ..., F_A(x_n)) = L(F_{\Pi_n A \Pi_n}(x_1), ..., F_{\Pi_n A \Pi_n}(x_n))$. Next, let us observe that since $\mathrm{Tr} A = \sum_{i \in \mathbb{N}} \sigma_i$ and $\|A\|_{\mathrm{HS}}^2 = \sum_{i \in \mathbb{N}} \sigma_i^2$ where $\sigma_i, i \in [n]$ are the eigenvalues of A, we have $\Omega(\Pi_n A \Pi_n) \leq \Omega(A)$. Putting everything together, this implies that $\forall A \in \mathbb{S}_+(\mathcal{H}_K), J(\Pi_n A \Pi_n) \leq J(A)$. Therefore, we have

$$\inf_{A \in \mathcal{S}_n(\mathcal{H}_K)_+} J(A) \leq \inf_{A \in \mathbb{S}_+(\mathcal{H}_K)} J(A).$$

Since $\mathcal{S}_n(\mathcal{H}_K)_+ \subset \mathbb{S}_+(\mathcal{H}_K)$, this finally implies

$$\inf_{A \in \mathcal{S}_n(\mathcal{H}_K)_+} J(A) = \inf_{A \in \mathbb{S}_+(\mathcal{H}_K)} J(A).$$

Step 2. Let us now show that $\inf_{A \in \mathcal{S}_n(\mathcal{H}_K)_+} J(A)$ has a solution. Again, we follow the steps of Marteau-Ferey et al. (2020). Let V_n be the injection $\mathcal{H}_K^n \hookrightarrow \mathcal{H}_K$. It holds $V_n V_n^* = \Pi_n$ and $V_n^* V_n = \mathbf{I}_{\mathcal{H}_K^n}$. Hence, we have

$$\mathcal{S}_n(\mathcal{H}_K)_+ = V_n \mathbb{S}_+(\mathcal{H}_K^n) V_n^* = \{V_n A V_n^* : A \in \mathbb{S}_+(\mathcal{H}_K^n)\}.$$

Let us now show that $\inf_{A \in \mathbb{S}_+(\mathcal{H}_K^n)} J(V_n A V_n^*)$ has a solution. Let us first observe that $\forall A \in \mathbb{S}_+(\mathcal{H}_K^n)$, we have $\Omega(V_n A V_n^*)$ (see Marteau-Ferey et al., 2020, Lemma 2), and thus $J(V_n A V_n^*) = L(F_{V_n A V_n^*}(x_1), ..., F_{V_n A V_n^*}(x_n)) + \Omega(A)$. Let $A_0 \in \mathbb{S}_+(\mathcal{H}_K^n)$ be such that $J_0 := J(V_n A_0 V_n^*) < \infty$, and let c_0 be a lower bound of L. From the eq. (6), there exists $R_0 > 0$ such that $\|A\|_{\mathrm{HS}} > 0 \implies \Omega(A) > J_0 - c_0$. This implies

$$\inf_{A \in \mathbb{S}_+(\mathcal{H}_K^n)} J(V_n A V_n^*) = \inf_{A \in \mathbb{S}_+(\mathcal{H}_K^n)} J(V_n A V_n^*) \ \text{ s.t. } \ \|A\|_{\mathrm{HS}} \leq R_0.$$

Since L and Ω are l.s.c., so is $A \mapsto J(V_n A V_n^*)$. Hence, it reaches its minimum on the compact set $\{A \in \mathbb{S}_+(\mathcal{H}_K^n) : \|A\|_{\mathrm{HS}} \leq R_0\}$, which is non-empty as it contains A_0. Hence, there exists $\tilde{A}_\star \in \mathbb{S}_+(\mathcal{H}_K^n)$ such that

$$J(V_n \tilde{A}_\star) = \inf_{A \in \mathbb{S}_+(\mathcal{H}_K^n)} J(V_n A V_n^*) \ \text{ s.t. } \ \|A\|_{\mathrm{HS}} \leq R_0,$$

which shows that $\inf_{A \in \mathbb{S}_+(\mathcal{H})} J(A) = J(A_\star)$ with $A_\star = V_n \tilde{A}_\star \in \mathcal{S}_n(\mathcal{H}_J)_+$. ■

To complete the proof of the Theorem, there remains to show the following lemma.

Lemma 12.

$$\mathcal{S}_n(\mathcal{H}_K)_+ = \left\{ \sum_{i,j=1}^n K_{x_i} \mathbf{B}_{ij} K_{x_j}^* : \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \dots & \mathbf{B}_{1n} \\ \vdots & & \vdots \\ \mathbf{B}_{n1} & \dots & \mathbf{B}_{nn} \end{pmatrix} \in \mathbb{S}_+(\mathbb{R}^{nd}) \right\}.$$

Proof We proceed by double inclusion. Recall the definitions of $\mathcal{S}_n$ and $\mathcal{S}_n^*$:

$$\mathcal{S}_n(h) = \left(K_{x_i}^\star h\right)_{1 \leq i \leq n}, \quad h \in \mathcal{H}_K,$$

and

$$\mathcal{S}_n^*(\alpha) = \sum_{i=1}^n K_{x_i}\alpha_i, \alpha \in \mathbb{R}^{n\times d}.$$

In particular, for $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \dots & \mathbf{B}_{1n} \\ \vdots & & \vdots \\ \mathbf{B}_{n1} & \dots & \mathbf{B}_{nn} \end{pmatrix} \in \mathbb{R}^{nd\times nd}$, it holds

$$\mathcal{S}_n^\star \mathbf{B} \mathcal{S}_n = \sum_{ij=1}^n K_{x_i}\mathbf{B}_{ij}K_{x_j}^\star. \tag{32}$$

a) $S_n(\mathcal{H}_K)_+ \subset \left\{ \sum_{i,j=1}^n K_{x_i}\mathbf{B}_{ij}K_{x_j}^* : \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \dots & \mathbf{B}_{1n} \\ \vdots & & \vdots \\ \mathbf{B}_{n1} & \dots & \mathbf{B}_{nn} \end{pmatrix} \in \mathbb{S}_+(\mathbb{R}^{nd}) \right\}$.
Let $\Pi_n A \Pi_n \in \mathcal{S}_n(\mathcal{H}_K)_+$, our goal is to show that there exists $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^d)$ such that $\Pi_n A \Pi_n = \mathcal{S}_n^\star \mathbf{B} \mathcal{S}_n$. Using Lemma 10, we have that Π_n can be written $\mathcal{S}_n^\star T_n$ with $T_n \in \mathcal{L}(\mathcal{H}_K, \mathbb{R}^{n\times d})$. Hence, defining $\mathbf{B} = T_n A T_n^\star$, we have $\mathcal{S}_n^\star \mathbf{B} \mathcal{S} = \Pi_n A \Pi_n$. Further, $A \succeq 0 \implies \mathbf{B} \succeq 0$. Given eq. (32), this shows the inclusion.

b) $\left\{ \sum_{i,j=1}^n K_{x_i}\mathbf{B}_{ij}K_{x_j}^* : \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \dots & \mathbf{B}_{1n} \\ \vdots & & \vdots \\ \mathbf{B}_{n1} & \dots & \mathbf{B}_{nn} \end{pmatrix} \in \mathbb{S}_+(\mathbb{R}^{nd}) \right\} \subset S_n(\mathcal{H}_K)_+$.
Let $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})$. It holds $A := \mathcal{S}_n^\star \mathbf{B} \in \mathbb{S}_+(\mathcal{H}_K)$. Further, since the range of $\mathcal{S}_n^\star$ is included in $\mathcal{H}_K^n$, it holds $\Pi_n S_n^* = \Pi_n$, and therefore $A \in \mathcal{S}_n(\mathcal{H}_K)_+$.

