Optimal quadrature-sparsification for integral operator approximation

Bertrand Gauthier, Johan a K Suykens

OPTIMAL QUADRATURE-SPARSIFICATION FOR INTEGRAL OPERATOR APPROXIMATION

BERTRAND GAUTHIER*† AND JOHAN A.K. SUYKENS‡†

Abstract. We address the problem of designing sparse quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels. For a given kernel, we introduce the notion of squared-kernel discrepancy between two measures, consisting in the squared Hilbert-Schmidt norm of the difference between the integral operators defined from the underlying kernel and the two measures, the integral operators being viewed as operators on the underlying reproducing kernel Hilbert space. In the framework of integral operators defined from measures supported by a fixed set of points, sparsity of the approximate quadrature can be enforced through an ℓ^1-type penalisation, and the computation of a penalised squared-kernel-discrepancy-optimal Nyström approximation is thus turned into a convex quadratic minimisation problem. The penalisation can be introduced under the form of a regularisation term or of a constraint, both formulations being equivalent; by analogy with spectral truncation, we in particular propose to penalise the trace of the approximate operator. The quadratic programs related to the regularised and constrained squared-kernel-discrepancy minimisation problems can be interpreted as the Lagrange dual formulations of distorted one-class support-vector machines related to the squared-kernel and the initial discrete measure. Numerical strategies for solving large-scale squared-kernel discrepancy minimisation problems are investigated and the efficiency of the approach is illustrated on a series of examples. We in particular demonstrate the ability of the proposed methodology to lead to accurate approximations of the main eigenpairs of kernel-matrices related to large-scale datasets.

Key words. sparse Nyström approximation, integral operator, RKHS, optimal quadrature-sparsification, squared-kernel discrepancy, L1-type penalisation, convex quadratic programming, one-class SVM.

AMS subject classifications. 47G10, 41A55, 46E22

1. Introduction.

The eigendecomposition of an integral operator defined from a symmetric and positive-semidefinite kernel and a discrete measure supported by N points is numerically equivalent to the diagonalisation of a $N \times N$ symmetric and positive-semidefinite matrix (notice that such a framework is encountered each time a pointwise quadrature is used to approximate an integral operator, or more simply when one aims at performing the eigendecomposition of a kernel-matrix). In the non-sparse case and for the direct "naive" approach, the amount of computations required to perform the diagonalisation of such a matrix scales as $\mathcal{O}(N^3)$, and therefore becomes quickly intractable when N increases; in addition, the memory required for the storage of the underlying matrix scales as $\mathcal{O}(N^2)$. By analogy with the classical framework of Nyström approximation (see, e.g., [8]), a common alternative consists in carrying out the eigendecomposition of a (weighted) principal submatrix of the initial matrix, with size $n \ll N$, and then expanding the result back up to dimension N; in the integral-operator framework, this operation is related to the definition of a sparse approximation of the initial discrete measure (the support of the approximate measure being included in the support of the initial measure).

Naturally, the selection of the considered sparse measure has a strong impact on the quality of the induced approximation, raising questions relative to the selection of an "appropriate" measure. In our framework, since the initial discrete measure may already be interpreted as a pointwise quadrature, we refer to this problem as *optimal quadrature-sparsification*; we in particular consider the issue of computing a sparse measure leading to an accurate approximation of the main eigendirections of the initial integral operator. As a main feature, our approach is deterministic, relies on convex quadratic optimisation, and leads to approximations enjoying an optimality property. By dealing with integral operators and kernel-matrices, we can potentially handle large-scale problems (i.e., large values of N) without suffering storage issues: we only require to have a quick access to any entry of the involved matrices, the required entries being computed on the fly.

To be more precise, we consider a general space $\mathscr{X}$ and a symmetric and positive-semidefinite kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$; the underlying reproducing kernel Hilbert space (RKHS, see for instance [1]) of real-valued functions on $\mathscr{X}$ is denoted by $\mathcal{H}$. For a set $S = \{x_k\}_{k=1}^N$ of N points in $\mathscr{X}$, we denote by $\mathbf{K}$ the underlying $N \times N$ kernel-matrix, with i, j-th entry $\mathbf{K}_{i,j} = K(x_i, x_j)$. Considering a

*bertrand.gauthier@esat.kuleuven.be (corresponding author)
‡johan.suykens@esat.kuleuven.be
†KU Leuven, ESAT-STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

measure $\mu = \sum_{k=1}^{N} \omega_k \delta_{x_k}$ supported by $\mathcal{S}$ (with $\omega_k > 0$ and where δ_{x_k} stands the for Dirac measure centered at x_k), we can naturally define the integral operator (see Section 2 for a detailed discussion)

$$T_\mu[f](x) = \int_{\mathscr{X}} K(x,t) f(t) \mathrm{d}\mu(t).$$

The operator T_μ is canonically related to the matrix $\mathbf{KW}$ where $\mathbf{W} = \mathrm{diag}(\boldsymbol{\omega})$ is the $N \times N$ diagonal matrix with diagonal entries $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_N)^T$; the eigendecomposition of T_μ can be easily obtained from the diagonalisation of the symmetric positive-semidefinite matrix $\mathbf{W}^{1/2}\mathbf{K}\mathbf{W}^{1/2}$ (notice that the diagonalisation of $\mathbf{K}$ corresponds to the case $\boldsymbol{\omega} = \mathbb{1}$, with $\mathbb{1} = (1, \cdots, 1)^T \in \mathbb{R}^N$). Our aim is to construct a sparse measure $\nu = \sum_{k=1}^{N} \upsilon_k \delta_{x_k}$, with $\upsilon_k \geqslant 0$ (ν is therefore also supported by $\mathcal{S}$) in order to use the eigendecomposition of T_ν to build an accurate approximation of the main eigenpairs of T_μ. By "sparse measure ν", we mean that we expect most of the components of the vector $\boldsymbol{\upsilon} = (\upsilon_1, \cdots, \upsilon_N)^T$ to be null, making the eigendecomposition of T_ν easily computable; the measure ν may therefore be interpreted as a sparse approximation of the initial measure μ.

In order to construct the measure ν, we propose to consider a constrained or regularised minimisation of the *squared-kernel discrepancy between μ and ν*, denoted by $D_{K^2}(\mu, \nu)$ and given by

$$D_{K^2}(\mu, \nu) = \|T_\mu - T_\nu\|^2_{\mathrm{HS},\mathcal{H}},$$

i.e., $D_{K^2}(\mu, \nu)$ is the squared Hilbert-Schmidt norm of the difference between the operators T_μ and T_ν interpreted as integral operators on the RKHS $\mathcal{H}$, see Section 2.3 (an similar framework is for instance considered in [18]). The connection between the squared-kernel discrepancy and the Nyström approximation problem is further highlighted by Theorem 2.1.

When μ and ν are discrete measures supported by $\mathcal{S}$ (related to the vectors $\boldsymbol{\omega}$ and $\boldsymbol{\upsilon} \in \mathbb{R}^N$, respectively), we have $D_{K^2}(\mu, \nu) = (\boldsymbol{\omega} - \boldsymbol{\upsilon})^T \mathbf{S} (\boldsymbol{\omega} - \boldsymbol{\upsilon})$, where $\mathbf{S} = \mathbf{K} * \mathbf{K}$ (Hadamard product) is the kernel-matrix defined by the (symmetric and positive-semidefinite) squared-kernel $K^2(\cdot,\cdot) = \big(K(\cdot,\cdot)\big)^2$ and $\mathcal{S}$. Remarkably, the squared-kernel discrepancy appears as a natural extension of the classical Frobenius-norm-based criterion for matrix low-rank approximation (see. e.g., [4, 11, 20, 2, 6]), as discussed in Section 3.1.

For a fixed measure μ, the minimum of the application $\nu \mapsto D_{K^2}(\mu, \nu)$ is trivially reached for $\nu = \mu$; an ℓ_1-type penalisation is therefore introduced in order to discard this trivial solution while enforcing the sparsity of the measure ν, see Section 3. The penalisation takes the form of a regularisation term or of an additional constraint, both formulations being equivalent. In this way, the sampling problem for Nyström approximation is turned into a quadratic program (QP), and the sparsity of the resulting optimal measure ν^* is controlled by the penalisation parameter (namely, α for the regularised formulation, or $\varkappa$ for the constrained formulation). By analogy with the framework of spectral truncation, we more particularly focus on the penalisation of the trace of the operator T_ν, while for instance enforcing $\mathrm{trace}(T_\nu)$ to be a fraction of $\mathrm{trace}(T_\mu)$.

Since the matrix $\mathbf{S}$ is a kernel-matrix, the QPs related to the regularised or constrained discrepancy minimisation can naturally be interpreted as the Lagrange duals of "distorted" one-class support vector machines (SVMs) related to the squared-kernel $K^2(\cdot,\cdot)$ and the initial discrete measured μ, as discussed in Section 5. This analogy with one-class SVMs also suggests the introduction of soft-margin-type extensions of the squared-kernel-discrepancy minimisations problems.

We present a careful analysis of the approach and propose two different numerical strategies for the computations of optimal (or nearly optimal) sparse measures ν^*. The first strategy, described in Section 6, is "direct" and consists in computing the regularisation path (see for instance [14]) related to the regularised or constrained formulations. The second strategy, see Section 7, applies more specifically to the constrained formulation and consists in using a vertex-exchange algorithm (see, e.g., [15, Chap. 9]) in order to solve the underlying QP. Importantly, in order to be able to tackle large-scale problems, our numerical strategies are "kernelised", in the sense that they do not require the storage of any large objects (like in particular the matrix $\mathbf{K}$ or $\mathbf{S}$). In order to enhance the sparsity of a given discrete measure ν while trying to keep the discrepancy as low as possible, in Section 7.2, we propose two greedy heuristics based on iterative pairwise-component merging.

Some experiments are carried out in Sections 8 and 9, illustrating the ability of the approach to lead to accurate sparse approximations of the main eigenpairs of any integral operator defined from a discrete measure with finite support.

The paper is organised as follows. Section 2 is devoted to some theoretical recalls about integral operators and to the introduction of the notion of squared-kernel discrepancy. In Section 3, we only consider discrete measures with support included in a fixed finite set S; in this framework, we introduce the QPs related to the regularised and constrained squared-kernel-discrepancy minimisation problems and discuss their properties. The approximation of the main eigenpairs of the initial operator T_μ from a low-discrepancy measure ν is investigated in Section 4. Section 5 is devoted to the study the one-class SVMs related to the regularised and constrained problems. Some numerical strategies for the resolution of the regularised and constrained problems are discussed in Sections 6 and 7. Sections 8 (two-dimentionnal toy example) and 9 (large-scale problems) are devoted to numerical experiments, and Section 10 concludes.

