SIAM J. MATH. DATA SCI.
Vol. 4, No. 1, pp. 153–178

Positive Semi-definite Embedding for Dimensionality Reduction and Out-of-Sample Extensions*

Michaël Fanuel†, Antoine Aspeel‡, Jean-Charles Delvenne‡, and Johan A. K. Suykens§

Abstract. In machine learning or statistics, it is often desirable to reduce the dimensionality of a sample of data points in a high dimensional space $\mathbb{R}^d$. This paper introduces a dimensionality reduction method where the embedding coordinates are the eigenvectors of a positive semi-definite kernel obtained as the solution of an infinite dimensional analogue of a semi-definite program. This embedding is adaptive and non-linear. We discuss this problem both with weak and strong smoothness assumptions about the learned kernel. A main feature of our approach is the existence of an out-of-sample extension formula of the embedding coordinates in both cases. This extrapolation formula yields an extension of the kernel matrix to a data-dependent Mercer kernel function. Our empirical results indicate that this embedding method is more robust with respect to the influence of outliers compared with a spectral embedding method.

Key words. diffusion maps, kernel methods, dimensionality reduction, semi-definite program

AMS subject classifications. 60J60, 42B35, 47A58, 30C40

DOI. 10.1137/20M1370653

1. Introduction.

Dimensionality reduction is often an essential step which precedes, for instance, a clustering procedure. This process consists in mapping a sample of n data points in $\mathbb{R}^d$ into a lower dimensional space. Among the possible approaches, this work addresses a special case of non-linear *adaptive* embedding, in the spirit of manifold learning. This is in contrast with linear dimensionality reduction methods such as, e.g., the techniques based on compressed sensing, which are data-oblivious. In the spirit of non-linear methods, our approach is closely related to diffusion maps [7, 9]. A central object of this definition is a

*Received by the editors September 30, 2020; accepted for publication (in revised form) November 23, 2021; published electronically February 10, 2022.
https://doi.org/10.1137/20M1370653
Funding: The authors thank the following organizations. EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923) and ERC AdG E-DUALITY (787960). This paper reflects only the authors' views; the Union is not liable for any use that may be made of the contained information. Research Council KUL: Optimization frameworks for deep kernel machines C14/18/068. Flemish Government: FWO: projects: GOA4917N (Deep Restricted Kernel Machines: Methods and Foundations). PhD/Postdoc grant Impulsfonds AI: VR 2019 2203 DOC.0318/1QUATER Kenniscentrum Data en Maatschappij. Ford KU Leuven Research Alliance Project KUL0076 (Stability analysis and performance improvement of deep reinforcement learning algorithms).

†Université de Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France (michael.fanuel@univ-lille.fr).

‡Université catholique de Louvain, Ecole polytechnique de Louvain, ICTEAM and CORE, Avenue Georges Lemaître, 4-6, Louvain-la-Neuve, B-1348, Belgium (antoine.aspeel@uclouvain.be, jean-charles.delvenne@uclouvain.be).

§KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (Johan.Suykens@esat.kuleuven.be).

diffusion kernel determining a diffusion process on the points of the dataset such that the probability to diffuse from one point to another is large only if the points are in a common neighborhood. Given a distribution of data points in $\mathbb{R}^d$, a diffusion embedding is obtained thanks to the spectral decomposition of the diffusion kernel, which is associated to an integral operator, or simply a square matrix in the discrete setting. Then, the mth largest eigenvalues of the diffusion kernel are selected in order to obtain an approximate embedding in $\mathbb{R}^m$, while the m embedding coordinates are given by the associated m eigenvectors. In this work, we discuss a semi-definite program (SDP) which is closely related to the spectral problem considered in diffusion maps. Beyond the embedding of a training dataset, diffusion maps allow for the out-of-sample extension of the embedding map so that any forthcoming point can be naturally embedded. This extrapolation is done thanks to a linear formula relying on the Nyström extension (see, e.g., [8, 29]).

The SDP embedding presented in this work shares many properties with a diffusion embedding, although its out-of-sample extension formula is non-linear. After introducing the notation used throughout this paper, we present the key results in a simplified setting.

1.1. Notation. Integral operators will be denoted by uppercase letters $(A, B, \dots)$, while associated integral kernels will be denoted by lower case letters $(a, b, \dots)$. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space. A linear operator $A : \mathcal{H} \to \mathcal{H}$ is said to be positive semi-definite (*psd*) if $\langle g, Ag \rangle \geq 0$ for all $g \in \mathcal{H}$, and in that case we write $A \succeq 0$. The nuclear norm of an operator on a Hilbert space is written $\|A\|_* = \mathrm{Tr}(\sqrt{A^* A})$. For convenience, we introduce the following space:

$$\mathcal{S}(\mathcal{H}) = \{A : \mathcal{H} \to \mathcal{H} \text{ s.t. } A \text{ is self-adjoint and } \|A\|_* < \infty\}.$$

We denote by $\mathcal{C}^m(X)$ the space of m-times continuously differentiable functions on $X \subset \mathbb{R}^d$. The smoothness of a function $d \in \mathcal{C}^m(X)$ is measured thanks to the semi-norm $|d|_{X,m} = \max_{|\alpha|=m} \sup_{x \in X} |\partial^\alpha d(x)|$, where ∂^α is the partial derivative with respect to the multi-index α. We will consider the Sobolev space $W_2^s(X)$ with the following inner product: $\langle g, g' \rangle = \langle g, g' \rangle_{L^2(X)} + \sum_{|\alpha|=s} \langle \partial^\alpha g, \partial^\alpha g' \rangle_{L^2(X)}$.

Matrices will be denoted by capital bold letters $(\mathbf{A}, \mathbf{B}, \dots)$, while italic bold letters will be used for vectors $(\boldsymbol{a}, \boldsymbol{b}, \dots)$. Then, we will write $\mathrm{Diag}(\boldsymbol{a})$ for the diagonal matrix with diagonal entries given by the entries of $\boldsymbol{a}$. Similarly, the vector $\mathrm{diag}(\mathbf{A})$ contains the diagonal elements of the matrix $\mathbf{A}$. Let $\mathrm{ddiag}(\mathbf{A})$ be the diagonal matrix constructed from the diagonal of $\mathbf{A}$ and with zero off-diagonal entries. The all-ones column vector is denoted by $\mathbf{1}$. Also, we write the set of integers $\{1, \dots, n\} = [n]$. Finally, we write $a \lesssim b$ if there exists a constant $c > 0$ such that $a < cb$.

1.2. Setting and outline of the main results. In this paper, we consider the domain X to be a bounded set of $\mathbb{R}^d$. For defining the general variational problem, we take X to be the ℓ_1-ball $[-c, c]^d$ with $c > 0$ since no smoothness assumption is needed in this case, while X is taken to be an ℓ_2-ball of diameter $2R > 0$ when we discuss the kernelized problem. The space of symmetric and nuclear operators acting on square-integrable functions on X is denoted by $\mathcal{S}(L^2(X))$ or simply $\mathcal{S}$ when no ambiguity is possible. Henceforth, we consider *psd* operators in $\mathcal{S}$ because they are associated to symmetric *psd* kernel functions, as we explain briefly below. Classically, a Hilbert–Schmidt operator A can be represented as an integral

operator $Ag(x) = \int_X a(x,y)g(y)\mathrm{d}y$ where $a \in L^2(X \times X)$. A stronger result exists if A is also nuclear, self-adjoint, and *psd*. Namely, a generalized Mercer theorem by Steinwart and Scovel [22], which is recalled as Theorem A.1 in the appendix, states that for all positive semi-definite operators $A \in \mathcal{S}$, there exists an associated integral kernel $a_M(x,y)$, which is defined *pointwisely*, i.e., for all $x, y \in X$. Such a representative a_M is called here a *Mercer* kernel of $A \in \mathcal{S}$, as indicated by the subscript $\cdot_M$. By contrast, a Hilbert–Schmidt operator does not necessarily admit a kernel that is defined pointwisely.

Let us give some motivations for these definitions in the context of manifold learning.

Motivation from diffusion embedding. In [9, 7, 8, 16] and subsequent works about diffusion geometry, the multiscale structure of data has been successfully studied thanks to the spectral properties of diffusion kernels. In the context of diffusion maps, an operator $\bar{A} \in \mathcal{S}$ is often defined thanks to its symmetric *diffusion kernel*[1]

$$\bar{a}(x,y) = \frac{e^{-\|x-y\|_2^2/\sigma^2}}{\sqrt{m(x)m(y)}} - \sqrt{\frac{m(x)}{m_t}}\sqrt{\frac{m(y)}{m_t}}, \tag{1.1}$$

where the function $m(x) = \int_X e^{-\|x-y\|^2/\sigma^2}\mathrm{d}y$ is a density and $m_t = \int_X m(x)\mathrm{d}x$ is a normalization. Notice that $\bar{A}$ is *psd* for the following reasons. Since the Gaussian kernel is strictly positive definite, the first term in (1.1) is *psd*, and, thanks to the normalization, the dominant eigenfunction of this first term is $\psi^{(0)}(x) = \sqrt{m(x)/m_t}$ with eigenvalue 1. Therefore, this eigenfunction is also an eigenvector of $\bar{A}$ but with eigenvalue zero, since the second term in (1.1) actually subtracts[2] a projector on the space generated by $\psi^{(0)}(x)$.

Diffusion maps are defined thanks to the spectral decomposition of $\bar{A}$, while the diffusion distance is associated to the ℓ^2-distance between two points in the diffusion embedding. Let $\{\lambda^{(\ell)}\}_{\ell\geq 1}$ and $\{\psi^{(\ell)}\}_{\ell\geq 1}$ be, respectively, the eigenvalues sorted in descending order and the associated eigenfunctions of $\bar{A}$. Then, the diffusion embedding $\Psi : X \to \ell^2$ is defined as $\Psi(x) = \big(\lambda^{(\ell)}\psi^{(\ell)}(x)\big)_{\ell\geq 1}$. As a simple consequence, the squared distance between the embedding of x and $y \in X$ is related to an $L^2(X)$ distance as follows:

$$D(x,y)^2 = \|\Psi(x) - \Psi(y)\|_{\ell^2}^2 = \|\bar{a}(x,\cdot) - \bar{a}(y,\cdot)\|_{L^2(X)}^2 = \int_X \big(\bar{a}(x,u) - \bar{a}(y,u)\big)^2\mathrm{d}u.$$

It is then common to compute only the eigenfunctions with the largest eigenvalues to yield a low dimensional embedding [9]. In particular, the operator with integral kernel $b_\star(x,y) = \psi^{(1)}(x)\psi^{(1)}(y)$—associated with the leading eigenfunction $\psi^{(1)}$ of $\bar{A}$—can be obtained as the solution of the problem $\sup_{B\in\mathcal{S}} \mathrm{Tr}(\bar{A}B)$ subject to $B \succeq 0$ and $\mathrm{Tr}(B) = 1$, as can be seen by introducing a spectral decomposition of $\bar{A}$.

In analogy with this problem, we propose another trace maximization problem with the same diffusion kernel but which involves additional constraints; i.e., the diagonal values of the kernel are constrained rather than the trace. Hence, to bound the diagonal, we introduce

[1] There is an alternative definition of diffusion maps in [7] which involves a non-symmetric kernel. Both operators are related by a conjugation.

[2] Although it is not present in the classical diffusion maps definition, the second term in (1.1) only subtracts an uninformative quantity related to the density of data.

a continuous and *strictly positive* function $d(x)$ that enters the variational problem (VP) hereafter. Our analysis is twofold:

- In section 1.2.1, an analysis is presented with weak smoothness assumptions. A simple discretization scheme is presented.
- In section 1.2.2, a kernelized approach is given within a reproducing kernel Hilbert space (RKHS) of continuously differentiable functions. This setting yields stronger statistical guarantees obtained thanks to results from [21] to which we refer extensively.

In particular, the latter approach allows for a built-in out-of-sample extension, whereas the former setting has a less natural extrapolation formula.

1.2.1. General variational problem. In view of this spectral problem, we define the following maximization problem:

$$\text{(VP)} \qquad \sup_{B \in \mathcal{S}(L^2(X))} \operatorname{Tr}(\bar{A}B) \text{ subject to } B \succeq 0 \text{ and } b_M(x,x) \leq d(x) \quad \text{almost everywhere,}$$

where $b_M(x,y)$ is a Mercer kernel associated to B as it is given by Theorem A.1 in the appendix. The inequality constraint in (VP) should be satisfied on X except possibly on a negligible set for the Lebesgue measure. As a main result of our paper, we show that the diagonal constraint makes sense and that the supremum hereinabove is attained by a positive semi-definite operator $B_\star \in \mathcal{S}$ with a Mercer kernel satisfying $b_{\star,M}(x,x) \leq d(x)$ almost everywhere.

Discrete problem. In practice, we solve an analogue of (VP) involving square symmetric matrices in order to calculate an embedding given by the eigenvectors of the solution. To start, we define a discrete analogue of (1.1). Given a sample of data points $\{x_i \in X\}_{i\in[n]}$ and a kernel matrix $[\mathbf{K}]_{ij} = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$ for $i,j \in [n]$, we can define the empirical normalized kernel by

$$\mathbf{A} = \operatorname{Diag}(1/\sqrt{\mathbf{m}})\mathbf{K}\operatorname{Diag}(1/\sqrt{\mathbf{m}}) \quad \text{with} \quad \mathbf{m} = \mathbf{K}\mathbf{1}, \tag{1.2}$$

where the above division is done elementwise. The subtracted kernel matrix reads $\bar{\mathbf{A}} = \mathbf{A} - \boldsymbol{v}^{(0)}\boldsymbol{v}^{(0)\top}$, where $\boldsymbol{v}^{(0)} = \sqrt{\mathbf{m}/(\mathbf{1}^\top \boldsymbol{m})}$ is the dominant eigenvector of $\mathbf{A}$ with eigenvalue 1. Given the function $d(x)$, a discrete counterpart of (VP) is the SDP

$$\text{(SDP)} \qquad \max_{\mathbf{B}\succeq 0} \operatorname{Tr}\left(\bar{\mathbf{A}}\mathbf{B}\right), \text{ subject to } \operatorname{diag}(\mathbf{B}) \leq \boldsymbol{d},$$

where the inequality constraint holds elementwise and with $[\boldsymbol{d}]_i = d(x_i)$ for all $i \in [n]$. The embedding coordinates $x_i \mapsto \boldsymbol{\Xi}_{i*} = [\boldsymbol{\chi}_i^{(1)}, \ldots, \boldsymbol{\chi}_i^{(r)}]^\top$ are given by the eigenvectors $\{\boldsymbol{\chi}^{(\ell)} \in \mathbb{R}^n\}_{1\leq \ell \leq r}$ of the solution $\boldsymbol{B}_\star$ with non-zero eigenvalue $\lambda^{(\ell)}$, which are normalized such that $\|\boldsymbol{\chi}^{(\ell)}\|_2^2 = \lambda^{(\ell)}$. The embedding obtained is illustrated in Figure 1 in a toy example. An advantage of the SDP embedding over the diffusion embedding is that the length of the embedding vectors is constrained by construction. Indeed, the squared length of the embedding vector is the corresponding diagonal element of the matrix $\boldsymbol{B}_\star$ that is upper bounded in (SDP). This constraint impedes a localization effect that can be observed in spectral embedding methods (see Figure 5, for instance). The dimensionality of the SDP embedding is the rank of $\boldsymbol{B}_\star$, which is presumably low as we argue here in light of our empirical simulations and of Proposition 3.1, which is given in section 3.1.