This concludes the proof. ■

Proof of the theorem. The proof of Theorem 2 follows by combining the two results above: from Proposition 11, we have that eq. (5) has a solution in $A^* \in S_n(\mathcal{H}_K)+$, which is unique if L is convex and $\lambda_2 > 0$. By Lemma 12, this solution may be written

$$A^* = \sum_{i,j=1}^n K_{x_i}\mathbf{B}_{ij}K_{x_j}^*, \tag{33}$$

for some $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \dots & \mathbf{B}_{1n} \\ \vdots & & \vdots \\ \mathbf{B}_{n1} & \dots & \mathbf{B}_{nn} \end{pmatrix} \in \mathbb{S}_+(\mathbb{R}^{nd})$. Further, if L is convex and $\lambda_2 > 0$, then the objection function of eq. (5) is strongly convex, and hence the minimizer is unique. Finally, plugging eq. (33) in eq. (31), for $x \in \mathcal{X}$ we have

$$\begin{aligned} F_{A^*}(x) &= K_x^* A^* K_x \\ &= \sum_{i,j=1}^n K_x^* K_{x_i} \mathbf{B}_{ij} K_{x_j}^* K_x \\ &= \sum_{i,j=1}^n K(x_i, x)\mathbf{B}_{ij}K(x_j, x) =: F_{\mathbf{B}}(x). \end{aligned}$$

Equation (7) is then a particularization of the expression above to the case $K(x, x') = k(x, x')\mathbf{I}_d$, where $k(\cdot, \cdot)$ is a scalar-valued kernel.

B.3 Proof of Theorem 3

We start by restating the following useful lemma.

Lemma 13 (Marteau-Ferey et al. (2020), Lemma 5)**.** *Let $\lambda_1, \lambda_2 \geq 0$. For $\mathbf{B} \in \mathrm{Sym}(\mathbb{R}^d)$, define*

$$\Omega(\mathbf{B}) = \begin{cases} \lambda_1 \mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2 & \text{if } \mathbf{B} \succeq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

(i) Assume $\lambda_2 > 0$. Then Ω is a closed convex function, whose Fenchel conjugate is given for $\mathbf{B} \in \mathrm{Sym}(\mathbb{R}^d)$ by

$$\Omega^\star(\mathbf{B}) = \frac{1}{2\lambda_2}\|[\mathbf{B} - \lambda_1 \mathbf{I}_d]_+|_F^2,$$

where $[\mathbf{B}]_+$ denotes the positive part of $\mathbf{B}$. $\Omega^\star$ is differentiable and $\frac{1}{\lambda_2}$-smooth. Its gradient is given by

$$\nabla\Omega^\star(\mathbf{B}) = \frac{1}{\lambda_2}[\mathbf{B} - \lambda_1 \mathbf{I}_d]_+.$$

(ii) Assume $\lambda_2 = 0$ and $\lambda_1 > 0$. Then Ω is a closed convex function, whose Fenchel conjugate is given for $\mathbf{B} \in \mathrm{Sym}(\mathbb{R}^d)$ by

$$\Omega^\star(\mathbf{B}) = \begin{cases} 0 & \text{if } \lambda_1 \mathbf{I}_d \succeq \mathbf{B}, \\ +\infty & \text{otherwise.} \end{cases}$$

Proof (i): this is exactly Marteau-Ferey et al. (2020)'s Lemma 5.
(ii): Let us adapt the proof Marteau-Ferey et al. (2020)'s Lemma 5. We may write

$$\Omega(\mathbf{B}) = \iota_{\mathbb{S}_+(\mathbb{R}^d)} + \lambda_1 \mathrm{Tr}\mathbf{B},$$

where $\iota_{\mathbb{S}_+(\mathbb{R}^d)}(\mathbf{B}) = 0$ if $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^d)$ and $+\infty$ otherwise. Since the trace is linear, it is in particular convex, and continuous. Further, $\iota_{\mathbb{S}_+(\mathbb{R}^d)}$ is closed since $\mathbb{S}_+(\mathbb{R}^d)$ is a closed non-empty convex subset of $\mathrm{Sym}(\mathbb{R}^d)$. This shows that Ω is closed and convex, and linear on its domain $\mathbb{S}_+(\mathbb{R}^d)$. Fix $\mathbf{B} \in \mathrm{Sym}(\mathbb{R}^d)$, we have

$$\begin{aligned} \Omega^*(\mathbf{B}) &\overset{\text{def}}{=} \sup_{\mathbf{A} \in \mathrm{Sym}(\mathbb{R}^d)} \mathrm{Tr}(\mathbf{A}\mathbf{B}) - \Omega(\mathbf{A}) \\ &= \sup_{\mathbf{A} \in \mathbb{S}_+(\mathbb{R}^d)} \mathrm{Tr}(\mathbf{A}(\mathbf{B} - \lambda_1 \mathbf{I}_d)) \\ &= \begin{cases} 0 & \text{if } \mathbf{B} - \lambda_1 \mathbf{I}_d \preceq 0, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

■

Proof [Theorem 3] Like Marteau-Ferey et al. (2020), we apply Theorem 3.3.5 of Borwein and Lewis (2006) with corresponding arguments represented in Table 1.

Indeed, the following properties are satisfied:

$\mathbf{E}$	$\mathrm{Sym}(\mathbb{R}^{nd})$
$\mathbf{Y}$	$\mathrm{Sym}(\mathbb{R}^d)^n$
$A : \mathbf{E} \to \mathbf{Y}$	$R : \mathbf{B} \in \mathrm{Sym}(\mathbb{R}^{nd}) \mapsto (F_{\mathbf{B}}(x_1), ..., F_{\mathbf{B}}(x_n)) \in \mathrm{Sym}(\mathbb{R}^d)^n$
$f : \mathbf{E} \to (-\infty, +\infty]$	$\Omega : \mathrm{Sym}(\mathbb{R}^{nd}) \to (-\infty, +\infty]$
$g : \mathbf{Y} \to (-\infty, +\infty]$	$L : \mathrm{Sym}(\mathbb{R}^d)^n \to (-\infty, +\infty]$
$p = \inf_{x \in \mathbf{E}} g(Ax) + f(x)$	$p = \inf_{\mathbf{B} \in \mathrm{Sym}(\mathbb{R}^{nd})} L(F_{\mathbf{B}}(x_1), ..., F_{\mathbf{B}}(x_n)) + \Omega(\mathbf{B})$
$d = \sup_{\phi \in \mathbf{Y}} -g^*(\phi) - f^*(-A^*\phi)$	$d = \sup_{\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n} -L^*(\Gamma) - \Omega^*(-R^*(\Gamma))$

Table 1: Corresponding arguments in Theorem 3.3.5 of Borwein and Lewis (2006) *(left)* and Theorem 3 *(right)*

- L is l.s.c., convex, and bounded below, which implies that it is closed (Borwein and Lewis, 2006, see);
- Ω is a non-negative closed convex function, with dual $\Omega^\star$ given in Lemma 13, which is differentiable and smooth when $\lambda_2 > 0$ (case (i));
- The domain of Ω is $\mathbb{S}_+(\mathbb{R}^{nd})$;
- R is linear, and for $\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n$, it holds $R^*\Gamma = \sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T$;
- Using the expressions of Ω^* and R^*, we may reformulate the dual d in Table 1.

  (i) When $\lambda_2 > 0$ and $\lambda_1 \geq 0$:

$$d = \sup_{\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n} -L^\star(\Gamma) - \frac{1}{2\lambda_2} \| [\sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T + \lambda_1 \mathbf{I}_{nd}]_- \|_F^2,$$

  (ii) When $\lambda_2 = 0$ and $\lambda_1 > 0$:

$$d = \sup_{\Gamma \in \mathrm{Sym}(\mathbb{R}^d)^n} -L^\star(\Gamma) \;\text{ s.t. }\; \sum_{i=1}^n \Psi_i \Gamma^{(i)} \Psi_i^T + \lambda \mathbf{I}_{nd} \succeq 0;$$

  (iii) Assume there exists $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})$ such that L is continuous in $(F_B(x_i))_{1 \leq i \leq n}$. Then there exists a point of continuity of g which is also in $R(\mathrm{dom}(f))$.