We have tried to make the paper as self-contained as possible; for the sake of readability, all the proofs are placed in Appendix A.

2. Notations, recalls and theoretical motivations. This section is devoted to recalls related to integral operators defined from a symmetric positive-semidefinite kernel; we then introduce the notion of squared-kernel discrepancy and highlight its relevance in the framework of Nyström approximation.

2.1. Integral operators. We assume $\mathscr{X}$ is a measurable space and we denote by $\mathcal{A}$ the underlying σ-algebra. We suppose that the (real-valued) kernel $K(\cdot,\cdot)$ is measurable on $\mathscr{X}\times\mathscr{X}$ for the product σ-algebra $\mathcal{A}\otimes\mathcal{A}$ (see for instance [19, Chap. 4]), so that, in particular, the RKHS $\mathcal{H}$ consists of measurable functions on $\mathscr{X}$. We also assume that the diagonal of $K(\cdot,\cdot)$, i.e., the function $x\mapsto K(x,x)$ is measurable on $(\mathscr{X},\mathcal{A})$. We denote by $\mathcal{M}$ the set of all measures on $(\mathscr{X},\mathcal{A})$ and we introduce

$$\mathscr{T}(K)=\left\{\mu\in\mathcal{M}\middle|\int_{\mathscr{X}}K(x,x)\mathrm{d}\mu(x)=\tau_\mu<+\infty\right\}.$$

For $\mu\in\mathscr{T}(K)$, we have $K(\cdot,\cdot)\in L^2(\mu\otimes\mu)$ since in particular

$$\|K\|^2_{L^2(\mu\otimes\mu)}=\int_{\mathscr{X}\times\mathscr{X}}\big(K(x,t)\big)^2\mathrm{d}\mu(x)\mathrm{d}\mu(t)\leqslant\tau_\mu^2.$$

In addition, for all $h\in\mathcal{H}$, we have $h\in L^2(\mu)$ and $\|h\|^2_{L^2(\mu)}\leqslant\tau_\mu\|h\|^2_{\mathcal{H}}$, i.e., $\mathcal{H}$ is continuously included in $L^2(\mu)$. We then define the symmetric and positive-semidefinite integral operator T_μ on $L^2(\mu)$, given by, for $f\in L^2(\mu)$ and $x\in\mathscr{X}$,

$$T_\mu[f](x)=\int_{\mathscr{X}}K(x,t)f(t)\mathrm{d}\mu(t),$$

so that $T_\mu[f]\in\mathcal{H}$ and, for $h\in\mathcal{H}$, $(h|T_\mu[f])_{\mathcal{H}}=(h|f)_{L^2(\mu)}$, see for instance [5] for more details.

We introduce the closed linear subspaces $\mathcal{H}_{0\mu}=\left\{h\in\mathcal{H}\middle|\|h\|_{L^2(\mu)}=0\right\}$ and $\mathcal{H}_\mu=\mathcal{H}_{0\mu}^{\perp_{\mathcal{H}}}$ (i.e., $\mathcal{H}_\mu$ is the orthogonal of $\mathcal{H}_{0\mu}$ in $\mathcal{H}$), leading to the orthogonal decomposition $\mathcal{H}=\mathcal{H}_\mu\oplus\mathcal{H}_{0\mu}$.

We denote by $\{\lambda_k|k\in\mathbb{I}_\mu^+\}$ the at most countable set of all strictly positive eigenvalues of T_μ (repeated according to their algebraic multiplicity), and let $\widetilde{\varphi}_k\in L^2(\mu)$ be the associated eigenfunctions, orthonormalised for $L^2(\mu)$. For $k\in\mathbb{I}_\mu^+$, let $\varphi_k=\frac{1}{\lambda_k}T_\mu[\widetilde{\varphi}_k]\in\mathcal{H}$ be the canonical extension of $\widetilde{\varphi}_k$, so that $\{\sqrt{\lambda_k}\varphi_k|k\in\mathbb{I}_\mu^+\}$ is an orthonormal basis (o.n.b.) of the subspace $\mathcal{H}_\mu$ of $\mathcal{H}$. The reproducing kernel of $\mathcal{H}_\mu$ (for the Hilbert structure of $\mathcal{H}$) is $K_\mu(x,t)=\sum_{k\in\mathbb{I}_\mu^+}\lambda_k\varphi_k(x)\varphi_k(t)$ and $\tau_\mu=\sum_{k\in\mathbb{I}_\mu^+}\lambda_k\geqslant 0$ (in fact, $\tau_\mu>0$ as far as $\mathbb{I}_\mu^+$ is not empty).

2.2. Discrete measures and kernel-matrices . Let $\mu=\sum_{k=1}^N\omega_k\delta_{x_k}$ be a discrete measure supported by S, with $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_N)^T\in\mathbb{R}^N$, $\omega_k>0$ (in what follows, we shall use the notation $\boldsymbol{\omega}>0$), where δ_{x_k} is the Dirac measure (evaluation functional) at x_k. We have $\mu\in\mathscr{T}(K)$, and for $f\in L^2(\mu)$ and $x\in\mathscr{X}$, using matrix notation,

$$T_\mu[f](x)=\sum_{k=1}^N\omega_kK(x,x_k)f(x_k)=\mathbf{k}^T(x)\mathbf{W}\mathbf{f},$$

with $\mathbf{W} = \mathrm{diag}(\boldsymbol{\omega})$, $\mathbf{k}(x) = \big(K(x_1, x), \cdots, K(x_N, x)\big)^T$ and $\mathbf{f} = \big(f(x_1), \cdots, f(x_N)\big)^T \in \mathbb{R}^N$. We can identify the Hilbert space $L^2(\mu)$ with the space $\mathbb{R}^N$ endowed with the inner product $(\cdot|\cdot)_{\mathbf{W}}$, where for $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^N$, $(\mathbf{x}|\mathbf{y})_{\mathbf{W}} = \mathbf{x}^T\mathbf{W}\mathbf{y}$; in this way, $f \in L^2(\mu)$ is assimilated to $\mathbf{f} \in \mathbb{R}^N$, and the operator T_μ then corresponds to the matrix $\mathbf{KW}$. We denote by $\lambda_1 \geqslant \cdots \geqslant \lambda_N \geqslant 0$ the eigenvalues of $\mathbf{KW}$ and by $\mathbf{v}_1, \cdots, \mathbf{v}_N$ the associated orthonormalised eigenvectors, i.e., $\mathbf{KW} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_N)$ and $\mathbf{P} = (\mathbf{v}_1| \cdots |\mathbf{v}_N)$. The vectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_N\}$ form an orthonormal basis of the Hilbert space $\big\{\mathbb{R}^N, (\cdot|\cdot)_{\mathbf{W}}\big\}$, i.e., $\mathbf{P}^T\mathbf{W}\mathbf{P} = \mathrm{Id}_N$, the $N \times N$ identity matrix. In particular, we have $\mathbf{K} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$, and $\mathbf{P}^T\mathbf{P} = \mathbf{W}^{-1}$. For $\lambda_k > 0$, the canonical eigenfunctions of T_μ are given by $\varphi_k = \frac{1}{\lambda_k}\mathbf{k}^T\mathbf{W}\mathbf{v}_k$.

For a general $\boldsymbol{\omega} > 0$, the matrix $\mathbf{KW}$ is non-symmetric and computing its diagonalisation may be numerically difficult. However, classically, since $\mathbf{KW}\mathbf{v}_k = \lambda_k\mathbf{v}_k$, we have

$$\mathbf{W}^{1/2}\mathbf{K}\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{v}_k = \lambda_k\mathbf{W}^{1/2}\mathbf{v}_k.$$

The symmetric matrix $\mathbf{W}^{1/2}\mathbf{K}\mathbf{W}^{1/2}$ thus defines a symmetric and positive-semidefinite operator on $\{\mathbb{R}^N, (\cdot|\cdot)_{\mathrm{Id}_N}\}$ with eigenvalues λ_k and orthonormalised eigenvectors $\mathbf{W}^{1/2}\mathbf{v}_k$. One can therefore easily deduce the eigendecomposition of the matrix $\mathbf{KW}$ viewed as an operator on $\{\mathbb{R}^N, (\cdot|\cdot)_{\mathbf{W}}\}$ from the eigendecomposition the symmetric matrix $\mathbf{W}^{1/2}\mathbf{K}\mathbf{W}^{1/2}$.

2.3. Hilbert-Schmidt norm and squared-kernel discrepancy. In view of Section 2.1, for $\mu \in \mathscr{T}(K)$, the operator T_μ can also be interpreted as an operator on $\mathcal{H}$; with a slight abuse of notation, we keep the same notation for "T_μ viewed as an operator on $L^2(\mu)$", and "T_μ viewed as an operator on $\mathcal{H}$". In both cases, T_μ is an Hilbert-Schmidt operator.

Let μ and $\nu \in \mathscr{T}(K)$; for an o.n.b. $\{h_k | k \in \mathbb{I}\}$ of $\mathcal{H}$ (with $\mathbb{I}$ a general, at most countable, index set), the Hilbert-Schmidt inner product between the operators T_μ and T_ν on $\mathcal{H}$ is given by

$$\big(T_\mu|T_\nu\big)_{\mathrm{HS},\mathcal{H}} = \sum_{k\in\mathbb{I}} \big(T_\mu[h_k]|T_\nu[h_k]\big)_{\mathcal{H}},$$

and we recall that the value of $\big(T_\mu|T_\nu\big)_{\mathrm{HS},\mathcal{H}}$ does not depend on the choice of the o.n.b. of $\mathcal{H}$, see, e.g., [17]. The underlying Hilbert-Schmidt norm (for operators on $\mathcal{H}$) is given by

$$\|T_\mu\|^2_{\mathrm{HS},\mathcal{H}} = \sum_{k\in\mathbb{I}} \|T_\mu[h_k]\|^2_{\mathcal{H}}.$$

Lemma 2.1. *For μ and $\nu \in \mathscr{T}(K)$, we have $(T_\mu|T_\nu)_{\mathrm{HS},\mathcal{H}} = \|K\|^2_{L^2(\mu\otimes\nu)}$. In addition, we have $\|T_\mu\|^2_{\mathrm{HS},\mathcal{H}} = \sum_{k\in\mathbb{I}^+_\mu} \lambda_k^2$, where $\{\lambda_k | k \in \mathbb{I}^+_\mu\}$ is the set of all strictly positive eigenvalues of T_μ.*

By definition, we also remark that if μ and $\nu \in \mathscr{T}(K)$ are such that $\mathcal{H}_\mu$ and $\mathcal{H}_\nu$ are orthogonal subspaces of $\mathcal{H}$, then $\|K\|^2_{L^2(\mu\otimes\nu)} = 0$.