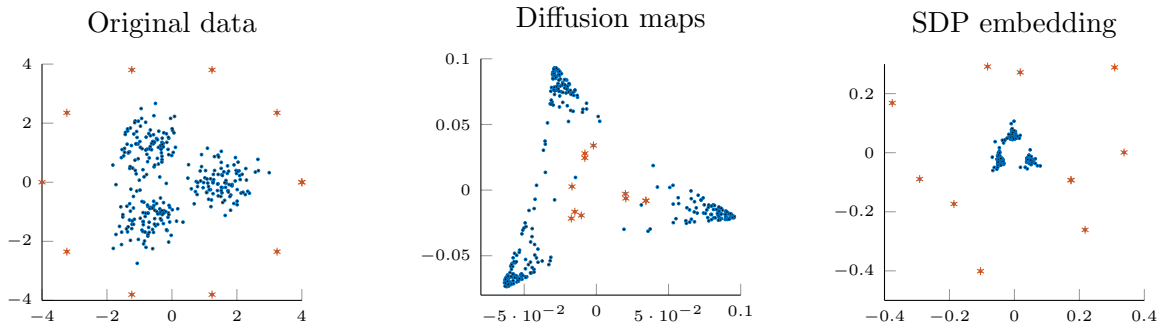

Figure 1. *SDP and diffusion maps embedding ($\sigma = 1.2$). The dataset, on the left-hand side (LHS), contains three clusters with 100 points each and 10 outliers (red stars).*

Out-of-sample extension. In this work, we also propose an extrapolation of the vectors $\boldsymbol{\chi}^{(\ell)}$ which allows us to embed additional data arising after the embedding of the initial n datapoints.[3] First, we can define an empirical diffusion kernel function whose Gram matrix $\boldsymbol{A}$ is given by (1.2), i.e.,

$$a_e(x,y) = \frac{e^{-\|x-y\|_2^2/\sigma^2}}{\sqrt{m_e(x)m_e(y)}} \text{ and } m_e(x) = \sum_{i=1}^{n} e^{-\|x-x_i\|_2^2/\sigma^2}, \tag{1.3}$$

where we indicated by a subscript $\cdot_e$ an extended empirical quantity. In the same way, the extension of any column of the Gram matrix of $a_e(x,y)$ is naturally defined by $[\boldsymbol{a}_e(x)]_i = a_e(x,x_i)$ in view of (1.3). This yields an extrapolation formula for the subtracted kernel. First, we extend the matrix $\bar{\mathbf{A}} = \mathbf{A} - \boldsymbol{v}^{(1)}\boldsymbol{v}^{(1)\top}$ by adding one additional row and column as follows:

$$\tilde{\mathbf{A}} = \begin{bmatrix} \bar{\mathbf{A}} & \bar{\boldsymbol{a}}_e(x) \\ \bar{\boldsymbol{a}}_e(x)^\top & \alpha(x) \end{bmatrix} \quad \text{with} \quad \bar{\boldsymbol{a}}_e(x) = (\mathbb{I} - \boldsymbol{v}^{(1)}\boldsymbol{v}^{(1)\top})\boldsymbol{a}_e(x), \tag{1.4}$$

and where the diagonal extension is $\alpha(x) = 1/m_e(x) - m_e(x)/(\mathbf{1}^\top \boldsymbol{m})$ with $x \in X$ in light of (1.1). Then, this extended empirical kernel (1.4) can serve to define the extension of the SDP embedding. Indeed, the proposed interpolation formula is the *normalized* Nyström extension [29],

$$\chi_e^{(\ell)}(x) = \sqrt{d(x)}\frac{\bar{\boldsymbol{a}}_e^\top(x)\boldsymbol{\chi}^{(\ell)}}{\sqrt{\bar{\boldsymbol{a}}_e^\top(x)\boldsymbol{B}_\star\bar{\boldsymbol{a}}_e(x)}} \text{ for all } x \in \mathcal{D}_X, \tag{1.5}$$

where $\mathcal{D}_X = \{x \in X | \boldsymbol{B}_\star\bar{\boldsymbol{a}}_e(x) \neq 0\}$. While the numerator in (1.5) is similar to the Nyström extension, the denominator here is rather different since it can be interpreted as a spherical normalization so that $\sum_{\ell=1}^{r}(\chi_e^{(\ell)}(x))^2 = d(x)$. In a word, this particular form of the out-of-sample extension is obtained thanks to the optimality conditions of (SDP). Our second

[3]The out-of-sample extension problem could also be addressed for (VP). We leave this question for future work.

main result, Theorem 3.2, given in section 3.2, indicates that the interpolation (1.5) is defined *almost everywhere* on X hereafter; that is, its domain is $\mathcal{D}_X = X \setminus X_0$ where X_0 is negligible. In essence, this is mainly due to the properties of the Gaussian kernel in (1.3), which belongs to a Hilbert space of analytic functions. Explicitly, the extended embedding then reads

$$x \mapsto \boldsymbol{\Xi}_e(x) = [\chi_e^{(1)}(x), \ldots, \chi_e^{(r)}(x)]^\top.$$

As a consequence, the extension of the matrix $\mathbf{B}_\star$ to a kernel function can be defined by $b_{\star,e}(x,y) = \boldsymbol{\Xi}_e(x)\boldsymbol{\Xi}_e(y)^\top$ for all $(x,y) \in \mathcal{D}_X \times \mathcal{D}_X$. The latter expression defines a data-dependent kernel function which has been estimated from the sample of n points in an unsupervised way. In this paper, we choose the upper bound on the diagonal as $d(x) = \alpha(x)$ in (1.4); this choice is motivated by Lemma 3.3 in section 3.2.

In Figure 1, the effect of outliers on diffusion maps and SDP embedding is illustrated on an artificial example, where we observe that the outliers remain far from the denser clusters for the SDP embedding, while their positions are not fixed in the diffusion maps embedding. Other illustrations of dimensionality reduction are given in section 5. Furthermore, although this is not an example of dimensionality reduction, the solution of a discrete approximation of (VP) for a simple toy model is also illustrated in Figures 2 and 3, where $X = [-1,1]$. Then, the square $[-1,1]^2$ is discretized in a square grid of 2001^2 points, and the result of the discrete optimization problem is displayed in Figure 2 for $\sigma = 0.1$. Figure 3 illustrates the out-of-sample extension.

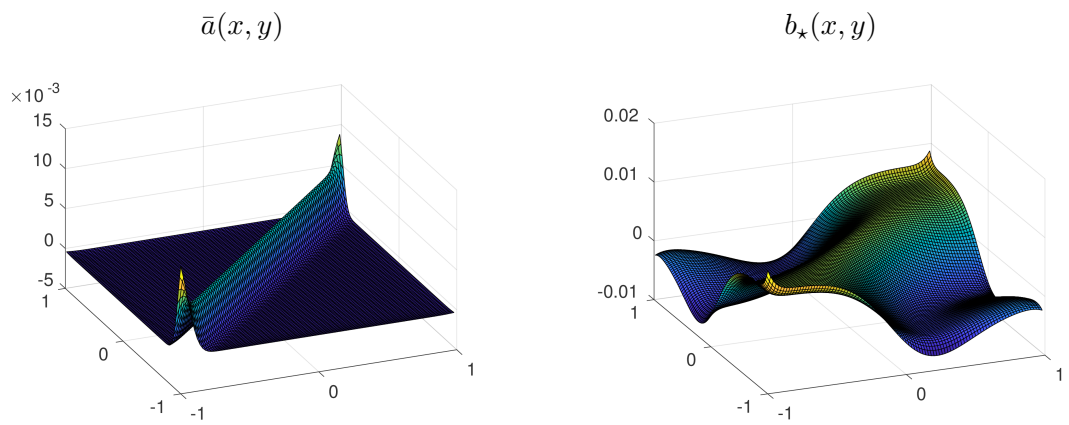

Figure 2. *For $\sigma = 0.1$, on the left, $\bar{a}(x,y)$ and, on the right, $b_\star(x,y)$ are plotted on the square $[-1,1]^2$. This plot illustrates the shape of the numerical solution of* (VP).

1.2.2. Kernelized variational problem with smoothness assumptions. The construction described hereinabove does not make assumptions on the smoothness of the integral operator in (VP). In particular, establishing a connection between (VP) and its discretized counterpart (SDP) is non-trivial. It is therefore advantageous to make additional assumptions on the integral kernel $b(x,y)$ of the estimated nuclear operator in order to leverage the well-known framework of estimation in an RKHS. Indeed, this setting is convenient for analyzing the discretization error thanks to the representer of [12]. Our discussion of the kernelized variational problem relies on several key techniques of [21], which addresses a different type of problem.

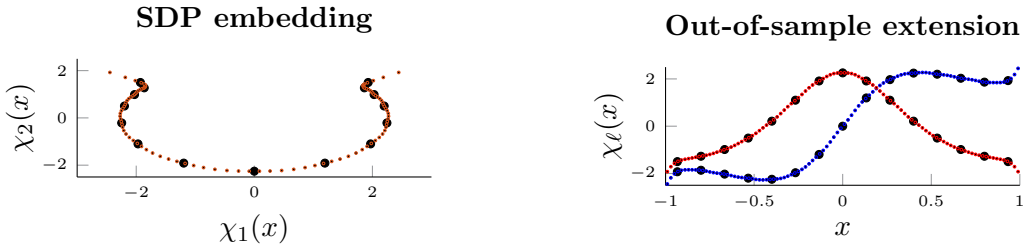

Figure 3. *Illustration of the extrapolation of the SDP embedding for a sparser sampling of the interval* $[-1, 1]$*, which accompanies Figure 2. The large black points denote training points, while the smaller colored dots are out-of-sample points. This indicates the usefulness of the out-of-sample extension. On the right-hand side, red points correspond to* $\ell = 1$ *and blue points to* $\ell = 2$*.*

For defining the kernelized problem, we consider that X is an open ℓ_2-ball of radius $R > 0$ in $\mathbb{R}^d$, so that the domain's boundary is "smooth." Denote by $k(x, y)$ the kernel of an RKHS $\mathcal{H}_k$, which we assume to be strictly positive definite, and let $\phi(x) = k(x, \cdot) \in \mathcal{H}_k$ be the canonical feature map, so that $\langle \phi(x), \phi(y) \rangle = k(x, y)$. Also, we assume that this kernel is continuous and uniformly bounded: $k(x, x) \leq \kappa^2$ for all $x \in X$ and with $\kappa > 0$. It is then common (see, e.g., [18, section 3]) to define the restriction operator $S : \mathcal{H}_k \to L^2(X)$ as $(Sg)(x) = g(x)/\sqrt{|X|}$, whereas its adjoint $S^* : L^2(X) \to \mathcal{H}_k$ is given by $S^*h = \int_X h(x)\phi(x)\mathrm{d}x/\sqrt{|X|}$. Thus, the operator SS^* belongs to $\mathcal{S}(L^2(X))$, and its integral kernel is $k(x, y)$. This remark motivates the following restriction: we can choose to estimate an operator $B \in \mathcal{S}(L^2(X))$ of the form

$$B = S\mathbb{B}S^*, \text{ where } \mathbb{B} \in \mathcal{S}(\mathcal{H}_k) \text{ is such that } \mathbb{B} \succeq 0.$$

By using the definition of the restriction operator, we see that B has the integral kernel $b(x, y) = \langle \phi(x), \mathbb{B}\phi(y) \rangle$, which is well defined for all x and y in X.

Smoothed variational problem. Now, we consider that $(\mathcal{H}_k, \langle \cdot, \cdot \rangle)$ is the Sobolev space $W_2^s(X)$ with $s > d/2$, as proposed, for instance, in [18]. First, for an integer m such that $0 < m < s - d/2$, we know that the space $W_2^s(X)$ is embedded in $\mathcal{C}^m(X)$, as shown in [21, Proposition 1]. More precisely, we can associate to each element of $W_2^s(X)$ a representative in $\mathcal{C}^m(X)$. Thus, we can define a kernelized variational problem

$$\text{(kVP)} \qquad \sup_{\mathbb{B} \in \mathcal{S}(\mathcal{H}_k)} f(\mathbb{B}) \triangleq \mathrm{Tr}(S\mathbb{A}S^*S\mathbb{B}S^*) \text{ s.t. } \mathbb{B} \succeq 0 \text{ and } \langle \phi(x), \mathbb{B}\phi(x) \rangle = d(x) \text{ for all } x \in X,$$

where we assume here that $d(x)$ can be written as $d(x) = \sum_{j=1}^p w_j(x)^2$ with p a finite integer, for all $x \in X$ with $w_j \in W_2^s(X)$ for all $1 \leq j \leq d$. The latter assumption makes sure that there exists a *psd* finite rank $\mathbb{B}_\star \in \mathcal{S}(\mathcal{H}_k)$ such that $d(x) = \langle \phi(x), \mathbb{B}_\star\phi(x) \rangle$ for all $x \in X$; see the proof of Corollary 1 in [21]. Next, to have an analogue of $\bar{A}$ in (VP), we choose

$$\mathbb{A} = \mathbb{I}_{\mathcal{H}_k} - u \otimes \bar{u} \text{ with } \|u\|_{\mathcal{H}_k} = 1,$$

where the second term is defined as follows: $(u \otimes \bar{u})g = \langle u, g \rangle u$ for all $g \in \mathcal{H}_k$. Still in analogy with (VP), we choose $u \in \mathcal{H}_k$ such that $u(x) > 0$ for all $x \in X$. Hence, we have the integral

operator

$$S\mathbb{A}S^*h(x) = \frac{1}{|X|}\int_X \Big(k(x,y) - u(x)u(y)\Big)h(y)\mathrm{d}y.$$

By construction, $\mathbb{A}$ is the projector onto $u^\perp$ in the RKHS, and, therefore, $0 \preceq \mathbb{A} \preceq \mathbb{I}_{\mathcal{H}_k}$.