We may thus apply theorem 3.3.5 of Borwein and Lewis (2006) to get:

- $d = p$;
- d is attained for a certain $\Gamma_* \in \mathrm{Sym}(\mathbb{R}^d)^n$. Indeed, there exists $\mathbf{B} \in \mathrm{dom}(\Omega)$ such that $R(\mathbf{B}) \in \mathrm{dom}(L)$. Thus, $L(R(\mathbf{B})) + \Omega(\mathbf{B}) < +\infty$ 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 attained by theorem 3.3.5.

Finally, if $\lambda_2 > 0$ and $\lambda_1 \geq 0$ (case (i)), then Ω^* is differentiable and since L and Ω are closed and convex, we may use Exercise 4.2.17 of Borwein and Lewis (2006) to show that the primal solution $\mathbf{B}_\star$ is given by

$$\mathbf{B}^\star = \nabla\Omega^*(-R^*\Gamma_*) = \frac{1}{\lambda_2}\Psi^{-1}[\sum_{i=1}^{n}\Psi_i\Gamma^{(i)}\Psi_i^T + \lambda_1\mathbf{I}_{nd}]_-\Psi^{-T}.$$

■

B.4 Proof of Theorem 4

We start by restating Theorem 4 under a more general form that is compatible with the vector-valued RKHSs introduced in Appendix A.

Theorem 14 (Universality). *Let $\mathcal{H}_K$ be a separable Hilbert space of functions taking values in $\mathbb{R}^d$ and $K : x \in \mathcal{X} \to K_x \in \mathcal{H}_K$ a universal map (Micchelli et al., 2006, see). Then*

$$\{F_A : x \mapsto K_x^\star A K_x \quad s.t. \quad A \in \mathbb{S}_+(\mathcal{H}^d), \mathrm{Tr}A < \infty\}$$

is a universal approximator of $d \times d$ PSD-valued functions over $\mathcal{X}$.

Proof Let $\mathcal{Z} \subset \mathcal{X}$ be a compact set, and $G \in \mathcal{C}(\mathcal{Z}, \mathbb{S}_+(\mathbb{R}^d))$ be a continuous PSD-valued function. For $x \in \mathcal{Z}$, let $R(x)$ denote the PSD square-root of $G(x)$. Since the matrix square root is continuous on $\mathbb{S}_+(\mathbb{R}^d)$ (Horn and Johnson, 2012), $R \in \mathcal{C}(\mathcal{Z}, \mathbb{S}_+(\mathbb{R}^d))$. For $i, j \in [d]$, define $r_{ij}(x) = e_i^T R(x) e_j$ (where $\{e_i, i \in [d]\}$ is the canonical ONB of $\mathbb{R}^d$), and $w_i(x) = \sum_{j=1}^d r_{ij}(x)e_j$. Then, $\forall i \in [d], w_i \in \mathcal{C}(\mathcal{Z}, \mathbb{R}^d)$. Let $\varepsilon > 0$ and $Q = d^2\sum_{i=1}^d(\varepsilon + 2\|w_i\|_\mathcal{Z})$. By universality of $x \in \mathcal{X} \to K_x \in \mathcal{H}_K$, there exist $f_i \in \mathcal{H}_K, i \in [d]$ such that $\|f_i - g\|_\mathcal{Z} \leq \frac{\varepsilon}{Q}, i \in [d]$ (as $f_i(x) = K_x^\star f_i$). Let $A = \sum_{i=1}^d f_i \otimes f_i \in \mathbb{S}_+(\mathcal{H}_K)$, and fix $x \in \mathcal{Z}$. It holds

$$\begin{aligned}
\|F_A(x) - G(x)\|_\infty &\leq d\|F_A(x) - G(x)\|_F \\
&= d\|\sum_{i=1}^d f_i(x)f_i(x)^T - w_i(x)w_i(x)^T\|_F \\
&\leq d\sum_{i=1}^d \|f_i(x)f_i(x)^T - w_i(x)w_i(x)^T\|_F \\
&\leq d\sum_{i=1}^d \|f_i(x) + w_i(x)\|_2\|f_i(x) - w_i(x)\|_2 \\
&\leq d^2\sum_{i=1}^d \|f_i(x) + w_i(x)\|_\infty\|f_i(x) - w_i(x)\|_\infty \\
&\leq d^2\sum_{i=1}^d (\varepsilon + 2\|w_i\|_\mathcal{Z})\|f_i - w_i\|_\mathcal{Z} \\
&\leq \frac{\varepsilon}{Q}d^2\sum_{i=1}^d(\varepsilon + 2\|w_i\|_\mathcal{Z}) \leq \varepsilon.
\end{aligned}$$

Since this bound is independent from $x \in \mathcal{Z}$, this shows universality. ■

B.5 Proof of Theorem 5

We prove a slightly more general version of Theorem 5 that is adapted to general vector-valued RKHSs. As for the results above, the original statement can be recovered by taking $K(x,x') = k(x,x')\mathbf{I}_d$, for which Assumption 3 is satisfied with $B_K = 1$.

Theorem 15. *Let $\mathcal{X}$ satisfy Assumption 1 for some $r > 0$. Let K be a matrix-valued kernel satisfying Assumption 3. Let $\hat{X} = \{x_1, ..., x_n\} \subset \mathcal{X}$ such that $h_{\hat{X},\mathcal{X}} \overset{\text{def}}{=} \sup_{x\in\mathcal{X}} \min_{i\in[n]} \|x - x_i\| \leq r \min\left(1, \frac{1}{18(m-1)^2}\right)$. Let $F \in C^m(\mathcal{X}, \mathrm{Sym}(\mathbb{R}^d))$ and assume there exists $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^n)$ such that*

$$F(x_i) = \Psi_i^T \mathbf{B} \Psi_i, \quad i \in [n],$$

where the Ψ_i's are defined in Section 3. Then, it holds

$$\forall x \in \mathcal{X}, \quad \lambda_{\min}(F(x)) \geq -\varepsilon, \quad \text{where} \quad \varepsilon = C h^m_{\hat{X},\mathcal{X}},$$

and $C = C_0(\lambda_{\max}(\mathbf{N}_F) + MD_mB_K\mathrm{Tr}\mathbf{B})$ with $C_0 = 3\frac{p^m}{m!}\max(1, 18(m-1)^2)^m$, and where $\mathbf{N}_F \in \mathbb{R}^{d\times d}$ denotes the matrix whose entries are the pseudo norms of the entries of F, $[\mathbf{N}_F]_{ij} = |F_{ij}|_{\mathcal{X},m}, 1 \leq i,j \leq d$.

Proof

Step 1. Let $u \in \mathcal{S}^{d-1}$, and let $g_u(x) = \langle u, (F(x) - K_x^* A K_x)\, u\rangle$. From the assumptions, $g_u \in C^m(\mathcal{X})$. Applying Theorem 13 from Rudi et al. (2020), we thus have

$$\sup_{x\in\mathcal{X}} |g_u(x)| \leq \varepsilon_u, \quad \varepsilon_u = cR_m(g_u)h^m_{\hat{X},\mathcal{X}},$$

where $c = 3\max(1, 18(m-1)^2)^m$ and $R_m(g_u) = \sum_{|\alpha|=m} \frac{1}{\alpha!} \sup_{x\in\mathcal{X}} |\partial^\alpha g_u(x)| \leq \frac{p^m}{m!}|g_u|_{\mathcal{X},m}$, with

$$|g_u|_{\mathcal{X},m} \overset{\text{def}}{=} \max_{|\alpha|=m} \sup_{x\in\mathcal{X}} |\partial^\alpha g_u(x)|.$$

Step 2: uniform bound on $|g_u|_{\hat{X},\mathcal{X}}, u \in \mathcal{S}^{d-1}$.