Definition 2.1. *For μ and $\nu \in \mathscr{T}(K)$, the squared-kernel discrepancy between μ and ν is*

$$D_{K^2}(\mu,\nu) = \|T_\mu - T_\nu\|^2_{\mathrm{HS},\mathcal{H}} = \|K\|^2_{L^2(\mu\otimes\mu)} + \|K\|^2_{L^2(\nu\otimes\nu)} - 2\|K\|^2_{L^2(\mu\otimes\nu)}.$$

The terminology "discrepancy" is motivated by the analogy with the notion of "kernel discrepancy" used for instance in [3]. By definition, $D_{K^2}(\mu,\nu)$ appears as a "measure of the similarity" between the operators T_μ and T_ν on $\mathcal{H}$; in particular, $D_{K^2}(\mu,\mu) = 0$. Notice that $\|K\|^2_{L^2(\mu\otimes\nu)} \leqslant \tau_\mu\tau_\nu$.

2.4. Squared-kernel discrepancy and Nyström approximation error. The following Theorem 2.1 further highlights the connection between the squared-kernel discrepancy and the error induced by the approximation of T_μ (the "true" integral operator) by T_ν (the approximate operator).

Theorem 2.1. *Let μ and $\nu \in \mathscr{T}(K)$ be such that $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ (i.e., for $h \in \mathcal{H}$, if $\|h\|_{L^2(\mu)} = 0$ then $\|h\|_{L^2(\nu)} = 0$). Let $\big\{\sqrt{\lambda_k}\varphi_k | k \in \mathbb{I}^+_\mu\big\}$ be the o.n.b. basis of $\mathcal{H}_\mu$ defined by T_μ, and $\{\sqrt{\vartheta_l}\psi_l | l \in \mathbb{I}^+_\nu\}$ be the o.n.b. of $\mathcal{H}_\nu$ defined by T_ν. We have*

$$D_{K^2}(\mu,\nu) = \sum_{k\in\mathbb{I}^+_\mu} \lambda_k \big\|T_\mu[\varphi_k] - T_\nu[\varphi_k]\big\|^2_{\mathcal{H}} \tag{2.1}$$

$$= \sum_{k\in\mathbb{I}^+_\mu} \lambda_k^2 + \sum_{k\in\mathbb{I}^+_\mu}\sum_{l\in\mathbb{I}^+_\nu} \big[\vartheta_l^2 - 2\lambda_k\vartheta_l\big]\big(\sqrt{\lambda_k}\varphi_k\big|\sqrt{\vartheta_l}\psi_l\big)^2_{\mathcal{H}}, \tag{2.2}$$

and, in addition, $\sum_{k\in\mathbb{I}_\mu^+} \lambda_k \left\|T_\mu[\varphi_k] - T_\nu[\varphi_k]\right\|^2_{L^2(\mu)} \leqslant \tau_\mu D_{K^2}(\mu,\nu)$.

For a fixed measure μ, by minimising the function $\nu \mapsto D_{K^2}(\mu,\nu)$ under the constraint $\mathcal{H}_\nu \subset \mathcal{H}_\mu$, we aim at minimising, for the RKHS norm, $\lambda_k\varphi_k - T_\nu[\varphi_k]$ for all $k \in \mathbb{I}_\mu^+$, while putting more emphasis on the approximation of the eigenpairs with largest eigenvalues (the eigenvalues λ_k playing the rule of penalisation weights). We can therefore in particular expect the main eigenfunctions of T_ν to be accurate approximations of the main eigendirections of T_μ (i.e., the one related to the largest eigenvalues). However, under the only condition $\mathcal{H}_\nu \subset \mathcal{H}_\mu$, $D_{K^2}(\mu,\nu)$ is minimal for $\nu = \mu$ (meaning that the best approximation of T_μ is T_μ itself); in order to avoid this trivial solution, one may add additional feasibility constraints for ν, or add a regularisation term to the cost $D_{K^2}(\mu,\nu)$. By analogy with spectral truncation, one possibility is to penalise $\tau_\nu = \mathrm{trace}(T_\nu) = \int_{\mathscr{X}} K(x,x)d\nu(x)$, see Section 3; more generally, in this work, we consider penalisations of the form $\int_{\mathscr{X}} f(x)d\nu(x)$, where f is a real-valued, positive, measurable function on $\mathscr{X}$.

3. Optimal quadrature-sparsification as quadratic programming. We assume that the measure μ is discrete and has a finite number of support points (framework of Section 2.2), and we enforce the support of the searched after approximation ν to be included in the support of μ. As already mentioned, since $D_{K^2}(\mu,\mu) = 0$, the minimisation of the squared-kernel discrepancy alone is not interesting (i.e., minimisation of $\nu \mapsto D_{K^2}(\mu,\nu)$ for μ fixed); in addition, we aim at obtaining an as sparse as possible discrete measure ν. In order to discard the trivial solution $\nu = \mu$, we introduce a linear penalisation.

More precisely, let $\mu = \sum_{k=1}^N \omega_k\delta_{x_k}$ and $\nu = \sum_{k=1}^N \upsilon_k\delta_{x_k}$ be discrete measures supported by $\mathcal{S}$, with $\boldsymbol{\omega} > 0$ and $\boldsymbol{\upsilon} \geqslant 0$ (the notation $\boldsymbol{\upsilon} \geqslant 0$ meaning that all the components of $\boldsymbol{\upsilon}$ are positive); we thus have $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ for any such $\boldsymbol{\upsilon}$ and

$$D_{K^2}(\mu,\nu) = (\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}),$$

where $\mathbf{S}$ is the kernel-matrix defined by the squared-kernel $K^2(\cdot,\cdot)$ and the set of points $\mathcal{S}$, i.e., with i,j entry $K^2(x_i,x_j) \geqslant 0$ ($\mathbf{S}$ is therefore a non-negative, positive-semidefinite, symmetric matrix). For a fixed measure μ (i.e., $\boldsymbol{\omega}$ is fixed), we define

$$D(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}),$$

the factor 1/2 being added for simplification purpose.

3.1. Relation with the classical Frobenius-norm-based criterion. In the framework of Section 3, the squared-kernel discrepancy appears as a natural extension of the classical Frobenius-norm-based criterion for matrix low-rank approximation (see for instance [4]). Indeed, assume that $\boldsymbol{\omega} = \mathbb{1}$, with $\mathbb{1} = (1,\cdots,1)^T \in \mathbb{R}^N$, so that $\mathbf{W} = \mathrm{Id}_N$ is the $N\times N$ identity matrix, and thus $\mathbf{KW} = \mathbf{W}^{1/2}\mathbf{KW}^{1/2} = \mathbf{K}$ (notice the analogy with the notations used in Section 2.2).

Selecting columns of the matrix $\mathbf{K}$ can be viewed as performing the product $\mathbf{KV}$, where $\mathbf{V} = \mathrm{diag}(\boldsymbol{\upsilon})$ is a $N\times N$ diagonal matrix with diagonal entries 0 or 1 (we are therefore also assuming that all the components of $\boldsymbol{\upsilon}$ are 0 or 1); in the same way, considering $\mathbf{VK}$ amounts to selecting rows of $\mathbf{K}$. Since such a sampling matrix $\mathbf{V}$ satisfies $\mathbf{V}^2 = \mathbf{V}$, selecting a principal submatrix of the matrix $\mathbf{K}$ can be written $\mathbf{V}^{1/2}\mathbf{KV}^{1/2} = \mathbf{VKV}$.

Lemma 3.1. *If* $\boldsymbol{\omega} = \mathbb{1}$ *and if the components of* $\boldsymbol{\upsilon}$ *are all* 0 *or* 1*, then* $(\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}) = \|\mathbf{K}-\mathbf{VKV}\|_F^2$*, where* $\|\cdot\|_F$ *stands for the Frobenius norm.*

3.2. Regularised squared-kernel-discrepancy minimisation. For a given penalisation vector $\mathbf{d} = (d_1,\cdots,d_N)^T \in \mathbb{R}^N$, $\mathbf{d} > 0$ (see Section 3.4), and $\alpha \geqslant 0$, we introduce the regularised squared-kernel-discrepancy minimisation problem

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D_\alpha(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}) + \alpha\mathbf{d}^T\boldsymbol{\upsilon} \text{ subject to } \boldsymbol{\upsilon} \geqslant 0. \tag{3.1}$$

Notice that $D_\alpha(\boldsymbol{\upsilon}) = D(\boldsymbol{\upsilon}) + \alpha\mathbf{d}^T\boldsymbol{\upsilon}$. In particular, when $\mathbf{S}$ is invertible, $D_\alpha(\cdot)$ is strongly convex (and, in this case, a solution to (3.1) is therefore necessarily unique). We also recall that, for a given α, the set of solutions to (3.1) is convex. The gradient of D_α at $\boldsymbol{\upsilon} \in \mathbb{R}^N$ is given by $\nabla D_\alpha(\boldsymbol{\upsilon}) = \mathbf{S}(\boldsymbol{\upsilon}-\boldsymbol{\omega}) + \alpha\mathbf{d}$. The following Lemma 3.2 recalls some simple properties verified by the solutions to problem (3.1).

Lemma 3.2. *Denote by $\boldsymbol{\upsilon}_\alpha^*$ a solution to* (3.1) *with $\alpha \geqslant 0$, we have:*

(i) for $\alpha = 0$, $\boldsymbol{\upsilon}_\alpha^* = \boldsymbol{\omega}$ is a solution to (3.1)*,*

(ii) if $\alpha \geqslant \max_k \left\{[\mathbf{S}\boldsymbol{\omega}]_k / d_k\right\}$, then $\boldsymbol{\upsilon}_\alpha^* = 0$ (with $[\mathbf{S}\boldsymbol{\omega}]_k$ the k-th component of $\mathbf{S}\boldsymbol{\omega}$),

(iii) for all $\alpha \geqslant 0$, we have $0 \leqslant \alpha \mathbf{d}^T \boldsymbol{\upsilon}_\alpha^* \leqslant \alpha \mathbf{d}^T \boldsymbol{\omega} - (\boldsymbol{\omega} - \boldsymbol{\upsilon}_\alpha^*)^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}_\alpha^*)$,

(iv) if $\widetilde{\boldsymbol{\upsilon}}_\alpha^*$ is another solution to (3.1)*, then $\mathbf{d}^T \widetilde{\boldsymbol{\upsilon}}_\alpha^* = \mathbf{d}^T \boldsymbol{\upsilon}_\alpha^*$,*

(v) if $[\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}]_k \geqslant 0$ and $\mathbf{S}_{k,k} > 0$ (see Remark 3.1), then $[\boldsymbol{\upsilon}_\alpha^*]_k = 0$,

(vi) the map $\alpha \mapsto D(\boldsymbol{\upsilon}_\alpha^*)$ is increasing, and $\alpha \mapsto \mathbf{d}^T \boldsymbol{\upsilon}_\alpha^*$ is decreasing.