Discretized objective. An additional advantage of this kernelized formulation is that it allows for a natural discretization. Indeed, given a discrete set $\widehat{X} = \{x_1, \dots, x_n\}$, we can define a discrete analogue of the restriction operator S as $S_n : \mathcal{H}_k \to \mathbb{R}^n$ such that $S_n g = (1/\sqrt{n})[g(x_1), \dots, g(x_n)]^\top$. Its adjoint is given by $S_n^*\mathbf{v} = (1/\sqrt{n})\sum_{i=1}^n \mathbf{v}_i\phi(x_i)$ for all $\mathbf{v} \in \mathbb{R}^n$. With the help of these definitions, the kernel matrix writes as $S_nS_n^* = (1/n)\mathbf{K}$. In layman's terms, if the $\{x_1, \dots, x_n\}$ are sampled independently from the uniform probability measure on X and if n is large enough, the event $\|S^*S - S_n^*S_n\|_{op} \lesssim 1/\sqrt{n}$ occurs with high probability (up to log terms). This can be shown thanks to a matrix Bernstein inequality; see, e.g., [19, Proposition 12] or the proof of Proposition 4.1 in what follows. With high probability for the same event, we have the informal discretization error bound

$$|\mathrm{Tr}\,(S\mathbb{A}S^*S\mathbb{B}S^*) - \mathrm{Tr}\,(S_n\mathbb{A}S_n^*S_n\mathbb{B}S_n^*)| \lesssim \mathrm{Tr}(\mathbb{B})/\sqrt{n} \tag{1.6}$$

if n is large enough; see Proposition 4.1 for a formal statement. Thus, we consider

(kVP-n)

$$\sup_{\mathbb{B}\in\mathcal{S}(\mathcal{H}_k)} f_n(\mathbb{B}) \triangleq \frac{1}{n}\mathrm{Tr}(\overline{\mathbf{K}}S_n\mathbb{B}S_n^*) \text{ s.t. } \mathbb{B} \succeq 0 \text{ and } \langle\phi(x), \mathbb{B}\phi(x)\rangle = d(x) \text{ for all } x \in X.$$

Above, $S_n\mathbb{B}S_n^*$ is a matrix whose element (i,j) is $\frac{1}{n}\langle\phi(x_i), \mathbb{B}\phi(x_j)\rangle$ and $\overline{\mathbf{K}} = \mathbf{K} - \mathbf{u}\mathbf{u}^\top$ is a matrix with $\mathbf{u} = [u(x_1)\dots, u(x_n)]^\top$. Note that thanks to the constraints, the objective of (kVP-n) is still upper bounded. The equality constraints are now discretized since solving the above optimization problem with continuous equality constraints is non-trivial.

Scattered constraints and regularization. To deal with the equality constraints, we use a sampling approach, inspired by [21, section 5.1], which requires additional regularity assumptions. The idea goes as follows. If the function $x \mapsto \langle\phi(x), \mathbb{B}\phi(x)\rangle - d(x)$ is smooth enough on X, restricting the equality constraints on $\widehat{X}$ yields a controlled approximation provided that $\widehat{X}$ covers X well enough. A key fact is that the function $x \mapsto \langle\phi(x), \mathbb{B}\phi(x)\rangle$ belongs to $W_2^s(X)$ since $\mathbb{B}$ is *psd* and nuclear, as shown in [21, Lemma 9]. After subsampling the constraints and complementing the objective with a regularization term for some $\lambda > 0$, the resulting problem reads

(reg-kVP-n)

$$\sup_{\mathbb{B}\in\mathcal{S}(\mathcal{H}_k)} f_n(\mathbb{B}) - \lambda\,\mathrm{Tr}(\mathbb{B}) \text{ subject to } \mathbb{B} \succeq 0 \text{ and } \langle\phi(x_i), \mathbb{B}\phi(x_i)\rangle = d(x_i) \text{ for all } i \in [n].$$

Notice that we could a priori consider a different discrete subset of X for enforcing scattered constraints; by taking the same discrete set $\widehat{X}$ for discretizing the constraints as for discretizing the objective of (kVP), we reduce the size of the final discrete optimization problem and we make sure that the constrained objective of (reg-kVP-n) is bounded from above. The extra regularization term penalizes large values of $\mathrm{Tr}(\mathbb{B})$ which helps to improve the bound given at the right-hand side (RHS) of (1.6). The constraints discretization entails an error which can be analyzed thanks to results about functions with scattered zeros [28]. Recall that the

domain X is given here by an open ℓ_2-ball of radius $R > 0$. Still following [21], if we have $d(x) - \langle \phi(x), \mathbb{B}\phi(x)\rangle = 0$ for all $x \in \widehat{X}$, then, in this case, it holds that

$$|d(x) - \langle \phi(x), \mathbb{B}\phi(x)\rangle| \lesssim h_{\widehat{X},X}^m \Big(\beta_m \operatorname{Tr}(\mathbb{B}) + |d|_{X,m}\Big) \text{ for all } x \in X, \tag{1.7}$$

for some positive number β_m, provided that the fill distance $h_{\widehat{X},X} = \max_{x\in X} \min_{i\in[n]} \|x_i - x\|_2$ is small enough; see Proposition 4.3 below. The RHS of (1.7) is bounded with high probability by a decreasing function of n. We now explain how. Let $\delta \in (0,1)$. If n is large enough (see Proposition 4.4) and if the elements of $\widehat{X}$ are sampled independently and uniformly from X, then with probability at least $1-\delta$ it holds that, up to a factor with a logarithmic dependence on n/δ, $h_{\widehat{X},X} \lesssim n^{-1/d}$, as given in [21, Lemma 4].

Representer theorem and finite dimensional problem. The representer theorem of [12] applies to the regularized problem (reg-kVP-n) so that the optimal operator can be found in the form

$$\mathbb{B} = \sum_{i,j=1}^n \mathbf{C}_{ij} \phi(x_i) \otimes \overline{\phi(x_j)} \text{ for some matrix } \mathbf{C} \succeq 0, \tag{1.8}$$

where we used the notation $\overline{\phi(x)} g = g(x)$ for all $g \in \mathcal{H}_k$ and $x \in X$. The representer (1.8) is useful to turn the variational problem into a finite dimensional optimization problem, as we explain in what follows. First, we define the Cholesky decomposition[4] $\mathbf{K} = \mathbf{R}^\top \mathbf{R}$. Second, we define the operator $V : \mathcal{H}_k \to \mathbb{R}^n$ as $V = \sqrt{n}(\mathbf{R}^{-1})^\top S_n$, which satisfies $VV^* = \mathbb{I}_n$ and is such that V^*V is the orthogonal projector onto the span of $\phi(x_i)$ for all $1 \le i \le n$. Hence, following [21, Lemma 2], a finite dimensional feature map associated to $\phi(x_i)$ writes simply as $\mathbf{\Phi}_i = V\phi(x_i) \in \mathbb{R}^n$, and note that the matrix $\mathbf{\Phi}$, whose ith column is $\mathbf{\Phi}_i$, is identically equal to $\mathbf{R}$. We do the change of variable $\mathbf{B} = \mathbf{R}\mathbf{C}\mathbf{R}^\top$, so that $\mathbb{B} = V^*\mathbf{B}V$. Consequently, the problem (reg-kVP-n) reduces to

$$\max_{\mathbf{B}\succeq 0} \frac{1}{n^2} \operatorname{Tr}(\overline{\mathbf{K}}\mathbf{\Phi}^\top \mathbf{B}\mathbf{\Phi}) - \lambda \operatorname{Tr}(\mathbf{B}) \text{ subject to } \mathbf{\Phi}_i^\top \mathbf{B}\mathbf{\Phi}_i = d(x_i) \text{ for all } i \in [n], \tag{1.9}$$

where the objective is equal to $f_n(\mathbb{B}) - \lambda \operatorname{Tr}(\mathbb{B})$, while the constraints are $\langle \phi(x_i), \mathbb{B}\phi(x_i)\rangle = \mathbf{\Phi}_i^\top \mathbf{B}\mathbf{\Phi}_i$ for all $i \in [n]$ for $\mathbb{B} = V^*\mathbf{B}V$.

Further, the out-of-sample formula is completely natural in this RKHS setting; that is,

$$b(x,y) = \sum_{i,j=1}^n k(x,x_i)[\mathbf{R}^{-1}\mathbf{B}\mathbf{R}^{-1\,\top}]_{ij} k(x_j,y) \text{ with } b(x_i,x_j) = \mathbf{\Phi}_i^\top \mathbf{B}\mathbf{\Phi}_j \text{ for all } i,j \in [n],$$

by construction. Algorithmically, the finite dimensional problem can be put in a form similar to (SDP) by performing the change of variables $\mathbf{B}' = \mathbf{\Phi}^\top \mathbf{B}\mathbf{\Phi}$, so that we have specifically

$$\max_{\mathbf{B}'\succeq 0} \operatorname{Tr}\left((\overline{\mathbf{K}} - \lambda n^2 \mathbf{K}^{-1})\mathbf{B}'\right) \text{ subject to } \mathbf{B}'_{ii} = d(x_i) \text{ for all } i \in [n].$$

Thus, although the definition of $\bar{\mathbf{A}}$ differs in (SDP) with respect to $\overline{\mathbf{K}} - \lambda n^2 \mathbf{K}^{-1}$, the above discrete optimization problem also involves a subtracted kernel matrix, and the amount of subtraction can be modified by varying the parameter $\lambda > 0$.

[4] Note that $\mathbf{K}$ is non-singular almost surely for the Sobolev kernel.

Our Theorem 4.5 below states that the operator associated with the solution of the discrete problem achieves a good approximation of the full kernelized problem with high probability; see section 4.

1.3. Outline. The rest of this paper is organized as follows. In section 2, we prove the existence of a well-defined solution of (VP) given in terms of an integral kernel of a *psd* symmetric nuclear operator in the absence of smoothness assumptions. Section 3 discusses the reasons why the discrete version (SDP) yields often to a dimensionality reduction, whereas the out-of-sample extension formula is also derived. The results related to the kernelized problem (kVP) are provided in section 4. Finally, empirical case studies are reported in section 5. The numerical method is given in the appendix together with basic elements of operator theory.

2. Discussion of the variational problem.

To start, we explain why the diagonal of a Mercer kernel of a *psd* operator in $\mathcal{S}$ cannot vanish unless it is trivially zero, so that the diagonal constraint in (VP) is not void.

In general, the kernel of a Hilbert–Schmidt operator is an element of $L^2(X\times X)$ which is not necessarily defined pointwisely everywhere. However, from the Mercer theorem of Steinwart and Scovel [22] (Theorem A.1 in the appendix), we know that for each *psd* symmetric nuclear operator K, there exists an integral kernel $k_M : X \times X \to \mathbb{R}$ which is defined pointwise, with eigendecomposition

$$k_M(x,y) = \sum_{\ell\geq 1} \lambda^{(\ell)}\phi^{(\ell)}(x)\phi^{(\ell)}(y) \text{ for all } x, y \in X.$$

Moreover, there exists an integral formula for the trace of a symmetric nuclear operator. Indeed, the trace of a *psd* symmetric nuclear K is given by the following integral formula:

$$\operatorname{Tr}(K) = \sum_{\ell\geq 1} \lambda^{(\ell)} = \sum_{\ell\geq 1} \lambda^{(\ell)} \int_X |\phi^{(\ell)}(x)|^2 \mathrm{d}x = \int_X \sum_{\ell\geq 1} \lambda^{(\ell)} |\phi^{(\ell)}(x)|^2 \mathrm{d}x, \tag{2.1}$$

where the integral and series have been exchanged thanks to Beppo Levi's theorem. This shows additionally that $k_M(x,x)$ is an element of $L^1(X)$, which is also defined for all $x \in X$, in light of the Mercer theorem. Moreover, the function $k_M(x,x)$ cannot vanish everywhere if K is not identically zero, since its trace is $\int_X k_M(x,x)\mathrm{d}x = \operatorname{Tr}(K)$, as given by (2.1).

We now state formally our main result about the existence of a *psd* symmetric nuclear operator attaining the supremum in (VP), and we introduce the proof techniques.

Theorem 2.1 (existence of an optimal nuclear operator). *Let $X = [-c,c]^d$ with $c > 0$. Let $\bar{A}$ be a symmetric nuclear operator and $d : X \to \mathbb{R}_{>0}$ be a continuous positive function in $L^1(X)$. There exist a* psd *symmetric nuclear operator $B_\star$ and an associated Mercer kernel $b^\star_M(x,y)$ such that*

$$\operatorname{Tr}(\bar{A}B_\star) = \sup_{B\in\mathcal{S}} \operatorname{Tr}(\bar{A}B) \text{ subject to } B \succeq 0 \text{ and } b_M(x,x) \leq d(x) \quad \text{for almost all } x \in X,$$

and $b^\star_M(x,x) \leq d(x)$ almost everywhere.

In order to prove Theorem 2.1, we first introduce a useful averaging technique and give some intermediate results about weak compactness.

2.1. Averaging kernels of nuclear operators. A key point is that the diagonal of an element of $L^2(X \times X)$ is not necessarily defined. In order to obtain a representative with a well-defined diagonal, we use an averaging technique developed by Brislawn in [4]. We recall that $X = [-c, c]^d$. Define $C_r = [-r, r]^d$. Then, the average of a locally L^1 function, namely $f \in L^1_{\text{loc}}(\mathbb{R}^d)$, is defined as follows:

$$\mathcal{A}_r f(x) = \frac{1}{|C_r|} \int_{C_r} f(x+s)\mathrm{d}s,$$

where $|C_r| = (2r)^d$, and $r > 0$ is small enough so that $x + s \in X$ for all $s \in [-r, r]^d$. The differentiation theorem of Lebesgue [4] states that $\lim_{r\to 0} \mathcal{A}_r f(x) = f(x)$ almost everywhere. More generally, the average of $f \in L^1_{\text{loc}}(\mathbb{R}^{2d})$ is $\mathcal{A}^{(2)}_r f(x,y) = \frac{1}{|C_r|^2} \int_{C_r \times C_r} f(x+s, y+t)\mathrm{d}s\mathrm{d}t$. Notice that C_r is chosen to be a cube for the following reason: a cube in $\mathbb{R}^{2d}$ is the product of two cubes in $\mathbb{R}^d$ with the same sides. Hence, if $b(x, y)$ is a kernel of a *psd* symmetric nuclear operator, the limit

$$\tilde{b}(x,y) = \lim_{r\to 0} \mathcal{A}^{(2)}_r b(x,y)$$

is defined pointwise almost everywhere on $X \times X$. Again, by applying the differentiation theorem of Lebesgue, the kernel and the averaged kernel satisfy $\tilde{b}(x, y) = b(x, y)$ almost everywhere on $X \times X$. Therefore, the averaging procedure is a way of choosing a pointwise representative of $b(x, y)$. Now, we consider the diagonal elements of the averaged kernel and follow an argument by Brislawn [4]. By using a Mercer representation of b, we find

$$\tilde{b}(x,x) = \lim_{r\to 0} (\mathcal{A}^{(2)}_r b)(x,x) = \lim_{r\to 0} \sum_{\ell \geq 1} \lambda^{(\ell)} |\mathcal{A}_r \phi^{(\ell)}(x)|^2 \overset{\star}{=} \sum_{\ell \geq 1} \lambda^{(\ell)} |\phi^{(\ell)}(x)|^2 \tag{2.2}$$

almost everywhere on X. Notice that we used in $\overset{\star}{=}$ that the series $\sum_{\ell\geq 1} \lambda^{(\ell)} |\mathcal{A}_r \phi^{(\ell)}(x)|^2$ converges absolutely and uniformly with respect to $r \in [0, +\infty)$ almost everywhere on X. This follows from the Hardy–Littlewood maximal theorem, as explained in detail in the proof of Theorem 3.1 in [4]. To summarize this subsection in other words, we found indeed that the diagonal of the averaged kernel, as given in (2.2), is defined almost everywhere.