1) Let $F_u(x) = \langle u, F(x)u\rangle$. We have $F_u \in C^m(\mathcal{X})$. Since $F_u(x) = \sum_{1\leq i,j\leq d} u_i u_j F_{ij}(x)$, we have

$$\begin{aligned}
|F_u|_{\mathcal{X},m} &= \max_{|\alpha|=m} \sup_{x\in\mathcal{X}} |\sum_{1\leq i,j\leq d} u_i u_j \partial^\alpha F_{ij}(x)| \\
&\leq \max_{|\alpha|=m} \sup_{x\in\mathcal{X}} \sum_{1\leq i,j\leq d} |u_i||u_j||\partial^\alpha F_{ij}(x)| \\
&\leq \sum_{1\leq i,j\leq d} |u_i||u_j||F_{ij}|_{\mathcal{X},m} \\
&\leq \lambda_{\max}(\mathbf{N}_F)
\end{aligned}$$

where $\mathbf{N}_F \in \mathbb{R}^{d\times d}$, $[\mathbf{N}_F]_{ij} = |F_{ij}|_{\mathcal{X},m}, 1 \leq i,j \leq d$ denotes the matrix whose entries are the pseudo norms of the entries of F. Hence $\sup_{u\in S^{d-1}} |F_u|_{\mathcal{X},m} \leq \lambda_{\max}(\mathbf{N}_F)$.

2) Let $r_{A,u}(x) = \langle u, K_x^* A K_x u \rangle$. We have $r_{A,u} \in C^m(\mathcal{X})$. Hence, from assumptions on k, we have $|r_{A,u}|_{\mathcal{X},m} \leq D_m \|r_{A,u}\|_{\mathcal{H}}$ (see Rudi et al., 2020, Remark 2). Let us now bound $\|r_{A,u}\|_{\mathcal{H}}$. A admits an eigen-decomposition $A = \sum_{p\in\mathbb{N}} \sigma_p f_p \otimes f_p$, with $\{f_p, p \in \mathbb{N}\}$ an orthonormal family of $\mathcal{H}_K$, and $\sigma_p, p \in \mathbb{N}$ is a non-increasing sequence of non-negative scalars converging to zero. Hence, we have

$$r_{A,u}(x) = \sum_{p\in\mathbb{N}} \sigma_p u^T f_p(x) u^T f_p(x).$$

Hence, by Assumption 3c,

$$\begin{aligned}\|r_{A,u}\|_{\mathcal{H}} &\leq \sum_{p\in\mathbb{N}} \sigma_p \|(u^T f_p(\cdot))^2\|_{\mathcal{H}} \\ &\leq M \sum_{p\in\mathbb{N}} \sigma_p \|u^T f_p(\cdot)\|_{\mathcal{H}}^2 \\ &\leq M B_K \mathrm{Tr} A,\end{aligned}$$

Indeed, for $f(x) = (f^{(1)}(x), ..., f^{(d)}(x))) \in \mathcal{H}_K$, we have

$$\begin{aligned}\|u^T f\|_{\mathcal{H}} &\leq \sum_{s=1}^{d} \|u_s f^{(s)}\|_{\mathcal{H}} \\ &\leq \sum_{s=1}^{d} |u_s| \|f^{(s)}\|_{\mathcal{H}}.\end{aligned}$$

Since $u \in \mathcal{S}^{d-1}$, the last line is maximized when $|u_s| = \frac{\|f^{(s)}\|_{\mathcal{H}}}{\left(\sum_{i=1}^d \|f^{(i)}\|_{\mathcal{H}}^2\right)^{1/2}}$. If f satisfies $\|f\|_{\mathcal{H}_K}^2 = 1$, this yields

$$\|u^T f\|_{\mathcal{H}} \leq \left(\sum_{i=1}^{d} \|f^{(i)}\|_{\mathcal{H}}^2\right)^{1/2} \leq B_K \|f\|_{\mathcal{H}_K} = B_K.$$

Hence we have $\sup_{x\in\mathcal{X}} |g_u(x)| \leq \tau$ for $\tau = \lambda_{\max}(\mathbf{N}_F) + M D_m B_K \mathrm{Tr} A$ that does not depend on u.

Step 3: Conclusion. Putting everything together, we have that

$$\forall x \in \mathcal{X}, \lambda_{\min} F(x) = \min_{u\in\mathcal{S}d-1} F_u(x) \geq -C_0(\lambda_{\max}(\mathbf{N}_F) + M D_m B_K \mathrm{Tr} A) h_{\hat{X},\mathcal{X}}^m,$$

with $C_0 = 3\frac{p^m}{m!}\max(1, 18(m-1)^2)^m$. In the statement of Theorem 15, we have further bounded $\lambda_{\max}(\mathbf{N}_F)$ from above by $\mathrm{Tr}\mathbf{N}_F = \sum_{i=1}^d |F_{ii}|_{\mathcal{X},m}$ for simplicity. ■

Appendix C. Modeling Convex Functions: Proofs

C.1 Proof of Theorem 6

We start by restating Theorem 6 under a slightly more general form, that allows matrix-valued kernels.

Theorem 16 (PSD sum-of-squares representation for convex functions)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{H}, \mathcal{H}_K$ be RKHSs of $\mathbb{R}$-valued and $\mathbb{R}^d$-valued functions on $\mathcal{X}$ satisfying Assumption 3, with $\mathcal{C}^s(\mathcal{X}) \subset \mathcal{H}$ for some $s \in \mathbb{N}, s > d/2$. Let $f \in \mathcal{C}^{s+2}(\mathcal{X})$ be strongly convex. Then, there exists $A \in \mathbb{S}_+(\mathcal{H}^d)$ such that $H_f(x) = F_A(x), \forall x \in \mathcal{X}$.*

Proof Since f is assumed strongly convex, there exists $\rho > 0$ such that $\forall x, \mathbf{H}_f(x) \succeq \frac{\rho}{2}\,\mathrm{Id}$. In particular, for all $x \in X$, the matrix $\mathbf{H}_f(x)$ admits a positive square root $\sqrt{\mathbf{H}_f(x)}$. Further, since $\sqrt{\cdot}$ is C^∞ on the closed set $\{\mathbf{A} \in \mathbb{S}_+(\mathbb{R}^d) : \mathbf{A} \succeq \rho\,\mathrm{Id}\}$ and $\frac{\partial^2 f}{\partial x_i \partial x_j} \in \mathcal{C}^s(\mathcal{X})$ for all $i, j \in [d]$, the functions $r_{i,j} : x \mapsto e_i^\top \sqrt{\mathbf{H}_f(x)} e_j$ are in $C^s(\mathcal{X})$ for all $i, j \in [d]$ (see Proposition 1 and Assumption 2b of Rudi et al., 2020), where $(e_1, ..., e_d)$ is the canonical ONB of $\mathbb{R}^d$. Define now the functions

$$w_i(x) \overset{\text{def}}{=} \sum_{j=1}^{d} r_{i,j}(x) e_j, \quad \forall (x, u) \in X \times \mathcal{S}^{d-1}, i \in [d].$$

From the above arguments, it holds that $w_i \in \mathcal{H}, i \in [d]$, and

$$\mathbf{H}_f(x) = \sum_{i=1}^{d} w_i(x) w_i(x)^T, \quad \forall x \in \mathcal{X}.$$

Following Assumption 2, we have that the $r_{i,j}$'s belong to $\mathcal{H}$. Hence, by Assumption 3, it holds $w_i(\cdot) \in \mathcal{H}_K, i \in [d]$. Finally, defining $A = \sum_{i=1}^{d} w_i \otimes w_i \in \mathbb{S}_+(\mathbb{R}^d)$, we have

$$\begin{aligned} F_A(x) &= K_x^\star A K_x \\ &= \sum_{i=1}^{d} K_x^\star (w_i \otimes w_i) K_x \\ &= \sum_{i=1}^{d} (K_x^\star w_i) \otimes (K_x^\star w_i) \\ &= \sum_{i=1}^{d} w_i(x) \otimes w_i(x) = \mathbf{H}_f(x), \end{aligned}$$

which concludes the proof. ■

C.2 Proof of Corollary 7

Proof This is a direct consequence of Theorem 5. Indeed, it holds $H_f(x_i) = F_{\mathbf{B}}(x_i), i \in [n]$, and H_f satisfies the hypothesis of Theorem 5. Hence, it holds

$$\forall x \in \mathcal{X}, \lambda_{\min}(H_f(x)) \geq -\eta \text{ with } \eta = C h_{\hat{X},\mathcal{X}}^m > 0,$$

and $C = C_0(\sum_{i=1}^d |(H_f)_{ii}|_{\mathcal{X},m} + MD_mB_K\mathrm{Tr}\mathbf{B})$ with $C_0 = 3\frac{d^m}{m!}\max(1, 18(m-1)^2)^m$. In particular, this implies that f is $-\eta$ strongly convex. ■