Since $\boldsymbol{\upsilon} \geqslant 0$, the term $\mathbf{d}^T \boldsymbol{\upsilon}$ can be interpreted as a weighted ℓ^1-regularisation (and α is the regularisation parameter). For appropriate $\mathbf{d}$ and α, we can thus expect a solution $\boldsymbol{\upsilon}_\alpha^*$ to (3.1) to be sparse (see, e.g., [10]). This intuition is confirmed by Lemma 3.2-(v); indeed, when all the diagonal entries of $\mathbf{S}$ are strictly positive, $\mathrm{card}(\{k | [\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}]_k < 0\})$ gives an upper bound on the number of strictly positive components of $\boldsymbol{\upsilon}_\alpha^*$ (notice that this bound is generally not tight). See Section 3.4 for a further discussion.

Remark 3.1. Assuming $\mathbf{S}_{k,k} = K^2(x_k, x_k) > 0$ for all $k \in \{1, \cdots, N\}$ (which we denote by $\mathrm{diag}(\mathbf{S}) > 0$) is equivalent to assuming $K(x_k, x_k) > 0$ for all k; we recall that for all $x \in \mathscr{X}$, we have $K(x, x) = \|K_x\|_{\mathcal{H}}^2 \geqslant 0$. This assumption is in practice not restrictive at all: indeed, if $K(x_k, x_k) = 0$, then $K_{x_k} = 0$ and thus $h(x_k) = 0$ for all $h \in \mathcal{H}$. In the framework of Section 3, such a point x_k may thus be removed from the sample $\mathcal{S}$ without inducing any modification of the operators T_μ and T_ν. ◁

3.3. Constrained squared-kernel-discrepancy minimisation. Instead of considering problem (3.1), we can equivalently introduce, for $\varkappa \geqslant 0$ (and $\varkappa \leqslant \mathbf{d}^T \boldsymbol{\omega}$, see Lemma 3.3)

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) \text{ subject to } \boldsymbol{\upsilon} \geqslant 0 \text{ and } \mathbf{d}^T \boldsymbol{\upsilon} = \varkappa. \tag{3.2}$$

Notice that problem (3.2) consists in minimising a convex function on a convex compact domain, so that a solution $\boldsymbol{\upsilon}_\varkappa^*$ to problem (3.2) always exists; in particular, if $\mathbf{S}$ is non-singular, then $\boldsymbol{\upsilon}_\varkappa^*$ is always unique (since $D(\cdot)$ is in this case strongly convex).

Lemma 3.3. *Let $\boldsymbol{\upsilon}_\alpha^*$ be a solution to problem* (3.1) *with $\alpha \geqslant 0$; then $\boldsymbol{\upsilon}_\alpha^*$ is a solution to problem* (3.2) *with $\varkappa = \mathbf{d}^T \boldsymbol{\upsilon}_\alpha^*$. Reciprocally, assume that $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to problem* (3.2) *with $0 < \varkappa \leqslant \mathbf{d}^T \boldsymbol{\omega}$; then $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to problem* (3.1) *with $\alpha = (\boldsymbol{\upsilon}_\varkappa^*)^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}_\varkappa^*)/\varkappa$. For $\varkappa = 0$, we have $\boldsymbol{\upsilon}_\varkappa^* = 0$, which is solution to problem* (3.1) *with $\alpha \geqslant \max_k \left\{[\mathbf{S}\boldsymbol{\omega}]_k / d_k\right\}$. For $0 \leqslant \varkappa \leqslant \mathbf{d}^T \boldsymbol{\omega}$, the map $\varkappa \mapsto D(\boldsymbol{\upsilon}_\varkappa^*)$ is decreasing.*

As an interesting feature, problem (3.2) can be efficiently solved thanks to a kernelised (i.e., without matrix storage) sparse descent direction QP solver, like for instance the vertex-exchange strategy, see [15, Chap. 9] and Section 7.1. A sequential strategy (based on the notion of regularisation path) for solving problems (3.1) and (3.2) is also discussed in Section 6

Notice that, in view of Lemma 3.2-(iii) and Lemma 3.3, considering $\varkappa = \rho \mathbf{d}^T \boldsymbol{\omega}$ with $\rho \in [0, 1]$ appears as a natural parameterisation for problem (3.2).

Remark 3.2. In case of non-uniqueness of the solution to problem (3.2) with $\varkappa \geqslant 0$, two solutions $\boldsymbol{\upsilon}_\varkappa^*$ and $\widetilde{\boldsymbol{\upsilon}}_\varkappa^*$ necessarily satisfy $\mathbf{S}(\boldsymbol{\upsilon}_\varkappa^* - \widetilde{\boldsymbol{\upsilon}}_\varkappa^*) = 0$. In exactly the same way, from Lemma 3.2-(iv) or Lemma 3.3, in case of non-uniqueness of the solution to problem (3.1) with $\alpha \geqslant 0$, two solutions $\boldsymbol{\upsilon}_\alpha^*$ and $\widetilde{\boldsymbol{\upsilon}}_\alpha^*$ necessarily satisfy $\mathbf{S}(\boldsymbol{\upsilon}_\alpha^* - \widetilde{\boldsymbol{\upsilon}}_\alpha^*) = 0$. ◁

3.4. Choice of the penalisation. In order to discard the solution $\boldsymbol{\upsilon} = \boldsymbol{\omega}$ (i.e., $\nu = \mu$), in Sections 3.2 and 3.3, we have introduced a penalisation based on the term $\mathbf{d}^T \boldsymbol{\upsilon}$ (the penalisation appearing in the cost or in the constraint), for a given $\mathbf{d} \in \mathbb{R}^N$ with $\mathbf{d} > 0$. Notice that the term $\mathbf{d}^T \boldsymbol{\upsilon}$ can be interpreted as the integral with respect to ν of a real valued function $f_{\mathbf{d}}$ on $\mathscr{X}$ satisfying $f_{\mathbf{d}}(x_k) = d_k$; in this case, we indeed have $\mathbf{d}^T \boldsymbol{\upsilon} = \int_{\mathscr{X}} f_{\mathbf{d}}(x) \mathrm{d}\nu(x)$. In practice, we ideally aim at obtaining a vector $\boldsymbol{\upsilon}^*$ which is both as sparse as possible and such that $D(\boldsymbol{\upsilon}^*)$ is as small as possible; this naturally raises questions relative to the choice of the penalisation vector $\mathbf{d}$ (see also Section 10).

We can first remark that for $\mathbf{d} = \theta \mathbf{S}\boldsymbol{\omega}$ with $\theta > 0$ and for $\alpha \leqslant 1/\theta$, $\boldsymbol{\upsilon}_\alpha^* = (1 - \alpha\theta)\boldsymbol{\omega}$ is a solution to (3.1) (and $\boldsymbol{\upsilon}_\alpha^* = 0$ for $\alpha > 1/\theta$); the solution $\boldsymbol{\upsilon}_\alpha^*$ is therefore non-sparse, and such a choice for the penalisation vector $\mathbf{d}$ appears to be of little interest. More generally, for $\mathbf{d} = \mathbf{S}\boldsymbol{\eta} > 0$ (with $\boldsymbol{\eta} \in \mathbb{R}^N$), a similar remark holds for all α such that $\boldsymbol{\omega} - \alpha\boldsymbol{\eta} \geqslant 0$; indeed, we have in this case $\boldsymbol{\upsilon}_\alpha^* = \boldsymbol{\omega} - \alpha\boldsymbol{\eta}$.

As mentioned in Remark 3.1, we may reasonably assume that $K(x_k, x_k) > 0$ for all k. In this case, by analogy with spectral truncation, choosing $d_k = K(x_k, x_k)$ appears as relatively interesting since it leads to $\mathbf{d}^T \boldsymbol{\upsilon} = \mathrm{trace}(T_\nu)$ (and $\mathrm{trace}(T_\mu) = \mathbf{d}^T \boldsymbol{\omega}$); in this case, we shall use the notation $\mathbf{d} = \mathrm{diag}(\mathbf{K})$. This choice therefore appears as relatively natural and interesting and is the one considered in the experiments of Sections 8 and 9.

4. Approximate eigendecomposition. We consider the framework of Section 3. Let $\boldsymbol{\upsilon} \geqslant 0$ with $\boldsymbol{\upsilon} \neq 0$ (we recall that $\boldsymbol{\upsilon} \in \mathbb{R}^N$); in practice, $\boldsymbol{\upsilon}$ will be a solution (or an approximate solution) to problem (3.1) or (3.2) (or of one of the extensions discussed in Sections 5.3), with $\mathbf{d} > 0$ and $\alpha \geqslant 0$ or $0 \leqslant \varkappa \leqslant \mathbf{d}^T \boldsymbol{\omega}$ (more particularly, since we are not interested in the case $\boldsymbol{\upsilon} = 0$, we shall consider $\alpha < max_k[\mathbf{S}\boldsymbol{\omega}]_k/d_k$ or $\varkappa > 0$). We recall that we denote by $\nu = \sum_{k=1}^N \upsilon_k \delta_{x_k}$ the measure related to $\boldsymbol{\upsilon}$.

We introduce the index set $I = \{i | \upsilon_i > 0\}$ and let $n = \mathrm{card}(I)$ be the number of strictly positive components of $\boldsymbol{\upsilon}$; we have in particular $\nu = \sum_{i \in I} \upsilon_i \delta_{x_i}$ (i.e., we have discarded all the points x_k such that $\upsilon_k = 0$). Following Section 2.4, the strictly positive eigenvalues $\{\vartheta_k | k \in \mathbb{I}_\nu^+\}$ of T_ν and their associated canonically extended eigenfunction $\psi_k \in \mathcal{H}$, orthonormalised for $L^2(\nu)$, can be easily obtained from the eigendecomposition of the $n \times n$ (symmetric and positive-semidefinite) principal submatrix $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I}$. We thus obtain the o.n.b. $\{\sqrt{\vartheta_l}\psi_l | l \in \mathbb{I}_\nu^+\}$ of the subspace $\mathcal{H}_\nu$ of $\mathcal{H}$ related to T_ν.

Depending on the considered application, (i.e., depending on the reason why the eigendecomposition of T_μ is needed), we may directly use the eigenpairs $\{(\vartheta_l, \psi_l) | l \in \mathbb{I}_\nu^+\}$ related to T_ν in order to approximate the eigenpairs $\{(\lambda_k, \varphi_k) | k \in \mathbb{I}_\mu^+\}$ related to T_μ. However, it is also possible to define more relevant approximations. In what follows, we discuss possible ways to approximate the main eigenpairs of the integral operator T_μ from the eigendecomposition of the integral operator T_ν.