2.2. Useful compactness result. For the proof of Theorem 2.1, we also need the following result, which relies on a classical compactness argument.

Lemma 2.2. *Let $s_\star$ be the supremum in Theorem 2.1. There is a sequence of nuclear operators $(B_\ell)_{\ell\geq 1}$ in $\mathcal{S}$ satisfying $B_\ell \succeq 0$ and $b_\ell(x,x) \leq d(x)$ such that $\mathrm{Tr}(\bar{A}B_\ell) \to s_\star$ when $\ell \to +\infty$. Furthermore, there exist a subsequence $(B_{\ell_k})_{k\geq 1}$ and a nuclear operator $B_\star \in S$ such that $\mathrm{Tr}\,(TB_{\ell_k}) \to \mathrm{Tr}\,(TB_\star)$, when $k \to +\infty$ for all compact operators T.*

Proof of Lemma 2.2. The existence of the sequence $(B_\ell)_{\ell\geq 1}$ is a consequence of the definition of the supremum. For all integers $\ell \geq 1$, since $B_\ell \succeq 0$, the trace of B_ℓ is its nuclear norm: $\|B_\ell\|_\star = \mathrm{Tr}(B_\ell)$. Therefore, the sequence $(B_\ell)_{\ell\geq 1}$ is within the ball $\|B\|_\star \leq \int_X d(x)\mathrm{d}x$. We now use the compactness for the weak* topology. Indeed, the space of nuclear operators $\mathcal{B}_1(L^2(X))$ is the dual of the space of compact operators $\mathcal{B}_0(L^2(X))$, and the duality pairing $\langle \cdot, \cdot \rangle$ is given by the trace, i.e., $\langle B, T\rangle = \mathrm{Tr}(BT)$ for all $T \in \mathcal{B}_0(L^2(X))$ and $B \in \mathcal{B}_1(L^2(X))$.

Since $\mathcal{B}_0(L^2(X))$ is a Banach space, we know that the ball of $\mathcal{B}_0(L^2(X))^*$ is compact for the weak* topology thanks to the Banach–Alaoglu theorem. Hence, there exist a subsequence $(B_{\ell_k})_{k\geq 1}$ and a nuclear operator $B_\star$ with $\|B_\star\|_\star \leq \int_X d(x)\mathrm{d}x$ such that we have $\lim_{k\to+\infty} \mathrm{Tr}\,(TB_{\ell_k}) = \mathrm{Tr}\,(TB_\star)$ for all compact operators T. ∎

2.3. Main part of the proof of Theorem 2.1. The proof structure goes as follows. First, we show that the supremum is attained by a nuclear operator which is self-adjoint and *psd*. Next, we prove that this operator admits an integral kernel which is bounded almost everywhere by an envelope function on $X \times X$. Finally, the averaging technique of section 2.1 is used to find an integral kernel satisfying the bound on its diagonal.

Proof of Theorem 2.1. Existence. By choosing different compact operators T in Lemma 2.2, we show different properties of $B_\star$. First, let T be the Hilbert–Schmidt integral operator of kernel $t(x,y) = g(x)f(y)$ with $f, g \in L^2(X)$ and $(B_{\ell_k})_{k\geq 1}$ such as in Lemma 2.2. Then, we find

$$\mathrm{Tr}\,(TB_{\ell_k}) = \langle g, B_{\ell_k} f\rangle_{L^2(X)} = \langle B_{\ell_k} g, f\rangle_{L^2(X)},$$

since B_{ℓ_k} is symmetric. By taking the limit $k \to +\infty$ in the above expression, we find that $B_\star$ is also symmetric. In particular, by taking the previous identity with $f = g$, since $0 \leq \langle f, B_{\ell_k} f\rangle_{L^2(X)}$ for all $f \in L^2(X)$, we show that $B_\star$ is *psd*. Hence, $B_\star$ is a *psd* symmetric nuclear operator, and, in view of Theorem A.1, we know that $B_\star$ admits an integral kernel given by $b_{\star,M}(x,y)$ for all $x, y \in X$, with a well-defined diagonal $b_{\star,M}(x,x)$ for all $x \in X$.

Bounding the kernel. Consequently, we have a sequence of kernels $(b_{\ell_k})_k$ in $L^2(X \times X)$ weakly convergent, that is, $\langle b_{\ell_k}, t\rangle_{L^2} \xrightarrow{k\to+\infty} \langle b_\star, t\rangle_{L^2}$ for all $t(x,y) \in L^2(X\times X)$. This sequence is bounded since $\|B_{\ell_k}\|_{HS} \leq \|B_{\ell_k}\|_*$, and therefore we can apply the Banach–Saks theorem. Indeed, we can construct the Cesaro means $\hat{b}_m = (1/m)\sum_{k=1}^m b_{n_k}$ such that we have a strong convergence $\|\hat{b}_m - b_\star\|_{L^2} \to 0$ for $m \to \infty$, and $\hat{b}_m(x,x) \leq d(x)$ almost everywhere. Naturally, $\hat{b}_m \in \mathcal{S}$. Since this sequence converges in $L^2(X \times X)$, there exists a subsequence $(\hat{b}_{m_k})_k$ such that we have a pointwise convergence $\hat{b}_{m_k}(x,y) \to b_\star(x,y)$ for $k \to \infty$ almost everywhere on $X \times X$. By construction, since each kernel is in an envelope such that $|\hat{b}_{m_k}(x,y)| \leq \sqrt{d(x)d(y)}$, we have by taking the limit $k \to \infty$ that $\hat{b}_\star(x,y) \leq \sqrt{d(x)d(y)}$ almost everywhere on $X \times X$.

Averaging the kernel. Now, we use the averaging techniques of section 2.1 in order to determine a representative with a well-defined diagonal. Let us define $\tilde{b}_\star(x,y) = \lim_{r\to 0}(\mathcal{A}_r^{(2)}\hat{b}_\star)(x,y)$ for almost all $(x,y) \in X \times X$. Then, in light of (2.2), its diagonal is upper bounded as follows:

$$\tilde{b}_\star(x,x) = \lim_{r\to 0}(\mathcal{A}_r^{(2)}\hat{b}_\star)(x,x) \leq \lim_{r\to 0} \frac{1}{|C_r|^2}\int_{C_r}\int_{C_r} \sqrt{d(x+s)d(y+t)}\mathrm{d}s\mathrm{d}t = d(x)$$

almost everywhere on X, where we used the envelope obtained hereinabove. This proves the desired result. ∎

3. SDP embedding and out-of-sample extension.

3.1. Intuition for the dimensionality reduction. Empirically, the solution of (SDP) is often a low rank matrix. In this section, we first provide a theoretical motivation for this observation. Finding the solution of (SDP) is in fact equivalent to solving a convex relaxation of a rank minimization problem in terms of the nuclear norm. Indeed, since the nuclear norm

is the convex envelope of the rank, Proposition 3.1 below illustrates why in practice we can expect the optimal solution of (SDP) to have a low rank.

Proposition 3.1 (equivalence with a nuclear norm minimization). *Consider* (SDP) *with* $\bar{\mathbf{A}} \in \mathbb{R}^{n\times n}$ *a psd matrix with a maximal eigenvalue strictly smaller than* 1, *and let* $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ *be the invertible matrix with orthogonal columns such that* $\mathbb{I} - \bar{\mathbf{A}} = \mathbf{\Sigma}\mathbf{\Sigma}^\top$. *Then, the optimal solution* $\mathbf{X}_\star$ *of*

$$\min_{\mathbf{X}\succeq 0} \|\mathbf{X}\|_*, \text{ subject to } \operatorname{diag}\left((\mathbf{\Sigma}^{-1})^\top \mathbf{X}\mathbf{\Sigma}^{-1}\right) = \boldsymbol{d},$$

has the same rank as the optimal solution $\mathbf{B}_\star$ *of* (SDP) *and is given by* $\mathbf{X}_\star = \mathbf{\Sigma}^\top \mathbf{B}_\star \mathbf{\Sigma}$.

Proof of Proposition 3.1. Notice first that maximizing $\operatorname{Tr}(\bar{\mathbf{A}}\mathbf{B})$ over $\mathbf{B} \succeq 0$ such that $\operatorname{diag}(\mathbf{B}) = \mathbf{d}$ is equivalent to minimizing $\operatorname{Tr}\left(\mathbf{B}(\mathbb{I} - \bar{\mathbf{A}})\right)$ since $\operatorname{Tr}(\mathbf{B})$ is fixed by the constraints. Then, we remark that $(\mathbb{I} - \bar{\mathbf{A}}) \succ 0$ because the maximal eigenvalue of $\bar{\mathbf{A}}$ is strictly smaller than 1 by assumption. Hence, a diagonalization procedure yields $\mathbb{I} - \bar{\mathbf{A}} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, where the diagonal matrix $\mathbf{D}$ is strictly positive definite. Therefore, we can also write $\mathbb{I} - \bar{\mathbf{A}} = \mathbf{\Sigma}\mathbf{\Sigma}^\top$, where $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ is invertible. As a consequence, the objective of the minimization problem

$$\min_{\mathbf{B}\succeq 0} \operatorname{Tr}\left(\mathbf{B}(\mathbb{I} - \bar{\mathbf{A}})\right), \text{ subject to } \operatorname{diag}(\mathbf{B}) = \mathbf{d},$$

can be written $\operatorname{Tr}\left(\mathbf{B}(\mathbb{I} - \bar{\mathbf{A}})\right) = \operatorname{Tr}\left(\mathbf{\Sigma}^\top \mathbf{B}\mathbf{\Sigma}\right)$, where $\mathbf{\Sigma}^\top \mathbf{B}\mathbf{\Sigma} \succeq 0$. Then, by performing the change of variables $\mathbf{X} = \mathbf{\Sigma}^\top \mathbf{B}\mathbf{\Sigma}$, we can rephrase the minimization problem as follows:

$$\min_{\mathbf{X}\succeq 0} \operatorname{Tr}\left(\mathbf{X}\right), \text{ subject to } \operatorname{diag}\left((\mathbf{\Sigma}^{-1})^\top \mathbf{X}\mathbf{\Sigma}^{-1}\right) = \mathbf{d}.$$

Finally, we notice that $\|\mathbf{X}\|_* = \operatorname{Tr}(\sqrt{\mathbf{X}\mathbf{X}^\top}) = \operatorname{Tr}(\mathbf{X})$ since $\mathbf{X}$ is symmetric and $\mathbf{X} \succeq 0$. ∎

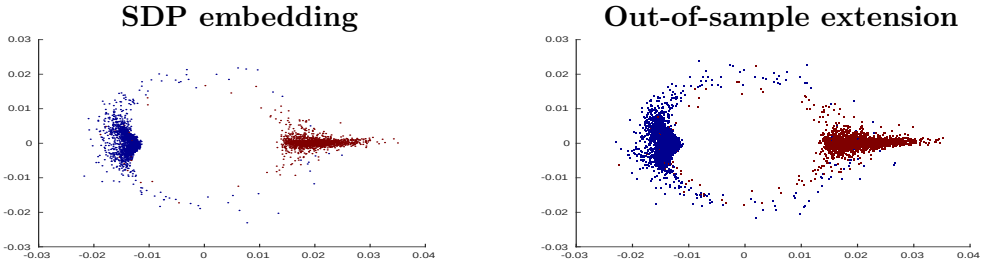

Figure 4. *SDP embedding for the training set of the MNIST dataset for the digits* 1 *(blue) and* 4 *(red) with a bandwidth* $\sigma = 10$. *The total number of points here is* 12584. *The embedding dimension is two, i.e.,* $\operatorname{rank}(\mathbf{B}_\star) = 2$. *On the LHS, a subsample of* 3775 *points is embedded by using the extension formula* (1.5), *while the figure on the RHS is the embedding of a subsample of* 8809 *points thanks to the out-of-sample extension. A* k*-nearest neighbors classifier was trained with* $k = 5$ *on the first sample, yielding a test error on the second sample of* 1%. *This indicates that the out-of-sample extension can be useful in supervised learning.*

3.2. Out-of-sample extension. The out-of-sample extension of an eigenvector is a function given in (1.5) which reduces to the initial vector when it is evaluated in the sample as stated in Theorem 3.2, which we prove hereafter.

Theorem 3.2 (out-of-sample extension). *Let $\boldsymbol{\chi}^{(\ell)}$ be an eigenvector of the solution of (SDP) with non-zero eigenvalue $\lambda^{(\ell)}$ for $\ell \in [r]$, such that $\|\boldsymbol{\chi}^{(\ell)}\|_2^2 = \lambda^{(\ell)}$. Let its extension $\chi_e^{(\ell)}(x)$ be given by (1.5), where $[\bar{\boldsymbol{a}}_e(x_i)]_j = [\bar{\mathbf{A}}]_{ij}$. Then, the following properties hold:*

(i) $\chi_e^{(\ell)}(x_i) = [\boldsymbol{\chi}^{(\ell)}]_i$ *for all* $i \in [n]$.