C.3 Proof of Theorem 8

Proof Let us consider the following problems:

$$\min_{\substack{f\in\mathcal{H}\\ A\in\mathbb{S}_+(\mathcal{H}^d)}} J(f) \;\text{ s.t. }\; H_f(x) = F_A(x), \;\; x\in\mathcal{X},$$

and its regularized and subsampled version

$$\min_{\substack{f\in\mathcal{H}\\ \mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})}} J(f) + \rho\|f\|_{\mathcal{F}}^2 + \lambda\mathrm{Tr}\mathbf{B} \;\text{ s.t. }\; H_f(x_j) = F_{\mathbf{B}}(x_j) \;\; j\in[n], \tag{34}$$

with respective solutions $(f_\star, A_\star)$ and $(\hat{f}, \hat{\mathbf{B}})$. $(f_\star, A_\star)$ is an admissible point of (34). This implies that

$$J(\hat{f}) + \rho\|\hat{f}\|_{\mathcal{F}}^2 + \lambda\mathrm{Tr}\hat{\mathbf{B}} \le J(f_\star) + \rho\|f_\star\|_{\mathcal{F}}^2 + \lambda\mathrm{Tr}A_\star \tag{35}$$

and thus

$$J(\hat{f}) - J(f_\star) \le \lambda\mathrm{Tr}A_\star + \rho\|f_\star\|_{\mathcal{F}}^2 - \lambda\mathrm{Tr}\hat{\mathbf{B}} - \rho\|\hat{f}_\star\|_{\mathcal{F}}^2 \le \lambda\mathrm{Tr}A_\star + \rho\|f_\star\|_{\mathcal{F}}^2.$$

From Corollary 7, we have that $\tilde{f} \overset{\text{def}}{=} x \mapsto \hat{f}(x) + \frac{\eta}{2}\|x\|^2$ is convex. Further, by Lipschitzness of J, it holds $|J(\hat{f}) - J(\tilde{f})| \le L\frac{\eta}{2}N(\|\cdot\|^2)$. Plugging this above, we get

$$\begin{aligned} J(\tilde{f}) - J(f_\star) &\le \lambda\mathrm{Tr}A_\star + \rho\|f_\star\|_{\mathcal{F}}^2 + |J(\hat{f}) - J(\tilde{f})| \\ &\le \lambda\mathrm{Tr}A_\star + \rho\|f_\star\|_{\mathcal{F}}^2 + L\tfrac{\eta}{2}C_N. \end{aligned}$$

From Corollary 7, we have $\eta \ge C_0(\sum_{i=1}^d |(H_{\hat{f}})_{ii}|_{\mathcal{X},m} + MD_mB_K\mathrm{Tr}\hat{\mathbf{B}})h_{\hat{X},\mathcal{X}}^m$. From the definition of $|\cdot|_{\mathcal{X},m}$ in eq. (1) and Assumption 2c, we have

$$\sum_{i=1}^d |(H_{\hat{f}})_{ii}|_{\mathcal{X},m} \le \sum_{i=1}^d |\hat{f}|_{\mathcal{X},m+2} \le dD_{m+2}\|\hat{f}\|_{\mathcal{F}}.$$

Further, since $J(f_\star) \le J(\hat{f})$, we have from eq. (35) that $\rho\|\hat{f}\|_{\mathcal{F}}^2 + \lambda\mathrm{Tr}\hat{\mathbf{B}} \le \rho\|f_\star\|_{\mathcal{F}}^2 + \lambda\mathrm{Tr}A_\star$. From this we get

$$\begin{aligned} \mathrm{Tr}\hat{\mathbf{B}} &\le \mathrm{Tr}A_\star + \frac{\rho}{\lambda}\|f_\star\|_{\mathcal{F}}^2 \\ \|\hat{f}\|_{\mathcal{F}}^2 &\le \sqrt{\frac{\lambda}{\rho} + \|f_\star\|_{\mathcal{F}}^2}. \end{aligned}$$

Let $R^2 = \frac{\lambda}{\rho}\mathrm{Tr}A_\star + \|f_\star\|_{\mathcal{F}}^2$ and $C_1 = \frac{1}{2}LC_NC_0\max(D_m, D_{m+2})$. Using the inequalities above, we get

$$\begin{aligned} J(\tilde{f}) - J(f_\star) &\leq \rho R^2 + L\tfrac{\eta}{2}C_N \\ &\leq \rho R^2 + \frac{C_1 M B_K h^m_{\hat{X},\mathcal{X}}}{\lambda}\rho R^2 + C_1 d h^m_{\hat{X},\mathcal{X}} R, \end{aligned}$$

which yields eq. (20). When $\rho = \lambda = C_1 M B_K h^m_{\hat{X},\mathcal{X}}$, we obtain from the above that

$$\begin{aligned} J(\tilde{f}) - J(f_\star) &\leq 2\rho R^2 + d\frac{\rho}{MB_K}R \\ &\leq 2\rho(R + \frac{d}{4MB_K})^2 \\ &\leq 4\rho(R^2 + \frac{d^2}{16M^2B_K^2}) \\ &\leq 4MC_1(\mathrm{Tr}A_\star + \|f_\star\|_{\mathcal{F}}^2 + \frac{d^2}{M^2B_K^2}), \end{aligned}$$

which completes the proof. ■

C.4 Proof of Proposition 9

Proof We divide the proof in two parts: we start by showing that strong duality holds, and then derive the dual formulation.

Strong duality. Problem (28) is finite-dimensional and convex, hence it suffices to show that a strictly feasible pair $(\alpha, \mathbf{B})$ exists, i.e. a pair $(\alpha, \mathbf{B}) \in \mathbb{R}^n \times \mathbb{S}_+(\mathbb{R}^{nd})$ with $\mathbf{B} \succ 0$ satisfying eq. (23). Let us show that this is implied by (H_1) (in fact, it is equivalent). Indeed, (H_1) implies that there exists $\alpha \in \mathbb{R}^n$ such that the Hessian of $f(\cdot) = \sum_{i=1}^n k(\cdot, x_i)$ is positive-definite in $x_i, i \in [n]$. For $i \in [n]$, let us denote $\mathbf{H}_i \in \mathbb{R}^{d\times d}$ the Hessian of f in x_i. By hypothesis, it holds $\mathbf{H}_i \succ 0, i \in [n]$. We must show that there exists $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd}), \mathbf{B} \succ 0$ such that $\Psi_i^T \mathbf{B} \Psi_i = \mathbf{H}_i, i \in [n]$. For a universal kernel k, we have that $\{\Psi_i, \in [n]\}$ is a linearly independent family of $\mathbb{R}^{nd\times d}$. For $i \in [n]$, let $\mathbf{L}_i \in \mathbb{R}^{nd\times n}$ be such that $\mathbf{L}_i^T\mathbf{L}_i = \mathbf{H}_i$. Since $\mathbf{H}_i \succ 0, i \in [n]$, the $\mathbf{L}_i$'s have rank d. Let us further choose the $\mathbf{L}_i$'s such that $\mathbf{L} \stackrel{\text{def}}{=} (\,\mathbf{L}_1 \mid \ldots \mid \mathbf{L}_n\,) \in \mathbb{R}^{nd\times nd}$ has full rank. Then, let $\mathbf{R} \in \mathbb{R}^{nd\times nd}$ be such that $\mathbf{R}\Psi_i = \mathbf{L}_i, i \in [n]$. Such a $\mathbf{R}$ exists since $(\,\Psi_1 \mid \ldots \mid \Psi_n\,) \in \mathbb{R}^{nd\times nd}$ has full rank if k is universal. Let finally $\mathbf{B} = \mathbf{R}^T\mathbf{R}$. It holds $\Psi_i^T\mathbf{B}\Psi_i = \Psi_i^T\mathbf{R}^T\mathbf{R}\Psi_i = \mathbf{L}_i^T\mathbf{L}_i, i \in [n]$, and $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{nd})$. Further, since $\mathbf{L}$ has full rank by assumption, $\mathbf{B}$ has full rank, which implies $\mathbf{B} \succ 0$. Therefore, $(\alpha, \mathbf{B})$ is a strictly feasible pair, and strong duality holds.