Normalised approximated eigenfunctions. Since the eigenfunctions φ_k of T_μ, with $k \in \mathbb{I}_\mu^+$, are such that $\|\widehat{\varphi}_l\|_{L^2(\mu)} = 1$, for all $l \in \mathbb{I}_\nu^+$, we can normalise the eigenfunctions ψ_k as well. We thus introduce the *normalised approximate eigenfunctions* of T_μ induced by T_ν, given by, for $l \in \mathbb{I}_\nu^+$,

$$\widehat{\varphi}_l = \psi_l / \|\psi_l\|_{L^2(\mu)}, \tag{4.1}$$

so that $\|\widehat{\varphi}_l\|_{L^2(\mu)} = 1$. Notice however that the true eigenfunctions $\{\varphi_k | k \in \mathbb{I}_\mu^+\}$ are orthogonal in $L^2(\mu)$, while the approximated eigenfunctions $\{\widehat{\varphi}_l | l \in \mathbb{I}_\nu^+\}$ are orthogonal in $L^2(\nu)$.

Remark 4.1. Controlling the orthogonality, in $L^2(\mu)$, between the approximate eigenfunctions $\widehat{\varphi}_l$ appears as a relatively inexpensive way to evaluate the accuracy of the approximate eigenfunctions. Indeed, accurate approximate eigenfunctions $\widehat{\varphi}_l$ should be almost mutually orthogonal in $L^2(\mu)$; this condition is however only a necessary condition. ◁

Rescaled eigenvalues. Let $\varrho \geqslant 0$ be such that $\mathrm{trace}(T_\nu) = \varrho\, \mathrm{trace}(T_\mu)$; if $\varrho > 0$ (i.e., $\boldsymbol{\upsilon} \neq 0$ and $\mathrm{diag}(\mathbf{K}) > 0$), for $l \in \mathbb{I}_\nu^+$, we may introduce the rescaled eigenvalues ϑ_l/ϱ, so that $\sum_{l \in \mathbb{I}_\nu^+} \vartheta_l/\varrho = \mathrm{trace}(T_\mu)$. As an interesting feature, such a rescaling involves any expensive computation, but the quality of the induced approximation is in general relatively poor.

One may also consider a corrected rescaling factor ϱ_c, taking for instance into account that we approximate only the n_ν largest eigenvalues of T_μ, where $n_\nu = \mathrm{card}(\mathbb{I}_\nu^+)$ is the number of strictly positive eigenvalues of T_ν. Depending on the application, another possibility may consists in rescaling the spectrum in such a way that the largest rescaled eigenvalue is equal to a given value.

Upsilon-test and induced approximated eigenvalues. By definition, for all $k \in \mathbb{I}_\mu^+$, we have

$$(\varphi_k | T_\mu[\varphi_k])_{L^2(\mu)} / \|T_\mu[\varphi_k]\|_{L^2(\mu)} = 1 \text{ and } \lambda_l = \|T_\mu[\varphi_k]\|_{L^2(\mu)}.$$

By analogy, for all $l \in \mathbb{I}_\nu^+$, considering the normalised approximate eigenfunctions $\widehat{\varphi}_l$ given by (4.1), we define

$$\Upsilon_l = (\widehat{\varphi}_l | T_\mu[\widehat{\varphi}_l])_{L^2(\mu)} / \|T_\mu[\widehat{\varphi}_l]\|_{L^2(\mu)}. \tag{4.2}$$

Since T_μ is a positive-semidefinite operator on $L^2(\mu)$, we have $\Upsilon_l \geqslant 0$; in addition, from the Cauchy-Schwarz inequality, we have $\Upsilon_l \leqslant 1$, so that $\Upsilon_l \in [0, 1]$. Notice that Υ_l simply consists in the $L^2(\mu)$

inner product between the two normalised vector $\widehat{\varphi}_l$ and $T_\mu[\widehat{\varphi}_l]/\|T_\mu[\widehat{\varphi}_l]\|_{L^2(\mu)}$. We also introduce

$$\widehat{\lambda}_l = \|T_\mu[\widehat{\varphi}_l]\|_{L^2(\mu)}. \tag{4.3}$$

So, the closer Υ_l is to 1, the closer $\widehat{\varphi}_l$ is to be an eigenfunction of T_μ. In addition, for $\Upsilon_l \approx 1$, $\widehat{\lambda}_l$ appears as a relevant approximation of the underlying eigenvalue of T_μ (i.e., the eigenvalue related to the eigenfunction approximated by $\widehat{\varphi}_l$). This intuition is confirmed by the following Lemma 4.1.

Lemma 4.1. *For $l \in \mathbb{I}_\nu^+$, consider $\widehat{\varphi}_l$, Υ_l and $\widehat{\lambda}_l$ given by* (4.1)*,* (4.2) *and* (4.3) *respectively; we have*

$$\|T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l\widehat{\varphi}_l\|^2_{L^2(\mu)} = 2\widehat{\lambda}_l^2(1-\Upsilon_l), \text{ and } \|T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l\widehat{\varphi}_l\|^2_{\mathcal{H}} = \widehat{\lambda}_l(\Upsilon_l + \|\widehat{\lambda}_l^{1/2}\widehat{\varphi}_l\|^2_{\mathcal{H}} - 2).$$

From Lemma 4.1, since $\|T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l\widehat{\varphi}_l\|^2_{\mathcal{H}} \geqslant 0$, we can also deduce that for $l \in \mathbb{I}_\nu^+$, we have $\|\widehat{\lambda}_l^{1/2}\widehat{\varphi}_l\|^2_{\mathcal{H}} \geqslant 2 - \Upsilon_l$; we recall that by definition, $\|\sqrt{\lambda_k}\varphi_k\|^2_{\mathcal{H}} = 1$, for all $k \in \mathbb{I}_\mu^+$.

The Upsilon-test appears has a very interesting and simple tool since it indicates the accuracy of the approximate eigenpairs of T_μ induced by T_ν. However, for $\mu = \sum_{k=1}^N \omega_k \delta_{x_k}$, computing $T_\mu[\widehat{\varphi}_l]$ consists in a kernelised matrix-vector product, with complexity scaling as $\mathcal{O}(N^2)$; this operation is therefore relatively costly, but may however be easily parallelised.

RKHS-norm-induced approximate eigenvalues. As already mentioned, the numerical complexity of the Upsilon-test (and therefore, of the computation of the approximate eigenvalues $\widehat{\lambda}_l$) may become prohibitive for very large N (depending on the computational power at disposal). One can however also define approximate eigenvalues involving computations with numerical complexity scaling as $\mathcal{O}(n^2)$, with n the number of support points of the considered approximate measure ν.

Indeed, since $\|\sqrt{\lambda_k}\varphi_k\|^2_{\mathcal{H}} = 1$, we have $\lambda_k = 1/\|\varphi_k\|^2_{\mathcal{H}}$. This therefore suggests to define the approximate eigenvalues, for $l \in \mathbb{I}_\nu^+$,

$$\widetilde{\lambda}_l = 1/\|\widehat{\varphi}_l\|^2_{\mathcal{H}}. \tag{4.4}$$

From Section 2.2, the computation of the approximate eigenvalues $\widetilde{\lambda}_l$ only involves the $n \times n$ principal kernel-submatrix $\mathbf{K}_{I,I}$, and may in addition be easily parallelised. If $\widehat{\varphi}_l$ is close to be a true normalised eigenfunction of T_μ, then we can expect $\widetilde{\lambda}_l$ to be a accurate approximation of the underlying eigenvalue of T_μ. For $l \in \mathbb{I}_\nu^+$, we have $\widehat{\lambda}_l \geqslant (2-\Upsilon_l)\widetilde{\lambda}_l$, with equality when $\Upsilon_l = 1$.

Remark 4.2. Consider any measure $\nu \in \mathscr{T}(K)$; the approximations $\widehat{\varphi}_l$, ϑ_l/ϱ, $\widehat{\lambda}_l$ and $\widetilde{\lambda}_l$ (i.e., the normalised approximated eigenfunctions of T_μ induced by T_ν, the rescaled eigenvalues, and approximated eigenvalues (4.3) and (4.4), respectively) remain unchanged if one replaces ν by $\theta\nu$ for any $\theta > 0$. ◁

5. Analogy with one-class SVM. Let $\mathcal{F}$ be the RHKS associated with the squared-kernel $K^2(\cdot,\cdot)$. Following [16], problems (3.1) and (3.2) can be interpreted as dual formulations of *squared-kernel one-class distorted SVMs* (or discrepancy-SVMs).

We introduce the function $g_{\boldsymbol{\omega}} \in \mathcal{F}$, defined by $g_{\boldsymbol{\omega}}(x) = \sum_{k=1}^N \omega_k K^2(x, x_k)$, we shall refer to $g_{\boldsymbol{\omega}}$ as the *primal distortion term*.

5.1. One-class SVM related to the regularised problem. We first discribe the SVM related to problem (3.1). For $g \in \mathcal{F}$, we consider the convex minimisation problem

$$\begin{array}{ll} \underset{g}{\text{minimise}} & \frac{1}{2}\|g\|^2_{\mathcal{F}} + (g|g_{\boldsymbol{\omega}})_{\mathcal{F}} \\ \text{subject to} & g(x_k) \geqslant -\alpha d_k \text{ for all } k \in \{1,\cdots,N\}. \end{array} \tag{5.1}$$

The application $g \mapsto \|g\|^2_{\mathcal{F}}$ being strictly convex, a solution to problem (5.1) is necessarily unique.

Lemma 5.1. *If $\boldsymbol{\upsilon}^*_\alpha$ is a solution to* (3.1) *with $\alpha \geqslant 0$, then $g^*_\alpha(x) = \sum_{k=1}^N [\boldsymbol{\upsilon}^*_\alpha - \boldsymbol{\omega}]_k K^2(x, x_k)$ is the solution to* (5.1)*. For all $k \in \{1,\cdots,N\}$ such that $[\boldsymbol{\upsilon}^*_\alpha]_k > 0$, we have $g^*_\alpha(x_k) = -\alpha d_k$ (notice that for all k, we have $g^*_\alpha(x_k) = [\mathbf{S}(\boldsymbol{\upsilon}^*_\alpha - \boldsymbol{\omega})]_k$).*

By introducing the change of variable $\check{g} = g + g_{\boldsymbol{\omega}} \in \mathcal{F}$, problem (5.1) leads to (up to an additive constant)

$$\begin{array}{ll} \underset{\check{g},\gamma}{\text{minimise}} & \frac{1}{2}\|\check{g}\|_{\mathcal{F}}^2 \\ \text{subject to} & \check{g}(x_k) - g_{\boldsymbol{\omega}}(x_k) \geqslant -\alpha d_k \text{ for all } k \in \{1, \cdots, N\}, \end{array} \tag{5.2}$$

which is an equivalent formulation for (5.1), with solution $\check{g}_\alpha^*(x) = \sum_{k=1}^N [\boldsymbol{\upsilon}_\alpha^*]_k K^2(x, x_k)$. Notice that, for all $x \in \mathscr{X}$, $K^2(x, \cdot) \geqslant 0$, so that $g_{\boldsymbol{\omega}}$ and $\check{g}_\alpha^*$ are positive functions on $\mathscr{X}$.