(ii) *If $\bar{\boldsymbol{a}}_e(x)$ is given by (1.3) and (1.4), then $\chi_e^{(\ell)}(x)$ is defined almost everywhere on $\mathbb{R}^d$.*

Proof. (i) Let $\boldsymbol{B}_\star$ be the solution of (SDP). Consider the dual certificate of (SDP), which is defined in more detail in Appendix B, $\mathbf{L}(\mathbf{B}) = \mathrm{Diag}(\boldsymbol{d})^{-1}\,\mathrm{ddiag}(\bar{\mathbf{A}}\mathbf{B}) - \bar{\mathbf{A}}$, and which satisfies $\mathbf{L}(\mathbf{B}_\star) \succeq 0$ as well as $\mathbf{L}(\mathbf{B}_\star)\mathbf{B}_\star = 0$. The latter gives in components

$$\frac{1}{d(x_i)}[\bar{\mathbf{A}}\mathbf{B}_\star]_{ii}[\mathbf{B}_\star]_{ik} = [\bar{\mathbf{A}}\mathbf{B}_\star]_{ik} \text{ for all } i,k \in [n]. \tag{3.1}$$

In particular, the identity $\mathbf{L}(\mathbf{B}_\star)\boldsymbol{\chi}^{(\ell)} = 0$ holds for all eigenvectors $\boldsymbol{\chi}^{(\ell)}$ of $\mathbf{B}_\star$, which gives in components

$$\frac{1}{d(x_i)}[\bar{\mathbf{A}}\mathbf{B}_\star]_{ii}[\boldsymbol{\chi}^{(\ell)}]_i = [\bar{\mathbf{A}}\boldsymbol{\chi}^{(\ell)}]_i \text{ for all } i \in [n]. \tag{3.2}$$

Let $i \in [n]$. Since $\mathbf{L}(\mathbf{B}_\star) \succeq 0$, its diagonal is non-negative, and therefore $\frac{1}{d(x_i)}[\bar{\mathbf{A}}\mathbf{B}_\star]_{ii} \geq \bar{a}(x_i,x_i) > 0$. Then, we consider the following identity:

$$\Big([\bar{\mathbf{A}}\mathbf{B}_\star]_{ii}\Big)^2 = \sum_{j=1}^n [\bar{\mathbf{A}}\mathbf{B}_\star]_{ii}[\mathbf{B}_\star]_{ij}\bar{\mathbf{A}}_{ij} \overset{\star}{=} \sum_{j=1}^n d(x_i)[\bar{\mathbf{A}}\mathbf{B}_\star]_{ij}\bar{\mathbf{A}}_{ij} = d(x_i)[\bar{\mathbf{A}}\mathbf{B}_\star\bar{\mathbf{A}}]_{ii},$$

where, in $\overset{\star}{=}$, we used (3.1). By using this identity and $[\bar{\mathbf{A}}\mathbf{B}_\star]_{ii} > 0$, we find the following equality: $[\bar{\mathbf{A}}\mathbf{B}_\star]_{ii} = \sqrt{d(x_i)}\sqrt{[\bar{\mathbf{A}}\mathbf{B}_\star\bar{\mathbf{A}}]_{ii}}$. Then, by considering (3.2), we find

$$[\boldsymbol{\chi}^{(\ell)}]_i = \sqrt{\frac{d(x_i)}{[\bar{\mathbf{A}}\mathbf{B}_\star\bar{\mathbf{A}}]_{ii}}}[\bar{\mathbf{A}}\boldsymbol{\chi}^{(\ell)}]_i \text{ for all } i \in [n],$$

which finishes the first part of the proof.

(ii) Let $\ell \in [n]$. Consider the equation $\bar{\boldsymbol{a}}_e^\top(x)\boldsymbol{\chi}^{(\ell)} = 0$. By substituting the definition of $\bar{\boldsymbol{a}}_e(x)$, we find

$$\boldsymbol{\chi}^{(\ell)^\top}\bar{\boldsymbol{a}}_e(x) = \boldsymbol{\chi}^{(\ell)^\top}(\mathbb{I} - \boldsymbol{v}^{(1)}\boldsymbol{v}^{(1)\top})\,\mathrm{Diag}(\sqrt{\boldsymbol{m}})^{-1}\mathbf{k}(x)/\sqrt{m_e(x)} = \boldsymbol{\beta}^\top\mathbf{k}(x)/\sqrt{m_e(x)},$$

where $\boldsymbol{\beta}$ is a suitable vector. Hence, an equivalent condition is $f(x) = \sum_j \beta_j k(x_j,x) = 0$. Since $k(x,y) = e^{-\|x-y\|_2^2/\sigma^2}$, we know that $f(x)$ belongs to the reproducing kernel Hilbert space (RKHS) associated to the Gaussian kernel. Hence, since the functions in this RKHS are analytic and f is not identically zero, the set $\{x \in X | f(x) = 0\}$ is a closed set whose interior is the empty set (cf. [20] and Corollary 4.44 in [22]). ∎

As explained in section 1.2, the bound on the diagonal of the integral kernel is chosen as follows: $d(x) = \bar{a}_e(x,x)$. We now prove Lemma 3.3, which states that this choice for $d(x)$ yields to a strictly positive function.

Lemma 3.3. *For all $x \in X$, we have $1/m_e(x) - m_e(x)/(\mathbf{1}^\top \mathbf{m}) > 0$.*

Proof. Denote the Gaussian kernel by $k(x,y) = \exp(-\|x-y\|_2^2/\sigma^2)$. Let $\bar{\mathbf{A}} = \mathbf{A} - \boldsymbol{v}^{(1)}\boldsymbol{v}^{(1)\top}$, with $[\mathbf{A}]_{ij} = k(x_i,x_j)/\sqrt{m_e(x_i)m_e(x_j)}$ for $i,j \in [n]$ and $m(x_i) = \sum_{j=1}^n k(x_i,x_j)$ for all $i \in [n]$, whereas $\boldsymbol{v}^{(1)} = \sqrt{\boldsymbol{m}/(\mathbf{1}^\top \boldsymbol{m})}$. Since $\bar{\mathbf{A}}$ is *psd*, we have $[\bar{\mathbf{A}}]_{ii} \geq 0$ for all $i \in [n]$, and, as a consequence, we obtain $m_e(x_i)^2 \leq k(x_i,x_i)\mathbf{1}^\top \boldsymbol{m}$ for all $i \in [n]$ Now, we consider the augmented dataset $\{x_1,\dots,x_n,x\}$ with $n+1$ point and where we define $x_{n+1} = x$. Let us denote quantities relative to the augmented set by a superscript $\cdot^{(a)}$. Naturally, in that case, we also have $m_e^{(a)2}(x_i) \leq k(x_i,x_i)\mathbf{1}^\top \boldsymbol{m}^{(a)}$ for all $i \in [n+1]$ with $m_e^{(a)}(x_i) = k(x_i,x) + m_e(x_i)$ and $\mathbf{1}^\top \boldsymbol{m}_e^{(a)} = \mathbf{1}^\top \boldsymbol{m} + 2m_e^{(a)}(x) + k(x,x)$ with $x = x_{n+1}$. Then, by using the above inequality with $i = n+1$, we find

$$\Big(k(x,x) + m(x)\Big)^2 \leq k(x,x)\left(\mathbf{1}^\top \boldsymbol{m} + 2m^{(a)}(x) + k(x,x)\right),$$

with $x = x_{n+1}$. After simplification, we find $m_e(x_{n+1})^2 \leq k(x,x)\mathbf{1}^\top \boldsymbol{m}$. Since this is true for any dataset $\{x_1,\dots,x_n,x\}$, we have the desired inequality $k(x,x)/m_e(x) - m_e(x)/(\mathbf{1}^\top \boldsymbol{m}) > 0$ for all $x \in X$. ∎

4. Kernelized SDP embedding.

4.1. Objective discretization.

We begin the discussion of the kernelized problem by stating and proving the following statistical guarantees about the objective discretization.

Proposition 4.1. *Let $\mathbb{A}$ and $\mathbb{B}$ be endomorphisms of $\mathcal{H}_k$ such that $\mathbb{B} \in \mathcal{S}(\mathcal{H}_k)$ and $\mathbb{B} \succeq 0$, while $\mathbb{A}$ is* psd *and satisfies $\|\mathbb{A}\|_{op} \leq 1$. Let $S : \mathcal{H}_k \to L^2(X)$ and $S_n : \mathcal{H}_k \to \mathbb{R}^n$ such as defined in section 1.2.2, where X is a bounded open set of $\mathbb{R}^d$. Let $\delta \in (0,1/2)$. Then, with probability at least $1-2\delta$, we have*

$$|\operatorname{Tr}(S\mathbb{A}S^*S\mathbb{B}S^*) - \operatorname{Tr}(S_n\mathbb{A}S_n^*S_n\mathbb{B}S_n^*)| \leq C_{n,\delta}\operatorname{Tr}(\mathbb{B}),$$

with $C_{n,\delta} = 2\kappa^2 c_{n,\delta} + c_{n,\delta}^2$ *and* $c_{n,\delta} = \dfrac{4\kappa^2 \log\left(\frac{2\kappa^2}{\lambda_{\max}(S^*S)\delta}\right)}{3n} + \sqrt{\dfrac{2\kappa^2\lambda_{\max}(S^*S)\log\left(\frac{2\kappa^2}{\lambda_{\max}(S^*S)\delta}\right)}{n}}$.

We emphasize that if n is large enough, the above result yields a discretization error decaying like $1/\sqrt{n}$ if we neglect the logarithmic dependence on δ.

Proof. To begin, we observe that $\mathbb{A}S^*S\mathbb{B}$ and $\mathbb{A}S_n^*S_n\mathbb{B}$ are *psd* nuclear operators since $\mathbb{B}$ is nuclear while $\mathbb{A}$ and S^*S are bounded and *psd*. Next, we use the cyclicity property of the trace, and, by using the triangle inequality, we have

$$\begin{aligned}|\operatorname{Tr}(\mathbb{A}S^*S\mathbb{B}S^*S) - \operatorname{Tr}(\mathbb{A}S_n^*S_n\mathbb{B}S_n^*S_n)| \leq{}& |\operatorname{Tr}(\mathbb{A}(S^*S - S_n^*S_n)\mathbb{B}(S^*S - S_n^*S_n))| \\ &+ |\operatorname{Tr}(\mathbb{A}(S^*S - S_n^*S_n)\mathbb{B}S_n^*S_n)| \\ &+ |\operatorname{Tr}(\mathbb{A}S_n^*S_n\mathbb{B}(S^*S - S_n^*S_n))|,\end{aligned}$$

where each of the terms on the RHS exists since $\mathbb{B}$ is nuclear and $\mathbb{A}$, S^*S, and $S_n^*S_n$ are bounded operators. First, we recall that $|\operatorname{Tr}(\mathbb{B}\mathbb{O})| \leq \operatorname{Tr}(|\mathbb{B}\mathbb{O}|) \leq \|\mathbb{O}\|_{op}\operatorname{Tr}(\mathbb{B})$ if $\mathbb{B}$ is nuclear

and *psd*, and $\mathbb{O}$ is bounded. Now, if $\|S_n^* S_n\|_{op} \leq c_0$ and $\|S^*S - S_n^* S_n\|_{op} \leq c$, by using the submultiplicativity of the operator norm and $\|\mathbb{A}\|_{op} \leq 1$, we find

$$|\mathrm{Tr}\,(\mathbb{A}S^*S\mathbb{B}S^*S) - \mathrm{Tr}\,(\mathbb{A}S_n^*S_n\mathbb{B}S_n^*S_n)| \leq (c^2 + 2c_0c)\,\mathrm{Tr}(\mathbb{B}). \tag{4.1}$$

Next we identify c_0 and c. First we remark that $S_n^*S_n = \frac{1}{n}\sum_{i=1}^n \phi(x_i) \otimes \overline{\phi(x_i)}$ and $S^*S = \frac{1}{|X|}\int_X \phi(x) \otimes \overline{\phi(x)}\mathrm{d}x$. Thus, $\|S_n^*S_n\|_{op} \leq \frac{1}{n}\sum_{i=1}^n k(x_i, x_i) \leq \kappa^2 = c_0$ almost surely since $k(x,x) \leq \kappa^2$ for all $x \in X$ by assumption. Next, we use a matrix Bernstein inequality for a sum of random operators to upper bound $\|S^*S - S_n^*S_n\|_{op}$.

Proposition 4.2 (Proposition 12 in [19]). *Let $\mathcal{H}$ be a separable Hilbert space, and let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed self-adjoint positive random operators on $\mathcal{H}$. Assume that $\mathbb{E}X = 0$ and that there exists a real number $\ell > 0$ such that $\lambda_{\max}(X) \leq \ell$ almost surely. Let Σ be a trace class positive operator such that $\mathbb{E}(X^2) \preceq \Sigma$. Then, for any $\delta \in (0,1)$,*

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \leq \frac{2\ell\beta}{3n} + \sqrt{\frac{2\|\Sigma\|_{op}\beta}{n}}, \quad \textit{where } \beta = \log\left(\frac{2\,\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{op}\delta}\right),$$

with probability $1-\delta$. If there further exists an ℓ' such that $\|X\|_{op} \leq \ell'$ almost surely, then, for any $\delta \in (0, 1/2)$,

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_{op} \leq \frac{2\ell'\beta}{3n} + \sqrt{\frac{2\|\Sigma\|_{op}\beta}{n}}, \quad \textit{where } \beta = \log\left(\frac{2\,\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{op}\delta}\right),$$

holds with probability $1-2\delta$.

Thus, we simply define the zero-mean random variable $X_i = \phi(x_i) \otimes \overline{\phi(x_i)} - \frac{1}{|X|}\int_X \phi(x) \otimes \overline{\phi(x)}\mathrm{d}x$ and find that $\|X\|_{op} \leq 2\kappa^2$ holds almost surely by using a triangle inequality. It is easy to verify that

$$\mathbb{E}[X_i^2] \leq \kappa^2\mathbb{E}[\phi(x_i) \otimes \overline{\phi(x_i)}] = \kappa^2\frac{1}{|X|}\int_X \phi(x) \otimes \overline{\phi(x)}\mathrm{d}x = \kappa^2 S^*S = \Sigma.$$

Thus we find $\mathrm{Tr}(\Sigma) = \kappa^2\int_X k(x,x)\mathrm{d}x/|X| \leq \kappa^4$ and $\|\Sigma\|_{op} \leq \kappa^2\lambda_{\max}(S^*S)$. By using these identifications, we have $\beta \leq \log(\frac{2\kappa^2}{\lambda_{\max}(S^*S)\delta})$. Then, we apply Proposition 4.2 and conclude that, with probability at least $1-2\delta$,

$$\|S^*S - S_n^*S_n\|_{op} \leq \frac{4\kappa^2\log\left(\frac{2\kappa^2}{\lambda_{\max}(S^*S)\delta}\right)}{3n} + \sqrt{\frac{2\kappa^2\lambda_{\max}(S^*S)\log\left(\frac{2\kappa^2}{\lambda_{\max}(S^*S)\delta}\right)}{n}} = c,$$

where $c = c_{n,\delta}$. We now use the structural inequality (4.1), and the result follows. ∎

4.2. Constraints discretization. The following result is directly adapted from Theorem 4 in [21] (and its proof).

Proposition 4.3 (see [21]). *Let X be an open ℓ_2-ball of diameter $2R$ in $\mathbb{R}^d$. Let k be the kernel of the RKHS $\mathcal{H}_k = W_2^s(X)$ with $s > d/2$ and $\phi(x)$ be its canonical feature map as defined in section 1.2.2. Let $0 < m < s - d/2$ be an integer. Let $\widehat{X} = \{x_1, \dots, x_n\} \subset X$ such that $h_{\widehat{X},X} \leq R \min(1, \frac{1}{18(m-1)^2})$. Let $d \in \mathcal{C}^m(X)$, and assume that $d(x_i) = \langle \phi(x_i), \mathbb{B}\phi(x_i) \rangle$ for all $i \in [n]$, where $\mathbb{B} \succeq 0$ is a nuclear endomorphism of $\mathcal{H}_k$. Then, it holds that*

$$|d(x) - \langle \phi(x), \mathbb{B}\phi(x) \rangle| \leq \alpha_{m,d} \left(\beta_m \operatorname{Tr}(\mathbb{B}) + |d|_{X,m} \right) h_{\widehat{X},X}^m \text{ for all } x \in X,$$

where $\alpha_{m,d}$ and β_m are constants.