Dual formulation. Since strong duality holds, we may apply Lagragian duality. The Lagrangian of eq. (28) is

$$\begin{aligned}\mathcal{L}(\alpha, \mathbf{B}, \Gamma) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{n}\alpha_i k(x_i, x_j))^2 + \rho\alpha^T\mathbf{K}\alpha + \lambda_1 \mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2 \\ - \sum_{i,j=1}^{n}\sum_{p,q=1}^{d}\Gamma_{pq}^{(j)}\alpha_i\frac{\partial^2 k(x_i, x_j)}{\partial x_j^p \partial x_j^q} + \sum_{j=1}^{n}\langle\Gamma^{(j)},\, \Psi_j^T\mathbf{B}\Psi_j\rangle_F.\end{aligned} \tag{36}$$

■

Let us derive the optimality conditions w.r.t. α. It holds

$$\nabla_\alpha\mathcal{L}(\alpha, \mathbf{B}, \Gamma) = \frac{2}{n}\mathbf{K}(\mathbf{K}\alpha - y) + 2\rho\mathbf{K}\alpha - \sum_{i=1}^{n}\sum_{p,q=1}^{d}\Gamma_{p,q}^{(i)}\frac{\partial^2 K(X, x_j)}{\partial x_j^p \partial x_j^q}.$$

Setting the gradient to zero, we get

$$\alpha = (\frac{1}{n}\mathbf{K}^2 + \rho\mathbf{K})^{-1}\left(\frac{1}{n}\mathbf{K}y + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{p,q=1}^{d}\Gamma_{pq}^{(j)}\frac{\partial^2 k(x_i, x_j)}{\partial x_j^p \partial x_j^q}\right)$$

Further, from Lemma 13 we have that

$$\inf_{\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})}\lambda_1\mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2 + \sum_{i=1}^{n}\langle\Gamma^{(i)},\, \Psi_j^T\mathbf{B}\Psi_j\rangle_F = -\frac{1}{2\lambda_2}\|[\sum_{i=1}^{n}\Psi_i\Gamma^{(i)}\Psi_i^T + \lambda_1\mathbf{I}_{nd}]_-\|_F^2.$$

Plugging everything in eq. (36), we get eq. (28).

C.5 Convex Regression with Reproducing Property of Derivatives

In this section, we consider convex regression with the Hessian representation (21). That is, given samples $(x_i, y_i) \in \mathcal{X}\times\mathbb{R}, i \in [n]$ and grid points $v_j \in \mathcal{X}, j \in [\ell]$ (that may but need not coincide with the x_i's) we solve the following problem

$$\begin{aligned}\min_{\substack{f\in\mathcal{H}\\ \mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})}} \quad & \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_1\mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2 + \rho\|f\|_{\mathcal{H}}^2 \\ \text{s.t.} \quad & \sum_{p,q=1}^{d}\langle f,\, \partial_{pq}k_{v_j}\rangle e_p e_q^T = \Psi_j^T\mathbf{B}\Psi_j, \quad j \in [\ell].\end{aligned} \tag{37}$$

The following proposition establishes the dual formulation of (37), which is more amenable to optimization.

Proposition 17 (Convex regression)**.** *Let k be a kernel satisfying Assumption 2 and such that $\mathcal{C}^s(\mathcal{X}) \subset \mathcal{H}$ for some $s \in \mathbb{N}_+$. Strong duality applies, and eq.* (37) *admits the following*

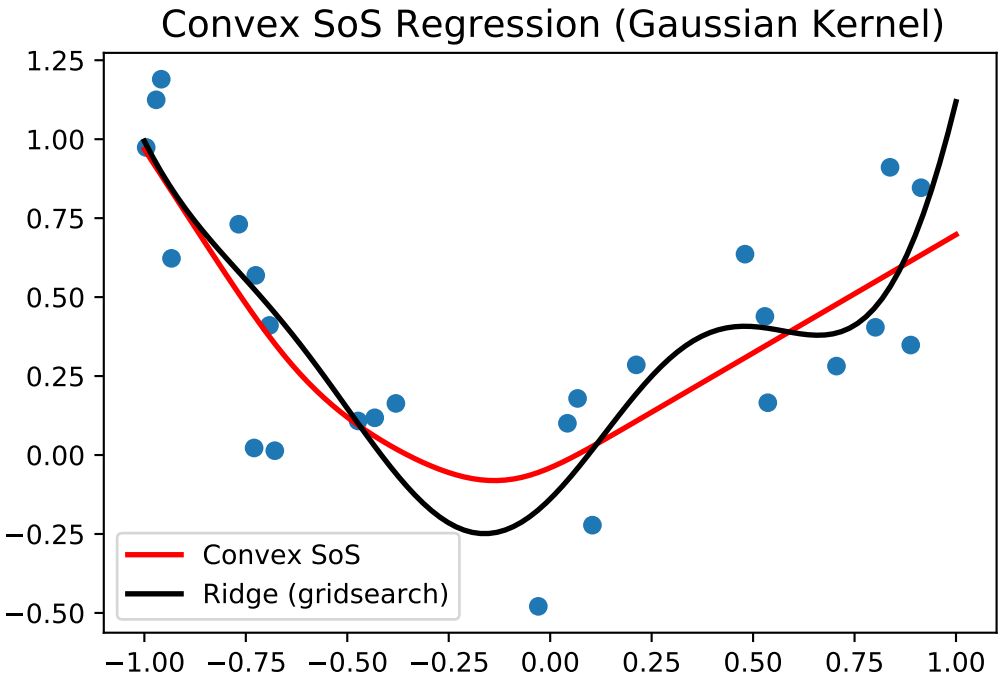


Figure 5: Convex regression with reproducing property: 1D example.

dual formulation:

$$\begin{aligned}\min_{\Gamma\in\mathrm{Sym}(\mathbb{R}^d)^\ell} \Big\{ & \frac{1}{4\rho}\sum_{p,q,r,s=1}^{d}\left(\Gamma_{pq}^{(\cdot)}\right)^T \partial^4_{pqrs}\mathbf{K}(V,V)\Gamma_{rs}^{(\cdot)} \\ & - \frac{1}{4n\rho}\sum_{p,q,r,s=1}^{d}\left(\Gamma_{pq}^{(\cdot)}\right)^T \partial^2_{pq}\mathbf{K}(X,V)^T\mathbf{W}\partial^2_{rs}\mathbf{K}(X,V)\Gamma_{rs}^{(\cdot)} \\ & + \frac{1}{2n\rho}y^T(\mathbf{I}_n+\rho\mathbf{W}-\frac{1}{n}\mathbf{K}\mathbf{W})(\sum_{p,q=1}^{d}\partial^2_{pq}\mathbf{K}(X,V)\Gamma_{pq}^{(\cdot)}) \\ & + \frac{1}{n}y^T(\frac{1}{n}\mathbf{K}\mathbf{W}-\mathbf{I}_n)y+\Omega^*(\sum_{j=1}^{\ell}\Psi_j\Gamma^{(j)}\Psi_j^T)\Big\},\end{aligned}$$

where for $p,q\in[d]$, $\Gamma_{pq}^{(\cdot)}\overset{\text{def}}{=}[\Gamma_{pq}^{(j)}]_{j\in[\ell]}\in\mathbb{R}^\ell$, *and with*

$$\mathbf{W}\overset{\text{def}}{=}(\frac{1}{n}\mathbf{K}+\rho\mathbf{I}_n)^{-1},\quad K_{ij}=k(x_i,x_j), i,j\in[n],$$

$$[\partial^2_{pq}\mathbf{K}(X,V)]_{ij}\overset{\text{def}}{=}\frac{\partial^2 k(x_i,v_j)}{\partial v_j^p\partial v_j^q},\quad p,q\in[d], i\in[n], j\in[\ell],$$

$$[\partial^4_{pqrs}\mathbf{K}(V,V)]_{ij}\overset{\text{def}}{=}\frac{\partial^4 k(v_i,v_j)}{\partial v_i^p\partial v_i^q\partial v_j^r\partial v_j^s}\quad p,q,r,s\in[d], i\in[n], j\in[\ell].$$

and

$$\Omega^*(\mathbf{B}) = \begin{cases} \frac{1}{2\lambda_2}\|[\mathbf{B}+\lambda_1\mathbf{I}_d]_-\|_F^2 & \text{if } \lambda_2 > 0 \text{ and } \lambda_1 \geq 0, \\ \iota_{\{\mathbf{B}+\lambda_1\mathbf{I}_{nd}\succeq 0\}} & \text{if } \lambda_2 = 0 \text{ and } \lambda_1 > 0. \end{cases}$$