5.2. One-class SVM related to the constrained problem. We now describe the SVM related to problem (3.2). For $g \in \mathcal{F}$ and $\gamma \in \mathbb{R}$, we introduce the problem

$$\begin{array}{ll} \underset{g,\gamma}{\text{minimise}} & \frac{1}{2}\|g\|_{\mathcal{F}}^2 + (g|g_{\boldsymbol{\omega}})_{\mathcal{F}} - \gamma \\ \text{subject to} & g(x_k) \geqslant \gamma d_k/\varkappa \text{ for all } k \in \{1, \cdots, N\}. \end{array} \tag{5.3}$$

Again, a solution to problem (5.3) is necessarily unique.

Lemma 5.2. *If $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to* (3.2)*, then $g_\varkappa^*(x) = \sum_{k=1}^N [\boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega}]_k K^2(x, x_k)$ and $\gamma_\varkappa^* = (\boldsymbol{\upsilon}_\varkappa^*)^T \mathbf{S}(\boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega})$ is the solution to* (5.3)*. For all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{\upsilon}_\varkappa^*]_k > 0$, we have $g_\varkappa^*(x_k) = \gamma_\varkappa^* d_k/\varkappa$.*

From Lemma 5.2, we have $\gamma_\varkappa^* = \|g_\varkappa^*\|_{\mathcal{F}}^2 + (g_\varkappa^*|g_{\boldsymbol{\omega}})_{\mathcal{F}}$. In view of Lemma 3.3, for $0 < \varkappa \leqslant \mathbf{d}^T\boldsymbol{\omega}$, we know that $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to (3.1) for $\alpha = -\gamma_\varkappa^*/\varkappa$, and since $\alpha \geqslant 0$, we therefore have $\gamma_\varkappa^* \leqslant 0$.

5.3. Soft-margin-type extensions. Pursuing the analogy with one-class SVMs, we can define soft-margin-type extensions of problems (5.1) and (5.3), i.e, we can consider models where the inequalities appearing in the constraints can, potentially, be violated, the level of violation being penalised. In this section, we only discuss extensions of problem (5.1), but a similar discussion can be done for problem (5.3).

We introduce $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_N)^T \in \mathbb{R}^N$; the components of $\boldsymbol{\xi}$ are referred to as *slack variables*. Instead of considering the contraints $g(x_k) \geqslant -\alpha d_k$, we can consider the relaxed constraints $g(x_k) \geqslant -\alpha d_k - \xi_k$, while penalising the values taken by ξ_k. The penalisation considered is related to a *loss function*, see for instance [19]. In what follows, we discuss the model obtained for two popular types of loss functions: the (weighted) hinge loss and the (generalised) square loss.

In view of our initial problem (3.1) (regularised squared-kernel-discrepancy minimisation), soft-margin extensions of problem (5.1) appear as tool to further constrain or penalise the measure ν (i.e., the vector $\boldsymbol{\upsilon}$) used to approximate the initial measure μ (i.e., the vector $\boldsymbol{\omega}$).

Weighted hinge loss. Let $\mathbf{c} \in \mathbb{R}^N$, with $\mathbf{c} \geqslant 0$; the soft-margin extension of problem (5.1) corresponding to a weighted hinge loss consists in the problem

$$\begin{array}{ll} \underset{g,\boldsymbol{\xi}}{\text{minimise}} & \frac{1}{2}\|g\|_{\mathcal{F}}^2 + (g|g_{\boldsymbol{\omega}})_{\mathcal{F}} + \mathbf{c}^T\boldsymbol{\xi} \\ \text{subject to} & g(x_k) \geqslant -\alpha d_k - \xi_k, \text{ with } \xi_k \geqslant 0, \text{ for all } k \in \{1, \cdots, N\}. \end{array} \tag{5.4}$$

The Lagrange dual of problem (5.4) is given by

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D_\alpha(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) + \alpha\mathbf{d}^T\boldsymbol{\upsilon} \text{ subject to } 0 \leqslant \boldsymbol{\upsilon} \leqslant \mathbf{c}. \tag{5.5}$$

Problem (5.5) only differs from problem (3.1) by the presence of the additional constraints $\boldsymbol{\upsilon} \leqslant \mathbf{c}$, which acts as an upper bound on the values taken by the components of $\boldsymbol{\upsilon}$ (i.e., an upper bound on the weights of the points in the quadrature related to ν).

Generalised square loss. Let $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ be a symmetric positive-definite matrix; the soft-margin extension of problem (5.1) corresponding to a generalised square loss consists in the problem

$$\begin{array}{ll} \underset{g}{\text{minimise}} & \frac{1}{2}\|g\|_{\mathcal{F}}^2 + (g|g_{\boldsymbol{\omega}})_{\mathcal{F}} + \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi} \\ \text{subject to} & g(x_k) \geqslant -\alpha d_k - \xi_k, \text{ for all } k \in \{1, \cdots, N\}. \end{array} \tag{5.6}$$

The Lagrange dual of problem (5.6) is given by

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D_{\alpha,\boldsymbol{\Sigma}}(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) + \tfrac{1}{2}\boldsymbol{\upsilon}^T\boldsymbol{\Sigma}\boldsymbol{\upsilon} + \alpha\mathbf{d}^T\boldsymbol{\upsilon} \text{ subject to } \boldsymbol{\upsilon} \geqslant 0. \tag{5.7}$$

In comparison to problem (3.1), the term $\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}$ is added to the initial cost $D_\alpha(\boldsymbol{v})$, and tends to "harmonise" the components of the underlying solution $\boldsymbol{v}^*$. In particular, notice that $D_{\alpha,\boldsymbol{\Sigma}}$ is strongly convex.

6. Regularisation path. Considering the framework of Sections 3 and 5 and following the classical results relative to the regularisation paths for Lasso or SVM models (see e.g., [13, 9]), we now discuss the regularisation paths related to problems (3.1) and (3.2). In what follows, we mainly consider problem (3.1) (i.e., the regularised model); results related to problem (3.2) can then be obtained from Lemma 3.3.

6.1. Generalities. Let $\boldsymbol{v}^*_\alpha$ be a solution to (3.1) for $\alpha \geqslant 0$; we introduce the index sets

$$J_\alpha = \{k | [\nabla D_\alpha(\boldsymbol{v}^*_\alpha)]_k = 0\} \text{ and } J^c_\alpha = \{1, \cdots, N\} \backslash J_\alpha,$$

so that, by definition, $[\nabla D_\alpha(\boldsymbol{v}^*_\alpha)]_{J^c_\alpha} = [\mathbf{S}(\boldsymbol{v}^*_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{d}]_{J^c_\alpha} > 0$. From Sections 3 and 5, the index set J_α is unique, even when the solution to (3.1) is not (i.e., in case of non-uniqueness of the solution, J_α does not depends on the solution $\boldsymbol{v}^*_\alpha$ considered). We shall refer to J_α as the *sparsity pattern* of the solutions to problem (3.1) for $\alpha \geqslant 0$. We also recall that for all k such that $[\boldsymbol{v}^*_\alpha]_k > 0$, we have $k \in J_\alpha$, see for instance Lemma 5.1.

Knowing J_α, a solution $\boldsymbol{v}^*_\alpha$ is characterised by the conditions

$$[\boldsymbol{v}^*_\alpha]_{J^c_\alpha} = 0, \text{ and } \mathbf{S}_{J_\alpha,J_\alpha}[\boldsymbol{v}^*_\alpha]_{J_\alpha} = [\mathbf{S}\boldsymbol{\omega}]_{J_\alpha} - \alpha\mathbf{d}_{J_\alpha} \text{ with } [\boldsymbol{v}^*_\alpha]_{J_\alpha} \geqslant 0, \tag{6.1}$$

where $\mathbf{S}_{J_\alpha,J_\alpha}$ stands for the $n_\alpha \times n_\alpha$ principal submatrix of $\mathbf{S}$ corresponding to the index set J_α, with $n_\alpha = \text{card}(J_\alpha)$, and where, for instance, $\mathbf{d}_{J_\alpha} \in \mathbb{R}^{n_\alpha}$ stands for the vector defined by the components of $\mathbf{d}$ with index in J_α. Note that the condition $[\boldsymbol{v}^*_\alpha]_{J_\alpha} \geqslant 0$ is involved only in case of non-uniqueness of the solution.

Lemma 6.1. *Let $\boldsymbol{v}^*_{\alpha_1}$ and $\boldsymbol{v}^*_{\alpha_2}$ be solutions to problem* (3.1) *with α_1 and $\alpha_2 \geqslant 0$, respectively. Asume that $J_{\alpha_1} = J_{\alpha_2}$, then for all $\theta \in [0,1]$, $\boldsymbol{v}^*_\alpha = \theta\boldsymbol{v}^*_{\alpha_1} + (1-\theta)\boldsymbol{v}^*_{\alpha_2}$ is a solution to problem* (3.1) *with $\alpha = \theta\alpha_1 + (1-\theta)\alpha_2$.*

If, for all $\alpha \geqslant 0$, the solution $\boldsymbol{v}^*_\alpha$ is unique (which is for instance the case when $\mathbf{S}$ is invertible), then in view of Lemma 6.1, the *regularisation map* $R : \alpha \mapsto \boldsymbol{v}^*_\alpha$ is a piecewise linear application from $\mathbb{R}_+$ into $\mathbb{R}^N$. In case of uniqueness of the solutions, the regularisation map is therefore piecewise continuous and has right and left limits for all α, these limits satisfying the optimality conditions for problem (3.1); by uniqueness of the solution, we can in this way shows that $R : \alpha \mapsto \boldsymbol{v}^*_\alpha$ is continuous. In case of non-uniqueness of the solution, Lemma 6.1 shows that the set of solutions related to a same sparsity pattern J is convex.