Proposition 4.4 (Lemma 4 in [21]). *Let $X \subset \mathbb{R}^d$ be a bounded set with diameter $2R$, for some $R > 0$, and such that $X = \cup_{x \in S} B_r(x)$, where S is a bounded subset of $\mathbb{R}^d$ for a given $r > 0$. Let $\widehat{X} = \{x_1, \dots, x_n\}$ be a set of independent points sampled from the uniform distribution on X. When $n \geq 2(\frac{6R}{r})^d (\log\left(\frac{2}{\delta}\right) + 2d \log\left(\frac{4R}{r}\right))$, then the following holds with probability at least $1 - \delta$:*

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

4.3. Putting it all together. Sections 4.1 and 4.2 are now used to state and prove our main result about the approximation of the full problem with strong smoothness assumptions.

Theorem 4.5 (approximation of the optimal objective). *Let $\mathcal{H}_k = W_2^s(X)$, where X is an open ℓ_2-ball of radius R in $\mathbb{R}^d$. Let an integer m be such that $0 < m < s - d/2$. Let $\mathbb{B}_{\star\star} \in \mathcal{S}(\mathcal{H}_k)$ be an optimizer of the full problem (kVP). Denote by $\mathbf{B}_\star$ the $n \times n$ matrix obtained by solving (1.9), and consider $V : \mathcal{H}_k \to \mathbb{R}^n$ defined as in section 1.2.2. Let $\delta \in (0, 1/3)$. Consider $C_{n,\delta}$ as given in Proposition 4.1. Then, we have the following:*

(i) *If $\lambda \geq 2C_{n,\delta}$, with probability at least $1 - 2\delta$, $|f(V^* \mathbf{B}_\star V) - f(\mathbb{B}_{\star\star})| \leq 3\lambda \operatorname{Tr}(\mathbb{B}_{\star\star})$.*

(ii) *If n is large enough such that $\frac{11}{n^{1/d}} \left(\log\left(\frac{n}{\delta}\right) + d \log(2)\right)^{\frac{1}{d}} \leq \min(1, \frac{1}{18(m-1)^2})$, and if $\lambda \geq 2C_{n,\delta}$, then, with probability at least $1 - \delta$, it holds that*

$$|d(x) - \langle \phi(x), V^* \mathbf{B}_\star V \phi(x) \rangle| \leq n^{-\frac{m}{d}} \gamma_{m,d,R} \left(\frac{3\lambda}{2C_{n,\delta}} \beta_m \operatorname{Tr}(\mathbb{B}_{\star\star}) + |d|_{X,m} \right) \left(\log\left(\frac{n}{\delta}\right) + d \log(2) \right)^{\frac{m}{d}}$$

for all $x \in X$ and for some constants $\gamma_{m,d,R}$ and $\beta_m > 0$.

Further, with the same assumptions on n and λ as in (i) *and* (ii)*, the two bounds in* (i) *and* (ii) *hold together with probability at least $1 - 3\delta$.*

Thus, in a nutshell, the operator $V^* \mathbf{B}_\star V$ has an objective value which is close to the optimal objective and approximately satisfies the constraints, with high probability, provided that $\lambda > 0$ decays as $1/\sqrt{n}$ up to logarithmic terms.

Proof. The proof technique mainly relies on [21, proof of Theorem 5] and is also used in [10].

Proof of (i). First, the full problem (kVP) has at least one feasible point since $d(x) = \sum_{j=1}^{p} w_j(x)^2$ with p a finite integer, for all $x \in X$ with $w_j \in W_2^s(X)$ for all $1 \leq j \leq d$. Indeed, $\sum_{j=1}^{p} w_j \otimes \overline{w_j}$ is feasible. The objective value in (1.9) for $\mathbf{B}_\star$ is the objective value of $V^*\mathbf{B}_\star V$ for (reg-kVP-n). Second, we know that there exists $\bar{\mathbf{B}} = V\mathbb{B}_{\star\star}V^*$ such that $f_n(V^*\bar{\mathbf{B}}V) = f_n(\mathbb{B}_{\star\star})$ and $\operatorname{Tr}(\bar{\mathbf{B}}) \leq \operatorname{Tr}(\mathbb{B}_{\star\star})$ as a consequence of Lemma 3 in [21]. Hence, by the optimality of $V^*\mathbf{B}_\star V$, we find

$$f_n(V^*\mathbf{B}_\star V) - \lambda \operatorname{Tr}(V^*\mathbf{B}_\star V) \geq f_n(V^*\bar{\mathbf{B}}V) - \lambda \operatorname{Tr}(V^*\bar{\mathbf{B}}V) \geq f_n(\mathbb{B}_{\star\star}) - \lambda \operatorname{Tr}(\mathbb{B}_{\star\star}).$$

Hence, since $VV^* = \mathbb{I}_n$, this gives $f_n(V^*\mathbf{B}_\star V) - f_n(\mathbb{B}_{\star\star}) \geq \lambda \operatorname{Tr}(\mathbf{B}_\star) - \lambda \operatorname{Tr}(\mathbb{B}_{\star\star})$. Now, we use Proposition 4.1 and find that the event $|f(\mathbb{B}_{\star\star}) - f_n(\mathbb{B}_{\star\star})| \leq C_{n,\delta} \operatorname{Tr}(\mathbb{B}_{\star\star})$, with $C_{n,\delta} = (2\kappa^2 c_{n,\delta} + c_{n,\delta}^2)$, occurs with probability $1 - 2\delta$. By combining this with the previous result, we find

$$\begin{aligned} f(\mathbb{B}_{\star\star}) - f_n(V^*\mathbf{B}_\star V) &= f(\mathbb{B}_{\star\star}) - f_n(\mathbb{B}_{\star\star}) + f_n(\mathbb{B}_{\star\star}) - f_n(V^*\mathbf{B}_\star V) \\ &\leq C_{n,\delta} \operatorname{Tr}(\mathbb{B}_{\star\star}) + \lambda \left(\operatorname{Tr}(\mathbb{B}_{\star\star}) - \operatorname{Tr}(\mathbf{B}_\star)\right) \qquad (4.2) \\ &\leq (C_{n,\delta} + \lambda) \operatorname{Tr}(\mathbb{B}_{\star\star}), \qquad (4.3) \end{aligned}$$

where we used $\mathbf{B}_\star \succeq 0$. Similarly, we have

$$\begin{aligned} f_n(V^*\mathbf{B}_\star V) - f(\mathbb{B}_{\star\star}) &= f_n(V^*\mathbf{B}_\star V) - f(V^*\mathbf{B}_\star V) + \underbrace{f(V^*\mathbf{B}_\star V) - f(\mathbb{B}_{\star\star})}_{\leq 0} \\ &\leq C_{n,\delta} \operatorname{Tr}(V^*\mathbf{B}_\star V) = C_{n,\delta} \operatorname{Tr}(\mathbf{B}_\star). \qquad (4.4) \end{aligned}$$

Using (4.2) with the latter inequality, we find $-C_{n,\delta} \operatorname{Tr}(\mathbf{B}_\star) \leq C_{n,\delta} \operatorname{Tr}(\mathbb{B}_{\star\star}) + \lambda \left(\operatorname{Tr}(\mathbb{B}_{\star\star}) - \operatorname{Tr}(\mathbf{B}_\star)\right)$. If $\lambda \geq 2C_{n,\delta}$, this becomes $C_{n,\delta} \operatorname{Tr}(\mathbf{B}_\star) \leq (C_{n,\delta} + \lambda) \operatorname{Tr}(\mathbb{B}_{\star\star})$. Hence, by combining this with (4.4) and recalling (4.3), we obtain, with probability at least $1 - 2\delta$,

$$|f_n(V^*\mathbf{B}_\star V) - f(\mathbb{B}_{\star\star})| \leq (C_{n,\delta} + \lambda) \operatorname{Tr}(\mathbb{B}_{\star\star}) \leq \frac{3}{2}\lambda \operatorname{Tr}(\mathbb{B}_{\star\star}). \qquad (4.5)$$

In light of the proof of Proposition 4.1, $|f_n(V^*\mathbf{B}_\star V) - f(V^*\mathbf{B}_\star V)| \leq C_{n,\delta} \operatorname{Tr}(V^*\mathbf{B}_\star V)$ occurs with probability at least $1 - 2\delta$ due to the same event as in (4.5). Thus, by using a triangle inequality and the same bound on $\operatorname{Tr}(\mathbf{B}_\star)$ as above, we obtain, with probability at least $1 - 2\delta$,

$$|f(V^*\mathbf{B}_\star V) - f(\mathbb{B}_{\star\star})| \leq |f(V^*\mathbf{B}_\star V) - f_n(V^*\mathbf{B}_\star V)| + |f_n(V^*\mathbf{B}_\star V) - f(\mathbb{B}_{\star\star})| \leq 3\lambda \operatorname{Tr}(\mathbb{B}_{\star\star}).$$

Proof of (ii). We simply apply Proposition 4.3 to bound $|d(x) - \langle \phi(x), V^*\mathbf{B}_\star V \phi(x) \rangle|$ in terms of the fill distance $h_{\widehat{X},X}$, and $\operatorname{Tr}(V^*\mathbf{B}_\star V) = \operatorname{Tr}(\mathbf{B}_\star)$, if $h_{\widehat{X},X} \leq R \min(1, \frac{1}{18(m-1)^2})$. The trace $\operatorname{Tr}(\mathbf{B}_\star)$ is upper bounded as above; i.e., if $\lambda \geq 2C_{n,\delta}$, we have

$$C_{n,\delta} \operatorname{Tr}(\mathbf{B}_\star) \leq \frac{C_{n,\delta} + \lambda}{C_{n,\delta}} \operatorname{Tr}(\mathbb{B}_{\star\star}) \leq \frac{3\lambda}{2C_{n,\delta}} \operatorname{Tr}(\mathbb{B}_{\star\star}).$$

As stated in Proposition 4.4, the fill distance is bounded with probability at least $1 - \delta$ by

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

Hence we require $\frac{11R}{n^{1/d}} \left(\log\left(\frac{n}{\delta}\right) + d\log(2)\right)^{\frac{1}{d}} \leq R \min(1, \frac{1}{18(m-1)^2})$. By combining these results, we obtain (ii). ∎

5. Illustrative examples of dimensionality reduction with weak smoothness assumptions. The datasets used in the simulations are Wine[5] ($n = 178$, $d = 13$), the digits 1 and 4 of the training data of MNIST[6] ($n = 12584$, $d = 784$), the HTRU2 dataset[7] ($n = 17898, d = 9$), which is a classification benchmark (1639 pulsars versus 16259 non-pulsars), and a commonly used artificial example with two half-moons[8] ($n = 400$, $d = 2$). Our code is available at https://github.com/mrfanuel/sdp-embedding.

As an illustration of the out-of-sample formula, the digits 1 and 4 of the training set of the MNIST dataset are visualized in Figure 4. On the left, 30 percent of the data is embedded thanks to the SDP embedding, while the out-of-sample formula is used in order to embed the remaining 70 percent of the dataset as displayed on the right of Figure 4. In the case displayed in Figure 1, the embedding dimension is two, since the rank of the optimal solution $\mathbf{B}_\star$ is equal to 2. In the case of the Wine dataset given in Figure 5, we observe again that the result of the SDP embedding still gives an interesting information (bottom left) when the bandwidth is chosen so small that diffusion maps only emphasize the outliers (bottom right).

5.1. Classification example. A larger scale example is given in Figure 6, which displays the embedding of the astrophysics HTRU2 dataset. As in the other examples, this SDP embedding has an annulus shape which can be intuitively interpreted as follows: the points on the same radius have a similar "centrality" in the dataset. In Figure 6, the minority class (pulsars, in red) can be seen as a spike on top of the annulus. A k-nearest neighbors classifier is trained 3 times on the SDP embedding of 70 percent of the dataset (uniform sample) with $k = 5$. The out-of-sample formula (1.5) is used to predict on the test set, i.e., the remaining 30 percent of the dataset. The results are reported in Table 1, where we observe that the diffusion and SDP embeddings yield almost the same precision and recall for this classification task.

5.2. Clustering example. SDP embedding is also applied to a commonly used artificial two-moons dataset displayed in Figure 7. After embedding, the dataset is clustered thanks to k-means. In Figure 8, we display the normalized mutual information (NMI) between the cluster label vector and the ground truth for several values of the Gaussian kernel's bandwidth. Compared with the diffusion embedding,[9] we observe that the clustering on the SDP embedding has a higher performance with a lower variance. In the same figure, a similar procedure is applied on a postprocessed embedding obtained by projecting all embedded points on a unit circle; see, e.g., [17]. As expected, the performance of diffusion embedding then gets closer to the SDP embedding performance, although the latter has still a smaller variance.

6. Related work. We now review a few related papers.

Optimization problem. The type of optimization problem in finite dimension solved in this paper has been studied in various works in different contexts. Several authors examined the connection between an SDP and the max-cut problem, $\mathbb{Z}_2$ or angular synchronization, as well

[5] https://archive.ics.uci.edu/ml/datasets/wine

[6] http://yann.lecun.com/exdb/mnist/

[7] https://archive.ics.uci.edu/ml/datasets/HTRU2

[8] See `twomoons` https://nl.mathworks.com/help/stats/label-data-using-semi-supervised-learning-techniques.html.

[9] The eigenvalue decomposition sometimes fails to converge when σ is small.