Further, the corresponding primal solution is

$$f^\star = \sum_{i=1}^n (\alpha_i + \frac{1}{2}\sum_{p,q=1}^d \delta_{p,q}^{(i)}) k_{x_i} + \frac{1}{2\rho}\sum_{j=1}^\ell \sum_{p,q=1}^d \Gamma_{pq}^{(j)} \partial_{pq} k_{v_j},$$

where

$$\begin{aligned} \alpha &= \frac{1}{n}\mathbf{W}y \\ \delta_{pq}^{(\cdot)} &= -\frac{1}{n\rho}\mathbf{W}\partial_{pq}^2 \mathbf{K}(X,V)\Gamma_{pq}^{(\cdot)} \in \mathbb{R}^n, \quad p,q \in [d]. \end{aligned}$$

Proof We start by showing that strong duality holds, and then derive the dual formulation.

Strong duality. Since (37) is a convex problem, it suffices to show that a strictly feasible point exists. By assumption, $\mathcal{H}$ contains $\mathcal{C}^s$ functions. Let $f \in \mathcal{C}^s(\mathcal{X})$ be such that f is strictly convex in $v_j, j \in [\ell]$ (e.g. pick f strongly convex and $\mathcal{C}^s$). Let $\mathbf{H}_j \stackrel{\text{def}}{=} \mathbf{H}_f(v_j) \succ 0, j \in [\ell]$. Following the construction in Appendix C.4, there exists $\mathbf{B} \in \mathbb{S}_+(\mathbb{R}^{\ell d}), \mathbf{B} \succ 0$ such that $\Psi_j \mathbf{B} \Psi_j = \mathbf{H}_j, j \in [\ell]$. Hence, $(f, \mathbf{B})$ is a strictly feasible point.

Dual formulation. Since strong duality holds, we may apply Lagrangian duality. We have

$$\begin{aligned} \mathcal{L}(f, \mathbf{B}, \Gamma) = &\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_1 \mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2 + \rho\|f\|_{\mathcal{H}}^2 \\ &- \langle f, \sum_{j=1}^\ell \sum_{p,q=1}^d \Gamma_{pq}^{(j)} \partial_{pq} k_{v_j} \rangle + \sum_{j=1}^\ell \langle \Gamma^{(j)}, \Psi_j^T \mathbf{B} \Psi_j \rangle_F. \end{aligned} \tag{38}$$

Let us derive the optimality conditions w.r.t. f. We have

$$\nabla_f \mathcal{L}(f, \mathbf{B}, \Gamma) = \frac{2}{n}\sum_{i=1}^n (\langle f, k_{x_i}\rangle_{\mathcal{H}} - y_i) k_{x_i} + 2\rho f - \sum_{j=1}^\ell \sum_{p,q=1}^d \Gamma_{pq}^{(j)} \partial_{pq} k_{v_j}.$$

Setting this gradient to zero, we obtain

$$f = \mathcal{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n y_i k_{x_i} + \frac{1}{2}\sum_{j=1}^\ell \sum_{p,q=1}^d \Gamma_{pq}^{(j)}, \partial_{pq} k_{v_j}) \right)$$

with $\mathcal{S} = \frac{1}{n}\sum_{i=1}^n k_{x_i} \otimes k_{x_i} + \rho\,\mathrm{Id}$. We can show that

$$\mathcal{S}^{-1}(\frac{1}{n}\sum_{i=1}^n y_i k_{x_i}) = \sum_{i=1}^n \alpha_i k_{x_i} \quad \text{with} \quad \alpha = \frac{1}{n}\mathbf{W}y,$$

and

$$\mathcal{S}^{-1}(\sum_{j=1}^{\ell}\sum_{p,q=1}^{d}\Gamma_{pq}^{(j)}\partial_{pq}k_{v_j}) = \sum_{i=1}^{n}\delta_{pq}^{(i)}k_{x_i} + \frac{1}{\rho}\sum_{j=i}^{\ell}\sum_{p,q=1}^{d}\Gamma_{pq}^{(j)}\partial_{pq}k_{v_i},$$

with

$$\delta_{pq}^{(\cdot)} = -\frac{1}{n\rho}\mathbf{W}\partial_{pq}^2\mathbf{K}(X,V)\Gamma_{pq}^{(\cdot)} \in \mathbb{R}^n, p, q \in [d],$$

where $\mathbf{W} \overset{\text{def}}{=} (\rho\mathbf{I}_n + \frac{1}{n}\mathbf{K})^{-1}$ and $[\partial_{pq}^2\mathbf{K}(X,V)]_{ij} \overset{\text{def}}{=} \frac{\partial^2 k(x_i,v_j)}{\partial v_j^p \partial v_j^q}$, $i \in [n], j \in [\ell], p, q \in [d]$. Let us now derive the optimality conditions in $\mathbf{B}$. We have

$$\begin{aligned}
&\inf_{\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})} \lambda_1 \mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2 + \sum_{j=1}^{\ell}\langle\Gamma^{(j)}, \Psi_j^T\mathbf{B}\Psi_j\rangle_F \\
&= \inf_{\mathbf{B}\in\mathbb{S}_+(\mathbb{R}^{nd})} \lambda_1 \mathrm{Tr}\mathbf{B} + \frac{\lambda_2}{2}\|\mathbf{B}\|_F^2 + \langle\mathbf{B}, \sum_{j=1}^{\ell}\Psi_j\Gamma^{(j)}\Psi_j^T\rangle_F \\
&= \Omega^*(\sum_{j=1}^{\ell}\Psi_j\Gamma^{(j)}\Psi_j^T),
\end{aligned}$$

where, from Lemma 13,

$$\Omega^*(\mathbf{B}) = \begin{cases} \frac{1}{2\lambda_2}\|[\mathbf{B} + \lambda_1\mathbf{I}_d]_-\|_F^2 & \text{if } \lambda_2 > 0 \text{ and } \lambda_1 \geq 0, \\ \iota_{\{\mathbf{B}+\lambda_1\mathbf{I}_{nd}\succeq 0\}} & \text{if } \lambda_2 = 0 \text{ and } \lambda_1 > 0. \end{cases}$$

Finally, plugging everything in eq. (38), we get the following problem:

$$\begin{aligned}
\inf_{\Gamma\in\mathrm{Sym}(\mathbb{R}^d)^n} \Big\{ &\frac{1}{4\rho}\sum_{p,q,r,s=1}^{d}(\Gamma_{pq}^{(\cdot)})^T\partial_{pqrs}^4\mathbf{K}(V,V)\Gamma_{rs}^{(\cdot)} \\
&- \frac{1}{4n\rho}\sum_{p,q,r,s=1}^{d}\left(\Gamma_{pq}^{(\cdot)}\right)^T\partial_{pq}^2\mathbf{K}(V,X)\mathbf{W}\partial_{rs}^2\mathbf{K}(X,V)\Gamma_{rs}^{(\cdot)} \\
&+ \frac{1}{2n\rho}y^T(\mathbf{I}_n + \rho\mathbf{W} - \frac{1}{n}\mathbf{K}\mathbf{W})(\sum_{p,q=1}^{d}\partial_{pq}^2\mathbf{K}(X,V)\Gamma_{pq}^{(\cdot)}) \\
&+ \frac{1}{n}y^T(\frac{1}{n}\mathbf{K}\mathbf{W} - \mathbf{I}_n)y\Big\},
\end{aligned}$$

with $[\partial_{pqrs}^2\mathbf{K}(V,V)]_{ij} \overset{\text{def}}{=} \frac{\partial^4 k(v_i,v_j)}{\partial v_i^p \partial v_i^q \partial v_j^r \partial v_j^s}$, $i, j \in [\ell], p, q, r, s \in [d]$. ■