When α decreases or increases, we refer to a change in the sparsity pattern J_α as an *event*; the values of α where an event occurs are sometime referred to as *kinks*. In particular remark that since there can not exists more than 2^N different subsets of $\{1, \cdots, N\}$, Lemma 6.1 induces that the number M_{ev} of events related to problem (3.1) necessarily satisfies $M_{ev} \leqslant 2^N$.

In the general case (i.e., even in case of non-uniqueness of the solutions), when α decreases, we can easily define the value α_0 and the underlying index set J_{α_0} at which the first event occurs. Indeed, from Lemma 3.2, we know that for $\alpha \geqslant \max_k [\mathbf{S}\boldsymbol{\omega}]_k/d_k$, we have $\boldsymbol{v}^*_\alpha = 0$. Therefore, the first event occurs at $\alpha_0 = \max_k [\mathbf{S}\boldsymbol{\omega}]_k/d_k$, and we have $J_{\alpha_0} = \{k | [\mathbf{S}\boldsymbol{\omega}]_k/d_k = \alpha_0\}$ (for larger values of α, the underlying sparsity pattern is the empty set). As detailed in the following Section 6.2, if the submatrix $\mathbf{S}_{J_{\alpha_0},J_{\alpha_0}}$ is invertible, we can easily compute the value α_1 at which the next event occurs, and also describe the related sparsity pattern J_{α_1}.

6.2. Regularisation direction for non-singular submatrix. We now discuss, in case of uniqueness of the solutions, the computation of the regularisation path for decreasing values of α, i.e., we asume that the events occurs successively at $\alpha_0 = \max_k [\mathbf{S}\boldsymbol{\omega}]_k/d_k > \alpha_1 > \cdots > \alpha_{M_{ev}-1} \geqslant 0$, the last event corresponding to the largest value of α such that $J_\alpha = \{1, \cdots, N\}$, since for $\alpha = 0$, we have $\boldsymbol{v}^*_\alpha = \boldsymbol{\omega} > 0$.

More precisely, considering a kink α_p with related sparsity pattern J_{α_p} (for $p \in \{0, \cdots, M_{ev} - 2\}$) and assuming that the submatrix $\mathbf{S}_{J_{\alpha_p},J_{\alpha_p}}$ is invertible, we describe how to compute the value $\alpha_{p+1} < \alpha_p$

corresponding to the next event, and how to characterise the related sparsity pattern $J_{\alpha_{p+1}}$. We assume that $n_{\alpha_p} = \mathrm{card}(J_{\alpha_p}) < N$ and, for simplicity, we use the notation $J = J_{\alpha_p}$. We recall that, by definition, α_p is the largest value of α such that $J_\alpha = J$.

From (6.1), we introduce the vector υ_α such that $[\upsilon_\alpha]_{J^c} = 0$ and $[\upsilon_\alpha]_J = \mathbf{S}_{J,J}^{-1}([\mathbf{S}\boldsymbol{\omega}]_J - \alpha \mathbf{d}_J)$; the vector υ_α is sometime referred to as the *regularisation direction*. By definition, α_{p+1} corresponds to the smallest α such that $0 \leqslant \alpha < \alpha_p$ and

$$[\upsilon_\alpha]_J \geqslant 0 \text{ and } [\mathbf{S}(\upsilon_\alpha - \boldsymbol{\omega}) + \alpha \mathbf{d}]_{J^c} \geqslant 0. \tag{6.2}$$

The set $J_{\alpha_{p+1}}$ is then obtained by removing from $J = J_{\alpha_p}$ all the indices $k \in J$ such that $[\upsilon_{\alpha_{p+1}}]_k = 0$, and by adding all the indices $k \in J^c$ such that $[\nabla D_{\alpha_{p+1}}(\upsilon_{\alpha_{p+1}})]_k = 0$; see Remark 6.1 for more details concerning the computation of α_{p+1}.

If $\mathbf{S}_{J_{\alpha_{p+1}}, J_{\alpha_{p+1}}}$ is invertible, we can next compute α_{p+2} and $J_{\alpha_{p+2}}$ in exactly the same way, and we may potentially iterate like this until we reach the last event, or, at least, as far as the encountered principal submatrices are invertible.

Remark 6.1. Consider (6.2) and introduce the $(N - n_{\alpha_p}) \times n_{\alpha_p}$ matrix $\mathbf{M} = \mathbf{S}_{J^c,J}\mathbf{S}_{J,J}^{-1}$. The condition $[\mathbf{S}(\upsilon_{\alpha_p} - \boldsymbol{\omega}) + \alpha \mathbf{d}]_{J^c} \geqslant 0$ yields $\alpha(\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}) \leqslant \mathbf{M}[\mathbf{S}\boldsymbol{\omega}]_J - [\mathbf{S}\boldsymbol{\omega}]_{J^c}$, and we can thus define

$$\alpha_+ = \max_l \left\{ \left[\mathbf{M}[\mathbf{S}\boldsymbol{\omega}]_J - [\mathbf{S}\boldsymbol{\omega}]_{J^c}\right]_l / \left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l \middle| \left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l < 0 \right\}. \tag{6.3}$$

In the same way, the condition $[\upsilon_\alpha]_J \geqslant 0$ gives $\alpha \mathbf{S}_{J,J}^{-1}\mathbf{d}_J \leqslant \mathbf{S}_{J,J}^{-1}[\mathbf{S}\boldsymbol{\omega}]_J$ and we obtain

$$\alpha_- = \max_m \left\{ \left[\mathbf{S}_{J,J}^{-1}[\mathbf{S}\boldsymbol{\omega}]_J\right]_m / \left[\mathbf{S}_{J,J}^{-1}\mathbf{d}_J\right]_m \middle| \left[\mathbf{S}_{J,J}^{-1}\mathbf{d}_J\right]_m < 0 \right\}. \tag{6.4}$$

The next event then occurs at $\alpha_{p+1} = \max\{\alpha_+, \alpha_-\}$. If $\alpha_{p+1} = \alpha_+$, the event consists in the entry of new indices in the sparsity parttern, and if $\alpha_{p+1} = \alpha_-$, some indices go out of the sparsity parttern. $\lhd$

Once α_p and $J = J_{\alpha_p}$ are known, the computation of the next event (i.e., of α_{p+1} and $J_{\alpha_{p+1}}$) requires the calculation of $\mathbf{S}_{J,J}^{-1}\mathbf{d}_J$ and $\mathbf{S}_{J,J}^{-1}[\mathbf{S}\boldsymbol{\omega}]_J$ (i.e., resolution of a linear system). Starting "from scratch" (i.e., without taking into account the computations already performed to obtain the information relative to the kink α_p) and using a direct method (by for instance considering the Cholesky decomposition of the symmetric and positive-definite matrix $\mathbf{S}_{J,J}$), the amount of computations required for this operation scale as $\mathcal{O}(n_{\alpha_p}^3)$. It is however possible to reduce the computational complexity by considering update formulae (by for instance iteratively updating the Cholesky decomposition of $\mathbf{S}_{J,J}$); in the favorable cases, the computational complexity may thus be reduced to $\mathcal{O}(n_{\alpha_p}^2)$ (while still considering a direct approach). Notice that an alternative may also consists in considering an indirect iterative approach, like for example conjugate gradient methods. Finally, the complexity of the two matrix-vector products involving the matrix $\mathbf{S}_{J^c,J}$ scale as $\mathcal{O}(n_{\alpha_p}(N - n_{\alpha_p}))$. The computation of the regularisation direction therefore becomes intractable once large values of n_{α_p} are reached; therefore, when N is large, the regularisation-path strategy may only be used to explore the range of very sparse approximate measure ν.

More generally, the regularisation-path strategy offers an interesting framework to compute solutions to problems (3.1) and (3.2) (or, at least, to obtain very accurate approximations of these solutions). However, the determination of the path is extremely sensitive to numerical errors. For instance, very small value of the gap $\alpha_p - \alpha_{p+1}$ between two consecutive kinks or the simultaneous entry or exit of indices of the sparsity pattern may lead to numerical precision issues.

See Sections 8 and 9 for applications of the regularisation-path strategy.

7. Numerical solver for the constrained problem. In this section, we discuss a strategy to compute approximate solutions to problem (3.2) (i.e., constrained discrepancy minimisation) efficiently and in a numerically tractable way (i.e., without storing the kernel matrix $\mathbf{S}$), for any given value of the parameter $\varkappa \geqslant 0$.

7.1. Vertex-exchange QP solver. Consider problem (3.2); for $\varkappa > 0$, we can define the change of variable $\widetilde{\boldsymbol{\upsilon}} = \mathbf{R}\boldsymbol{\upsilon}$, with $\mathbf{R} = \mathrm{diag}(\mathbf{r})$, and $\mathbf{r} = (r_1, \cdots, r_N)^T = \mathbf{d}/\varkappa$, i.e., $\mathbf{R}$ is a diagonal matrix with i-th diagonal entry r_i. In this way, problem (3.2) is turned into (up to an additive constant), for $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^N$,

$$\underset{\widetilde{\boldsymbol{\upsilon}}}{\text{minimise}}\ C(\widetilde{\boldsymbol{\upsilon}}) = \tfrac{1}{2}\widetilde{\boldsymbol{\upsilon}}^T\mathbf{A}\widetilde{\boldsymbol{\upsilon}} - \mathbf{b}^T\widetilde{\boldsymbol{\upsilon}} \text{ subject to } \widetilde{\boldsymbol{\upsilon}} \geqslant 0 \text{ and } \mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1, \tag{7.1}$$

with $\mathbf{A} = \mathbf{R}^{-1}\mathbf{S}\mathbf{R}^{-1}$ and $\mathbf{b} = \mathbf{R}^{-1}\mathbf{S}\boldsymbol{\omega}$. We refer to (7.1) as the *canonical QP* related to the constrained squared-kernel-discrepancy minimisation (3.2). Since $\mathbf{A}_{i,j} = K^2(x_i, x_j)/(r_i r_j)$, any entry of $\mathbf{A}$ can be easily obtained from only the knowledge of the squared-kernel $K^2(\cdot,\cdot)$, the set S and the vector $\mathbf{r}$. We shall therefore not store the matrix $\mathbf{A}$, but rather compute on the fly any required entry of $\mathbf{A}$ ("kernelised solver"). Notice that since $\boldsymbol{\omega}$ is non-sparse, for large N, the computation of the *distortion term* $\mathbf{S}\boldsymbol{\omega}$ is computationally demanding ($\mathcal{O}(N^2)$ complexity), but it may be parallelised. Once $\mathbf{b}$ is known, the gradient $\nabla C(\widetilde{\boldsymbol{\upsilon}}) = \mathbf{A}\widetilde{\boldsymbol{\upsilon}} - \mathbf{b}$ can be easily obtained for any sparse feasible $\widetilde{\boldsymbol{\upsilon}}$.