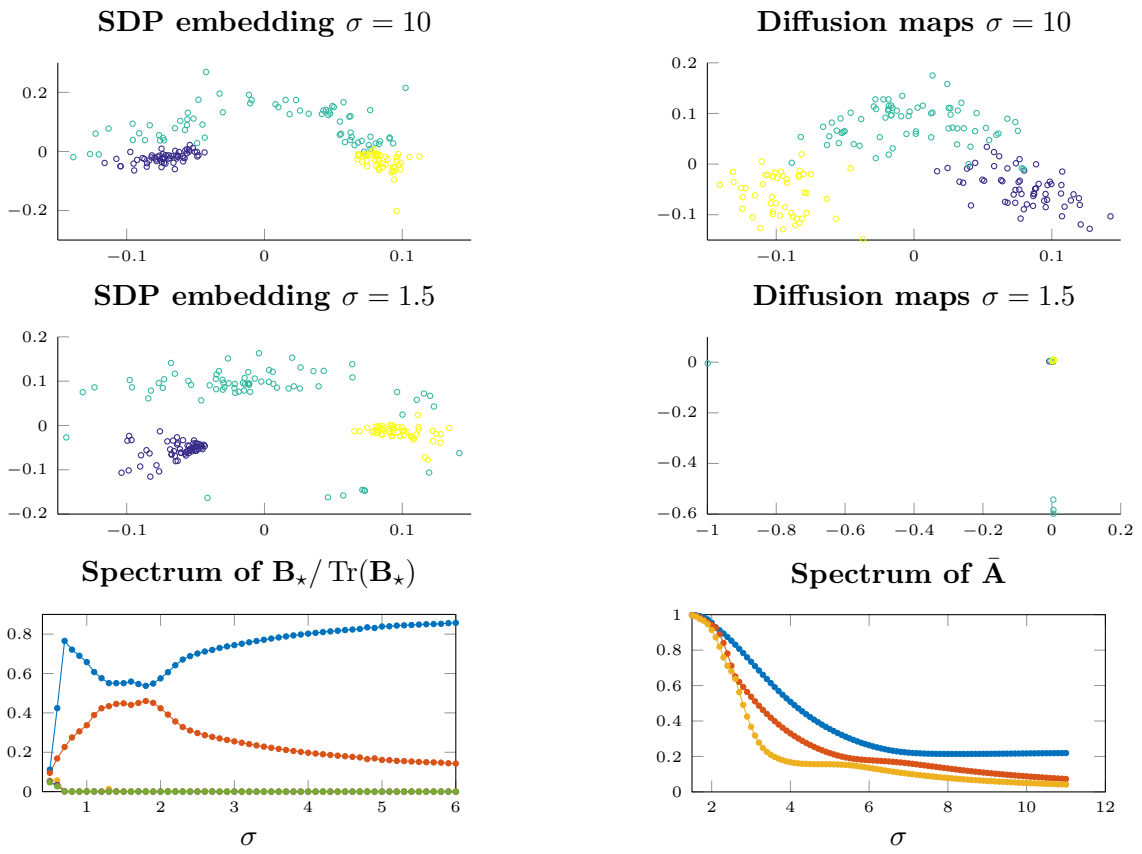

Figure 5. *Comparison of the SDP embedding and the diffusion maps for the Wine dataset after a standardization of the data points. Several different values of the bandwidth σ are used. In the embedding plots, each color refers a different class. At the bottom left, the non-zero eigenvalues of $B_\star / \mathrm{Tr}(B_\star)$ are plotted, whereas, at the bottom right, only the three largest eigenvalues of $\bar{\mathbf{A}}$ (see (1.2)) are displayed. On the RHS, the diffusion embedding is displayed for normalized eigenvectors.*

Table 1

Classification results for the SDP and diffusion embedding (with two components) of the HTRU2 dataset. The positive class here is "pulsars" (minority). The standard deviation over 3 runs is given in parentheses.

SDP embedding				Diffusion embedding			
$\sigma = 10$		$\sigma = 5$		$\sigma = 10$		$\sigma = 5$	
precision	recall	precision	recall	precision	recall	precision	recall
0.90(0.01)	0.76(0.02)	0.91(0.01)	0.79(0.01)	0.90(0.01)	0.78(0.02)	0.91(0.01)	0.79(0.01)

as community detection [5, 3, 11, 6]. Also, in the context of the angular synchronization problem, the paper [2] investigates when an SDP relaxation has a global rank one solution which solves exactly an angular synchronization problem. The conclusions presented here can be viewed as a generalization of previous ideas applied to graph data [5, 3, 2, 11, 1] in the context of kernel methods. Concerning the scalability of (structured) SDPs, [30] recently

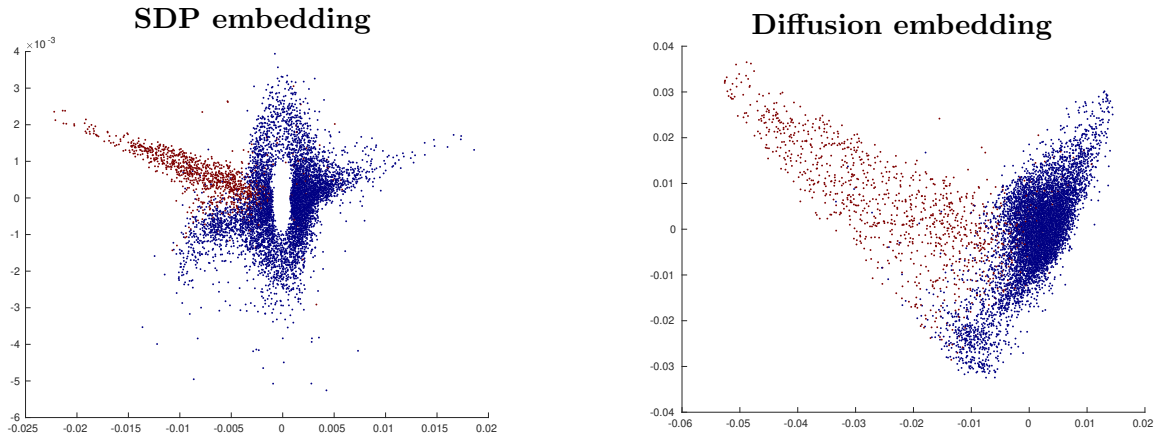

Figure 6. *Embedding of the training set of standardized HTRU2 dataset for $\sigma = 10$. Pulsars are in red, and non-pulsars are in blue. Left: The SDP embedding dimension is the number of numerically significant eigenvalues of $\mathbf{B}_\star / \operatorname{Tr}(\mathbf{B}_\star)$, namely 2 in this case. Right: Diffusion embedding with 2 normalized eigenvectors and $\sigma = 10$.*

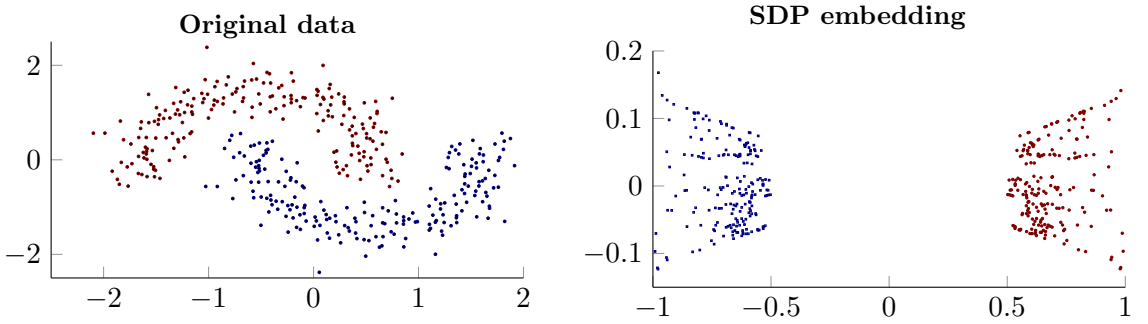

Figure 7. *Illustration of the two-moons dataset with 200 points in each cluster. Raw data (left). SDP embedding of the standardized data with the Gaussian kernel for $\sigma = 0.1$ (right).*

provided an efficient optimization strategy with theoretical guarantees for solving large SDPs with the help of sketching. The latter work might be a source of inspiration for scaling up the approach presented here.

Learning a positive semi-definite matrix. Another dimensionality reduction technique associated to an SDP is the so-called maximum variance unfolding (MVU), also called semidefinite embedding, which was introduced in [27] and was used for computer vision in [26]. Although similar in spirit, the objective function of MVU differs from the objective considered in this paper. Furthermore, an asset of our approach is the out-of-sample formula.

Learning a kernel in an RKHS. Learning the kernel is a topic of interest in the context of supervised learning and has been investigated, e.g., in [15, 14] in the framework of reproducing kernel Hilbert spaces (RKHSs). To the best of our knowledge, the variational problem in infinite dimensions defined in this paper has not yet been studied in the literature. Let us also mention that, in another context, kernels of nuclear operators have been studied, for instance, in [24, 25] in connection with the virtual continuity.

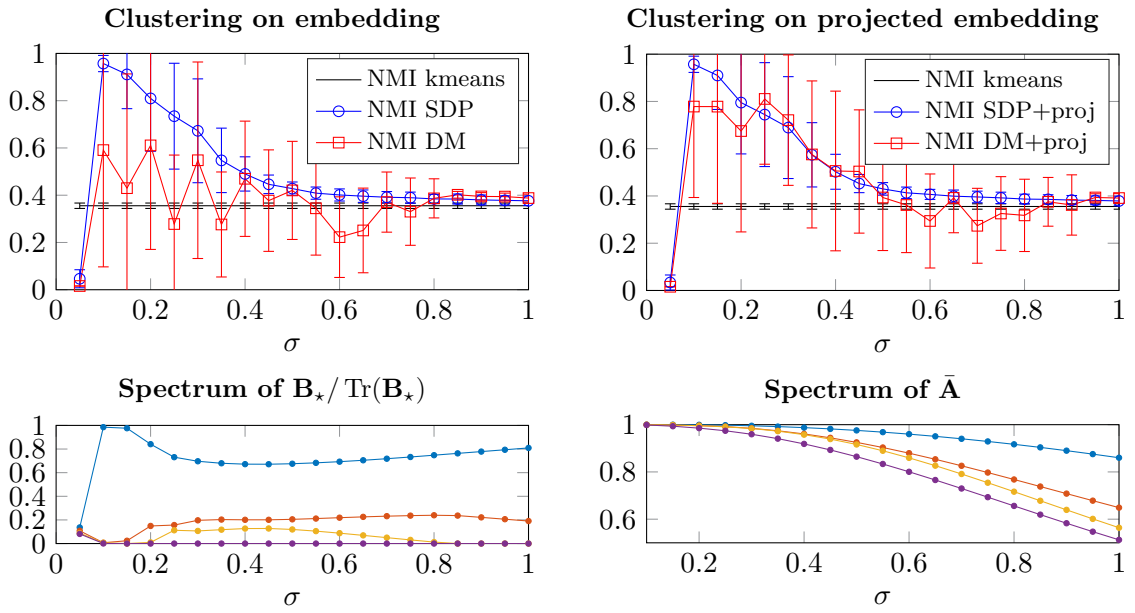

Figure 8. *Clustering performance on the two-moons dataset. Top row: Normalized mutual information (NMI, the larger the better) between the ground truth and the clustering obtained on the embeddings, versus the kernel bandwidth σ. On the LHS, k-means clustering with $k = 2$ is run once on a 2-dimensional SDP embedding (blue circles), a 2-dimensional DM embedding (red squares), and on the raw data (black), as a baseline. On the RHS, a similar procedure is performed except that the embedding is postprocessed by projecting all points on a circle before applying k-means. The markers denote the NMIs averaged over 10 runs, whereas the error bars are standard deviations. Bottom row: We display, as a function of σ, the four leading eigenvalues of $\mathbf{B}_\star/\operatorname{Tr}(\mathbf{B}_\star)$ on the left and the four leading eigenvalues of the matrix $\bar{\mathbf{A}}$, defined in (1.2), on the right. Again, from the observation of the eigenvalues of $\mathbf{B}_\star/\operatorname{Tr}(\mathbf{B}_\star)$, we notice that the effective dimension of the SDP embedding is very low.*

State-of-the-art dimensionality reduction. Two leading dimensionality reduction methods are t-distributed stochastic neighbor embedding (t-SNE) [23] and uniform manifold approximation and projection (UMAP) [13]. These methods are highly successful in practice and scalable. Advantages of our approach are its interpretability and the out-of-sample formula, while a major drawback is that SDP embedding does not achieve as good visualizations in two dimensions as t-SNE or UMAP.

7. Conclusions.

In this work, we discussed a novel dimensionality reduction technique which is based on a semi-definite program. Two approaches were presented: one with weak assumptions about smoothness, and one with strong assumptions about smoothness. An out-of-sample formula is also provided in both cases. It is observed numerically that the embedding is robust to the presence of outliers. Possible future research includes the empirical study of the learning performance of the finite dimensional feature maps defined by the out-of-sample extension formula given by the kernelized problem of section 4. The numerical simulations reported here only considered the Gaussian kernel with weak smoothness assumptions. It would be instructive to analyze the role of the regularization parameter of the kernelized problem as well as the influence of the regularity parameter of the Sobolev kernel.

Appendix A. Useful elements of operator theory.

A.1. Hilbert–Schmidt and nuclear operators. Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a separable Hilbert space with orthonormal basis (ONB) $\{\phi_\ell\}_{\ell\geq 1}$. We say that $A : \mathcal{H} \to \mathcal{H}$ is a compact linear operator if for any bounded sequence $(f_\ell)_\ell \in \mathcal{H}$, the sequence $(Af_\ell)_\ell \in \mathcal{H}$ admits a convergent subsequence. We denote the adjoint of A by A^*. An operator A is *psd* if $\langle v, Av\rangle \geq 0$ for all $v \in \mathcal{H}$. The operator A is Hilbert–Schmidt if $\sum_{\ell\geq 1}\langle A\phi_\ell, A\phi_\ell\rangle < \infty$. The Hilbert–Schmidt operators form a Hilbert space for the inner product $\langle A, B\rangle_{HS} = \sum_{\ell\geq 1}\langle A\phi_\ell, B\phi_\ell\rangle$. A Hilbert–Schmidt operator A is nuclear (or trace class) if $\sum_{\ell\geq 1}\langle\sqrt{A^*A}\phi_\ell, \phi\ell\rangle < \infty$, while the trace of this operator reads $\mathrm{Tr}(A) = \sum_{\ell\geq 1}\langle A\phi_\ell, \phi_\ell\rangle$. Finally, the trace class or nuclear norm of an operator is $\|A\|_* = \mathrm{Tr}(\sqrt{A^*A})$. Notice that these definitions are in fact independent of the choice of ONB for $\mathcal{H}$.

A.2. Reproducing kernel Hilbert space. Let X be a set. A kernel is a symmetric function $k : X \times X \to \mathbb{R}$. A kernel is positive semi-definite (*psd*) if for all finite samples $\{x_1, \dots, x_m\}$ of points in X, the corresponding Gram matrix $[\mathbf{K}]_{ij} = k(x_i, x_j)$, which is *psd*. Let $\phi(x) : X \to \mathbb{R}$ be the function $k(x,\cdot)$. The RKHS $\mathcal{H}_k$ is given by the completion of span$\{\phi(x)$ s.t. $x \in X\}$ endowed with the inner product $\langle\phi(x), \phi(y)\rangle = k(x, y)$.