Appendix D. Additional Experiments and Numerical Details

D.1 PSD-Valued Least Squares Regression

We select hyperparameters λ_1, λ_2 and exponential kernel bandwidth σ using leave-one-out cross-validation, over the following grid: $\lambda_1 \in \{10^{-n}, n = 0, ..., 8\} \cup \{0\}$, $\lambda_2 \in \{10^{-n}, n =$

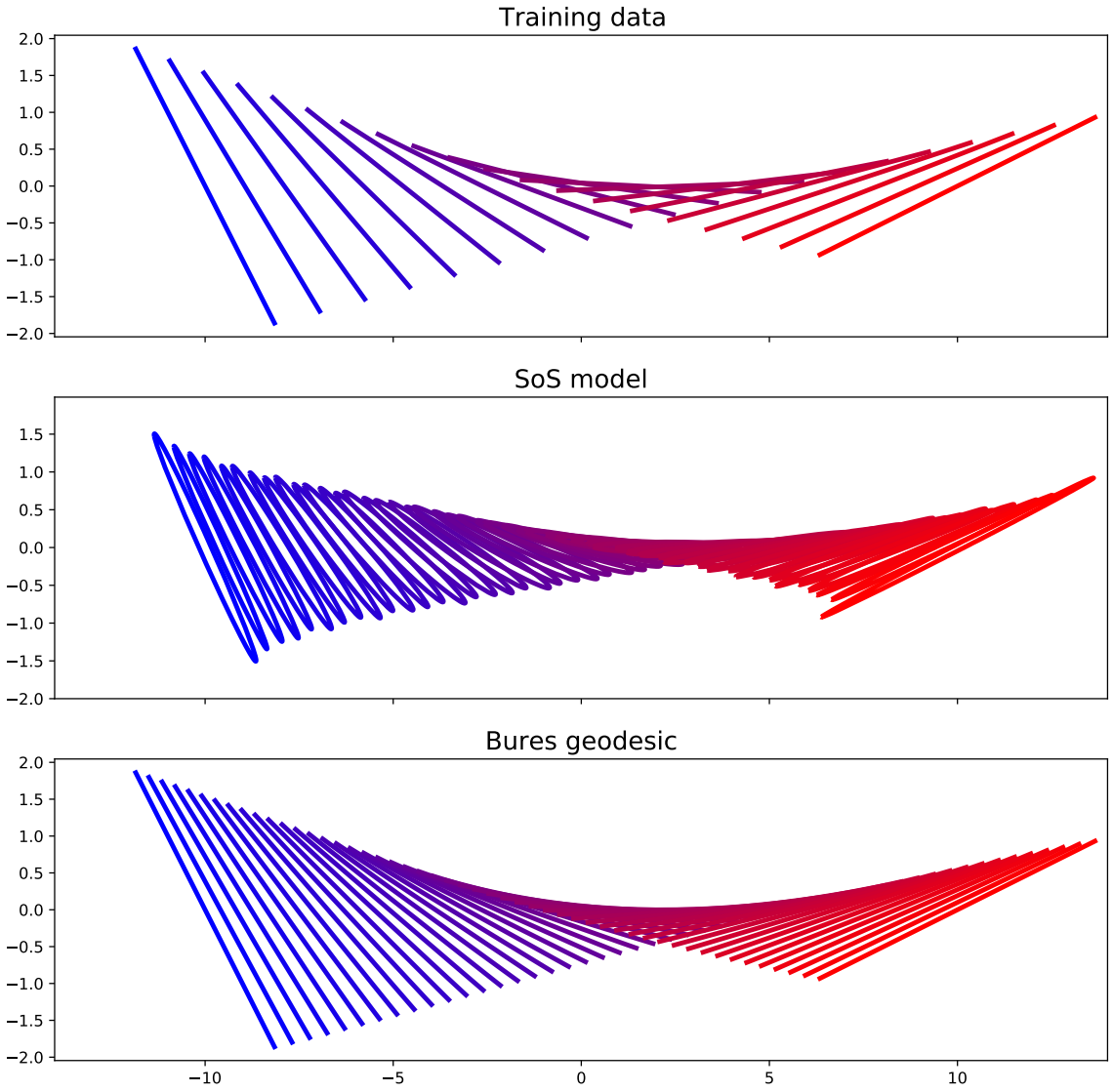


Figure 6: Interpolation of a rank 1 Bures geodesic from 12 data points (top). The matrices are represented as the level sets of (potentially degenerate) Gaussian distributions: $\{x : \mathcal{N}(x; 0, \Sigma) \leq r\}$. We use the exponential kernel and select all hyperparameters using cross-validation. The learned model is represented in the middle figure, and the full geodesic is plotted in the bottom figure for comparison.

$0, ..., 8\}$, $\sigma \in \{1, 0.1, 0.01\}$. In Figure 1, the parameters selected by CV are $\lambda_1 = 0, \lambda_2 = 10^{-5}, \sigma = 0.1$ and in Figure 6, $\lambda_1 = 0, \lambda_2 = 10^{-7}, \sigma = 0.1$.

D.2 Convex Regression

We select hyperparameters ρ, λ_2 and Gaussian kernel bandwidth σ^2 using 5-fold cross-validation, over the following grid: $\lambda_2 \in \{10^{-n}, n = 3, ..., 7\}$, $\sigma^2 \in \{1, 5, 10\}$. λ_1 is fixed to 0. When $n > 25$, we perform Nyström approximation with rank $r = 25$. In Figure 3, we display the average scores and their standard deviations on 10 independent sets of samples. In Figure 4, the hyperparameters selected by CV are $\lambda_2 = 10^{-3}, \rho = 10^{-5}$ and $\sigma^2 = 10$.

References

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

Mauricio A Álvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

Pierre-Cyril Aubin-Frankowski and Zoltan Szabo. Handling hard affine SDP shape constraints in RKHSs. *arXiv preprint arXiv:2101.01519*, 2021.

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

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Jonathan Borwein and Adrian S Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer Science & Business Media, 2006.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Yize Chen, Yuanyuan Shi, and Baosen Zhang. Optimal control via neural networks: A convex approach. In *International Conference on Learning Representations*, 2018.

Philippe Choné and Hervé VJ Le Meur. Non-convergence result for conformal approximation of variational problems subject to a convexity constraint. 2001.

Clifford Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619, 1954.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge university press, 2012.

Thomas Lachand-Robert and Édouard Oudet. Minimizing within convex bodies using a convex hull method. *SIAM Journal on Optimization*, 16(2):368–379, 2005.

Ives Macêdo and Rener Castro. Learning divergence-free and curl-free vector fields with matrix-valued kernels. Technical report, Instituto Nacional de Matematica Pura e Aplicada, 2008.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

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

Quentin Mérigot and Edouard Oudet. Handling convexity-like constraints in variational problems. *SIAM Journal on Numerical Analysis*, 52(5):2466–2487, 2014.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Jean-Marie Mirebeau. Adaptive, anisotropic and hierarchical cones of discrete convex functions. *Numerische Mathematik*, 132(4):807–853, 2016.

Boris Muzellec, Francis Bach, and Alessandro Rudi. A note on optimizing distributions using kernel mean embeddings. *arXiv preprint arXiv:2106.09994*, 2021.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge University Press, 2016.

Jean-Charles Rochet and Philippe Choné. Ironing, sweeping, and multidimensional screening. *Econometrica*, pages 783–826, 1998.

Alessandro Rudi and Carlo Ciliberto. PSD representations for effective probability models. *arXiv preprint arXiv:2106.16116*, 2021.

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

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Emilio Seijo and Bodhisattva Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

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

Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge university press, 2004.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*. Citeseer, 2001.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1-2):456–463, 2008.