A.3. Generalized Mercer theorem.

Theorem A.1 (Mercer's theorem for *psd* symmetric nuclear operators, Theorem 3.10 in [22]). *Let X be a measurable space, μ be a measure on X, and $K : L^2(X,\mu) \to L^2(X,\mu)$ be a psd, symmetric, and nuclear operator. We write $\{\lambda^{(\ell)}\}_{\ell\geq 1}$ for the at most countably many non-zero eigenvalues of K, where we included geometric multiplicities. Then there exists a measurable kernel k on X with separable RKHS $\mathcal{H}_k$ such that $(Kf)(\cdot) = \int_X k(\cdot, y)f(y)\mathrm{d}\mu(y)$, with $f \in L^2(X,\mu)$ and where the left-hand side (LHS) of the equation is considered to be a μ-equivalence class in $L^2(X,\mu)$. In addition, k enjoys a Mercer representation; that is, there exists an ONB $\{\sqrt{\lambda^{(\ell)}}\phi^{(\ell)}\}_{\ell\geq 1}$ of $\mathcal{H}_k$ such that $\{\phi^{(\ell)}\}_{\ell\geq 1}$ is an orthonormal system in $L^2(X,\mu)$ consisting of the eigenfunctions corresponding to the eigenvalues $\{\lambda^{(\ell)}\}_{\ell\geq 1}$ of K such that*

$$k(x, y) = \sum_{\ell\geq 1} \lambda^{(\ell)}\phi^{(\ell)}(x)\phi^{(\ell)}(y) \text{ for all } x, y \in X.$$

Moreover, the integral operator K is defined pointwise, that is, $Kf(x) = \int_X k(x, y)f(y)\mathrm{d}\mu(y)$, for all $f \in L^2(X,\mu)$ and all $x \in X$.

Appendix B. Dual certificate.

By using the duality theory of SDPs, we can show that the optimality of a matrix $\mathbf{B}$ can be certified thanks to a so-called dual certificate. Indeed, we can write the Lagrangian associated to (SDP), $\mathcal{L}(\mathbf{B}, \boldsymbol{y}) = \mathrm{Tr}\left(\bar{\mathbf{A}}\mathbf{B}\right) + \boldsymbol{y}^\top(\mathrm{diag}(\mathbf{B}) - \boldsymbol{d})$. The primal optimization problem is $p_\star = \max_{\mathbf{B}\succeq 0}\min_{\boldsymbol{y}\in\mathbb{R}^n}\mathcal{L}(\mathbf{B}, \boldsymbol{y})$. The dual problem, which is classically given by $d_\star = \min_{\boldsymbol{y}\in\mathbb{R}^n}\max_{\mathbf{B}\succeq 0}\mathcal{L}(\mathbf{B}, \boldsymbol{y})$, can be simplified as

$$d_\star = \min_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{d}^\top\boldsymbol{y}, \text{ subject to } \mathrm{Diag}(\boldsymbol{y}) - \bar{\mathbf{A}} \succeq 0.$$

Since the matrix $\mathbf{B} = \mathrm{Diag}(\boldsymbol{d}) \succ 0$ is strictly feasible for (SDP) (Slater condition), the optimal values of the dual and primal problems coincide, namely $d_\star = p_\star$. The duality gap can be

written as follows:

$$d_\star - p_\star = \boldsymbol{d}^\top \boldsymbol{y}_\star - \operatorname{Tr}\left(\bar{\mathbf{A}}\mathbf{B}_\star\right) = \operatorname{Tr}\Big((\operatorname{Diag}(\boldsymbol{y}_\star) - \bar{\mathbf{A}})\mathbf{B}_\star\Big).$$

Since by feasibility $\mathbf{B}_\star \succeq 0$ and $\operatorname{Diag}(\boldsymbol{y}_\star) - \bar{\mathbf{A}} \succeq 0$, a vanishing duality gap yields $(\operatorname{Diag}(\boldsymbol{y}_\star) - \bar{\mathbf{A}})\mathbf{B}_\star = 0$. In particular, the diagonal of this matrix has to vanish: $\operatorname{Diag}(\boldsymbol{y}_\star)\operatorname{Diag}(\boldsymbol{d}) - \operatorname{ddiag}(\bar{\mathbf{A}}\mathbf{B}_\star) = 0$. As is explained in [2], where an analogous calculation is done, the only solution is $\boldsymbol{y}_\star = \operatorname{Diag}(\boldsymbol{d})^{-1}\operatorname{Diag}(\bar{\mathbf{A}}\mathbf{B})$. This yields a dual certificate given by

$$\mathbf{L}(\mathbf{B}) = \operatorname{Diag}(\boldsymbol{d})^{-1}\operatorname{ddiag}(\bar{\mathbf{A}}\mathbf{B}) - \bar{\mathbf{A}}, \tag{B.1}$$

which certifies the optimality of $\boldsymbol{B}$, and the following conditions hold: $\mathbf{L}(\mathbf{B})\mathbf{B} = 0$ (complementary slackness) and $\mathbf{L}(\mathbf{B}) \succeq 0$ (dual feasibility).

Algorithm 1. Projected power method [3].

1: Input: Symmetric *psd* $\mathbf{J} \in \mathbb{R}^{n\times n}$; and $\mathbf{H_0} \in \mathbb{R}^{n\times r_0}$ such that $\mathcal{P}(\mathbf{H_0}) = \mathbf{H_0}$.
2: **for** $n = 1, 2, \ldots$ **do**
3: $\quad \mathbf{H_n} = \mathcal{P}(\mathbf{J}\mathbf{H_{n-1}})$.
4: **end for**

Appendix C. Numerical method. We address here the numerical solution of the optimization problem (SDP) thanks to a low rank factorization [5]. Although interior point algorithms can solve SDPs with good theoretical guarantees, we propose here another method with the advantage that it can empirically solve problems of larger sizes. We first use the change of variables $\bar{\mathbf{A}}_d = \operatorname{Diag}(\boldsymbol{d})^{1/2}\bar{\mathbf{A}}\operatorname{Diag}(\boldsymbol{d})^{1/2}$ and $\mathbf{B}_d = \operatorname{Diag}(\boldsymbol{d})^{-1/2}\mathbf{B}\operatorname{Diag}(\boldsymbol{d})^{-1/2}$. Then, we factorize $\mathbf{B}_d = \mathbf{H}\mathbf{H}^\top$ with $\mathbf{H} \in \mathbb{R}^{n\times r_0}$ and propose solving instead

$$\max_{\mathbf{H}\in\mathbb{R}^{n\times r_0}} \operatorname{Tr}\left(\mathbf{H}^\top\bar{\mathbf{A}}_d\mathbf{H}\right) \text{ subject to } \|\mathbf{H}_{i*}\|_2 = 1, \text{ for all } i \in [n], \tag{C.1}$$

where $\mathbf{H}_{i*}$ denotes the ith row of $\mathbf{H} \in \mathbb{R}^{n\times r_0}$. Inspired by the projected gradient method, a natural projection operator on the feasible $\mathcal{M} = (\mathbb{S}^{r_0-1})^n$ is simply obtained by projecting each factor of the Cartesian product on the unit sphere $\mathbb{S}^{r_0-1}$; i.e., let $i \in [n]$; then $[\mathcal{P}(\mathbf{H})]_{i*} = \mathbf{H}_{i*}/\|\mathbf{H}_{i*}\|_2$ normalizes the rows of the matrix $\mathbf{H} \in \mathbb{R}^{n\times r_0}$. If one row of the matrix $\mathbf{H}$ is a row of zeros, $\mathcal{P}$ returns a random row vector. Hence, the method for maximizing $\operatorname{Tr}\left(\mathbf{H}^\top\bar{\mathbf{A}}_d\mathbf{H}\right)$ summarized in Algorithm 1 consists of a succession of matrix multiplications by $\mathbf{J} = \bar{\mathbf{A}}_d$ and projection steps $\mathcal{P}$. The sequence of iterates of Algorithm 1 has increasing objective values, as explained in [3]. Once a solution $\mathbf{H}_\star$ is found by using Algorithm 1, the embedding coordinates are found by computing the singular value decomposition of $\mathbf{H_d}$. To summarize, the numerical algorithm used to solve a rank constrained version of (SDP) is the following: The initial point for Algorithm 1 is obtained as $\mathbf{H_0} = \mathcal{P}(\mathbf{M_0})$, where $\mathbf{M_0} \in \mathbb{R}^{n\times r_0}$ is generated with independent entries in $[-1, 1]$ chosen uniformly at random. Algorithm 1 yields $\mathbf{H}_\star \in \mathbb{R}^{n\times r_0}$ after convergence, and the optimality of the candidate solution $\mathbf{B}_\star = \mathbf{H_\Xi}\mathbf{H_\Xi}^\top$ with $\mathbf{H_\Xi} = \operatorname{Diag}(\boldsymbol{d})^{1/2}\mathbf{H}_\star$ can be certified by the dual certificate (B.1). Finally, a singular value decomposition of $\mathbf{H_\Xi}$ is performed in order to obtain the embedding coordinates.

Acknowledgments. M.F. acknowledges stimulating discussions with A. Themelis and thanks the reviewers and associate editor for their suggestions.

REFERENCES

[1] A. Aspeel, *Community Detection in Large-Scale Time-Varying Networks, A Modularity Based Approach*, Master thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium, 2017.
[2] A. S. Bandeira, N. Boumal, and A. Singer, *Tightness of the maximum likelihood semidefinite relaxation for angular synchronization*, Math. Program., 163 (2017), pp. 145–167.
[3] N. Boumal, *Nonconvex phase synchronization*, SIAM J. Optim., 26 (2016), pp. 2355–2377, https://doi.org/10.1137/16M105808X.
[4] C. Brislawn, *Kernels of trace class operators*, Proc. Amer. Math. Soc., 104 (1988), pp. 1181–1190.
[5] S. Burer and R. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program., 95 (2003), pp. 329–357.
[6] S. Chrétien, M. Cucuringu, G. Lecué, and L. Neirac, *Learning with semi-definite programming: Statistical bounds based on fixed point analysis and excess risk curvature*, J. Mach. Learn. Res., 22 (2021), pp. 1–64, https://jmlr.org/papers/v22/21-0021.html.
[7] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.
[8] R. R. Coifman and S. Lafon, *Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 31–52.
[9] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, F. Warner, and S. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 7426–7431.
[10] M. Fanuel and R. Bardenet, *Nonparametric estimation of continuous DPPs with kernel methods*, in Advances in Neural Information Processing Systems 34 (virtual event), NeurIPS, 2021, https://proceedings.neurips.cc/paper/2021/hash/ca8a2d76a5bcc212226417361a5f0740-Abstract.html.
[11] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, *Phase transitions in semidefinite relaxations*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. E2218–E2223.
[12] U. Marteau-Ferey, F. R. Bach, and A. Rudi, *Non-parametric models for non-negative functions*, in Advances in Neural Information Processing Systems 33 (Vancouver, Canada), NeurIPS, San Diego, CA, 2020, https://proceedings.neurips.cc/paper/2020/file/968b15768f3d19770471e9436d97913c-Paper.pdf.
[13] L. McInnes J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, preprint, https://arxiv.org/abs/1802.03426, 2018.
[14] C. Micchelli, M. Pontil, Q. Wu, and D.-X. Zhou, *Error bounds for learning the kernel*, Anal. Appl. (Singap.), 14 (2016), pp. 849–868.
[15] C. A. Micchelli and M. Pontil, *Learning the kernel function via regularization*, J. Mach. Learn. Res., 6 (2005), pp. 1099–1125.
[16] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, *Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators*, in Proceedings of the 18th International Conference on Neural Information Processing Systems, NeurIPS, San Diego, CA, 2005, pp. 955–962.
[17] T. Qin and K. Rohe, *Regularized spectral clustering under the degree-corrected stochastic blockmodel*, in Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS 2013, Volume 2, Curran Associates, Red Hook, NY, 2013, pp. 3120–3128.
[18] L. Rosasco, M. Belkin, and E. D. Vito, *On learning with integral operators*, J. Mach. Learn. Res., 11 (2010), pp. 905–934, https://jmlr.org/papers/v11/rosasco10a.html.
[19] A. Rudi, R. Camoriano, and L. Rosasco, *Less is more: Nyström computational regularization*, in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2015, pp. 1657–1665, http://papers.nips.cc/paper/5936-less-is-more-nystrom-computational-regularization.pdf.
[20] A. Rudi, E. De Vito, A. Verri, and F. Odone, *Regularized kernel algorithms for support estimation*, Front. Appl. Math. Statist., 3 (2017), 23, https://doi.org/10.3389/fams.2017.00023.
[21] A. Rudi, U. Marteau-Ferey, and F. Bach, *Finding Global Minima via Kernel Approximations*, pre-

print, https://arxiv.org/abs/2012.11978, 2020.

[22] I. Steinwart and C. Scovel, *Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs*, Constr. Approx., 35 (2012), pp. 363–417.

[23] L. van der Maaten and G. Hinton, *Visualizing data using t-SNE*, J. Mach. Learn. Res., 9 (2008), pp. 2579–2605, https://www.jmlr.org/papers/v9/vandermaaten08a.html.

[24] A. M. Vershik, P. B. Zatitskiy, and F. V. Petrov, *Virtual continuity of measurable functions of several variables and embedding theorems*, Funct. Anal. Appl., 47 (2013), pp. 165–173.

[25] A. M. Vershik, P. B. Zatitskiy, and F. V. Petrov, *Integration of virtually continuous functions over bistochastic measures and the trace formula for nuclear operators*, St. Petersburg Math. J., 27 (2016), pp. 393–398.

[26] K. Weinberger and L. Saul, *Unsupervised learning of image manifolds by semidefinite programming*, in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, Vol. 2, IEEE, Washington, DC, 2004, p. II, https://doi.org/10.1109/CVPR.2004.1315272.

[27] K. Q. Weinberger, F. Sha, and L. K. Saul, *Learning a kernel matrix for nonlinear dimensionality reduction*, in Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04, ACM, New York, 2004, p. 106, https://doi.org/10.1145/1015330.1015345.

[28] H. Wendland, *Scattered Data Approximation*, Cambridge Monogr. Appl. Comput. Math. 17, Cambridge University Press, Cambridge, UK, 2004, https://doi.org/10.1017/CBO9780511617539.

[29] C. K. I. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*, in Advances in Neural Information Processing Systems 13, NeurIPS, San Diego, CA, 2001, pp. 682–688.

[30] A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, and V. Cevher, *Scalable semidefinite programming*, SIAM J. Math. Data Sci., 3 (2021), pp. 171–200, https://doi.org/10.1137/19M1305045.