The extreme points of the polytope defined by the constraints in (7.1) are the vectors $\{\mathbf{e}_i\}_{i=1}^N$, where $\mathbf{e}_i \in \mathbb{R}^N$ is the i-th element of the canonical basis of $\mathbb{R}^N$ (that is $[\mathbf{e}_i]_i = 1$, all the other components being equal to zero). For a feasible $\widetilde{\boldsymbol{\upsilon}}$, let $J_{\widetilde{\boldsymbol{\upsilon}}} = \{k | \widetilde{\upsilon}_k > 0\}$ be the index set of all strictly positive components of $\widetilde{\boldsymbol{\upsilon}}$. An iteration of the vertex-exchange algorithm consists in searching ($\mathcal{O}(N)$ complexity)

$$i^* = \underset{i}{\mathrm{argmin}}[\nabla C(\widetilde{\boldsymbol{\upsilon}})]_i \text{ and } j^* = \underset{j \in J_{\widetilde{\boldsymbol{\upsilon}}}}{\mathrm{argmax}}[\nabla C(\widetilde{\boldsymbol{\upsilon}})]_j,$$

defining the sparse descent direction $\boldsymbol{\delta} = \mathbf{e}_{i^*} - \mathbf{e}_{j^*}$ (i.e., weight is transfered from the j^*-th to the i^*-th component of $\widetilde{\boldsymbol{\upsilon}}$); in case of non-uniqness of the extrema, an index is simply selected at random among the ones satisfying the condition. The step size is then classically obtained by line search, the optimal step size σ being given by $\sigma = \min\left\{\widetilde{\upsilon}_{j^*}, -\left(\boldsymbol{\delta}^T\nabla C(\widetilde{\boldsymbol{\upsilon}})\right)/(\boldsymbol{\delta}^T\mathbf{A}\boldsymbol{\delta})\right\}$. In particular, since the descent direction $\boldsymbol{\delta}$ is sparse, the computation of the optimal step size is numerically inexpensive, and the same holds for the gradient update. Indeed, we have $\nabla C(\widetilde{\boldsymbol{\upsilon}} + \sigma\boldsymbol{\delta}) = \nabla C(\widetilde{\boldsymbol{\upsilon}}) + \sigma\mathbf{A}\boldsymbol{\delta}$, so that the gradient update only involves two columns of $\mathbf{A}$. The kernelised vertex-exchange strategy thus appears as a interesting candidate to scale up to relatively large N and may be use as a complement of the regularisation-path strategy described in Section 6.

Denoting by $\widetilde{\boldsymbol{\upsilon}}^*$ a solution to (7.1), the convergence of the vertex-exchange algorithm can be easily verified (see, e.g., [7]) by simply remarking that since $\widetilde{\boldsymbol{\upsilon}} \geqslant 0$ and $\mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1$, by definition of j^*, we have $\widetilde{\boldsymbol{\upsilon}}^T\nabla C(\widetilde{\boldsymbol{\upsilon}}) \leqslant \mathbf{e}_{j^*}^T\nabla C(\widetilde{\boldsymbol{\upsilon}})$, so that

$$C(\widetilde{\boldsymbol{\upsilon}}) - C(\widetilde{\boldsymbol{\upsilon}}^*) \leqslant -(\mathbf{e}_{i^*} - \widetilde{\boldsymbol{\upsilon}})^T\nabla C(\widetilde{\boldsymbol{\upsilon}}) \leqslant -(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})^T\nabla C(\widetilde{\boldsymbol{\upsilon}}),$$

and these inequalities can also be used to check distance from optimality. In Sections 8 and 9, the accuracy of an approximate solution $\widetilde{\boldsymbol{\upsilon}}$ is indicated by $\epsilon = (\widetilde{\boldsymbol{\upsilon}} - \mathbf{e}_{i^*})^T\nabla C(\widetilde{\boldsymbol{\upsilon}})$ (Frank-Wolfe error bound).

7.2. Enhancing sparsity through components merging. The canonical QP formulation introduced in Section 7.1 offers a convenient framework to attempt to enhance the sparsity of the approximation while trying to keep the discrepancy of the related measure ν as low as possible. Let $\widetilde{\boldsymbol{\upsilon}} \geqslant 0$ be such that $\mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1$. In practice, $\widetilde{\boldsymbol{\upsilon}}$ will be an exact or approximate solution to problem (7.1), or any vector related to an interesting low-discrepancy configuration $\boldsymbol{\upsilon}$ through the change of variable $\widetilde{\boldsymbol{\upsilon}} = \mathbf{R}\boldsymbol{\upsilon}$, with $\mathbf{R} = \mathrm{diag}(\mathbf{r})$ and $\mathbf{r} = \mathbf{d}/(\mathbf{d}^T\boldsymbol{\upsilon})$, see Section 7.1. We assume that $\widetilde{\boldsymbol{\upsilon}}$ has $n = n_0$ strictly positive components and we introduce $I = \{i | \widetilde{\upsilon}_i > 0\}$. In practice (see Section 8), we observe that it appears possible, to a certain extent, to merge together some components of $\widetilde{\boldsymbol{\upsilon}}$ while inducing a negligible increase of the cost $C(\cdot)$. In what follows, we discuss two simple greedy heuristics based on the sequential merging of pairs of components of $\widetilde{\boldsymbol{\upsilon}}$.

We assume that $n > 1$. For an ordered pair $\{i, j\}$, with i and $j \in I$ and $i \neq j$, we define $\widetilde{\boldsymbol{\upsilon}}_{\{i,j\}} = \widetilde{\boldsymbol{\upsilon}} + \widetilde{\upsilon}_j(\mathbf{e}_i - \mathbf{e}_j)$, i.e., $\widetilde{\boldsymbol{\upsilon}}_{\{i,j\}}$ has $n-1$ strictly positive components, the j-th component of $\widetilde{\boldsymbol{\upsilon}}$ being absorbed by the i-th; we refer to this operation as the $\{i, j\}$-merging of $\widetilde{\boldsymbol{\upsilon}}$. We have

$$C(\widetilde{\boldsymbol{\upsilon}}_{\{i,j\}}) = C(\widetilde{\boldsymbol{\upsilon}}) + \tfrac{1}{2}\widetilde{\upsilon}_j^2(\mathbf{e}_i - \mathbf{e}_j)^T\mathbf{A}(\mathbf{e}_i - \mathbf{e}_j) + \widetilde{\upsilon}_j(\mathbf{e}_i - \mathbf{e}_j)^T\nabla C(\widetilde{\boldsymbol{\upsilon}}).$$

Thus, knowing $\nabla C(\widetilde{\boldsymbol{v}})$, the computation $C(\widetilde{\boldsymbol{v}}_{\{i,j\}})$ is numerically inexpensive (since only four entries of the matrix $\mathbf{A}$ and two entries of $\nabla C(\widetilde{\boldsymbol{v}})$ are involved).

We can then search for the merging associated with the smallest value of $C(\widetilde{\boldsymbol{v}}_{\{i,j\}})$, with i and $j \in I$, and $i \neq j$. Depending on n_0 and on the computational power at disposal, we may either consider

- *strong-pairwise-merging*: search for the best ordred pair $\{i^*, j^*\} = \operatorname{argmin}_{i \neq j} = (\widetilde{\boldsymbol{v}}_{\{i,j\}})$, the amount of computations involved scaling as $\mathcal{O}(n^2)$; or
- *weak-pairwise-merging*: fix $j^* = \operatorname{argmin}_{j \in I} \widetilde{v}_j$, and search for $i^* = \operatorname{argmin}_{i \neq j^*} = (\widetilde{\boldsymbol{v}}_{\{i,j^*\}})$, the amount of computations involved scaling as $\mathcal{O}(n)$.

We thus obtain a "best" pairwise merging $\{i^*, j^*\}$ for $\widetilde{\boldsymbol{v}}$. We next update all the involved objects, i.e., $\widetilde{\boldsymbol{v}} \leftarrow \widetilde{\boldsymbol{v}}_{\{i^*,j^*\}}$, $I \leftarrow I \backslash \{j^*\}$, $n \leftarrow n-1$ and $\nabla C(\widetilde{\boldsymbol{v}}) \leftarrow \nabla C(\widetilde{\boldsymbol{v}}_{\{i^*,j^*\}})$, and we may potentially iterate like this until $n = 1$ (i.e., after $n_0 - 1$ iterations), or at least, until we have reached a satisfactory sparsity-discrepancy trade-off.

We thus obtain a sequence of merged-vectors $\{\widetilde{\boldsymbol{v}}_{[0]}, \widetilde{\boldsymbol{v}}_{[1]}, \cdots, \widetilde{\boldsymbol{v}}_{[n_0-1]}\}$, where $\widetilde{\boldsymbol{v}}_0$ is our initial vector, $\widetilde{\boldsymbol{v}}_{[1]}$ results from the merging of two components of $\widetilde{\boldsymbol{v}}_{[0]}$, etc.; by construction, $\widetilde{\boldsymbol{v}}_{[k]} \geqslant 0$ and $\mathbb{1}^T \widetilde{\boldsymbol{v}}_{[k]} = 1$ for all k, and $\widetilde{\boldsymbol{v}}_{[k]}$ has $n_0 - k$ strictly positive components. Finally, instead of considering the Nyström approximation induced by $\boldsymbol{v} = \mathbf{R}^{-1}\widetilde{\boldsymbol{v}}_0$, we may consider a sparser vector $\boldsymbol{v}_{[k]} = \mathbf{R}^{-1}\widetilde{\boldsymbol{v}}_{[k]}$ (notice that $\boldsymbol{v}_{[k]}$ and $\widetilde{\boldsymbol{v}}_{[k]}$ have the same number of strictly positive components); see Sections 8 and 9 for illustrations.

8. Two-dimensional example. We assume that $S = \{x_k\}_{k=1}^N$ consists of the $N = 2016$ first points of a uniform Halton sequence on $[-1,1]^2$ (see Figure 8.2), and we set $\omega_k = 1/N$ for all k; in particular, the measure $\mu = \sum_k \omega_k \delta_{x_k}$ thus appears as a quadrature approximation of the uniform probability measure on $[-1,1]^2$. We consider the Gaussian kernel $K(x,y) = \exp(-\ell\|x-y\|^2)$, where $\|x-y\|$ is the Euclidean norm on $\mathbb{R}^2$, and we set $\ell = 1/0.16$. An overview of the spectrum of the operator T_μ is given in Figure 8.1. In all Section 8, we consider $\mathbf{d} = \operatorname{diag}(\mathbf{K}) = \mathbb{1}$, i.e., we penalise the trace of T_v; we have $\mathbf{d}^T\boldsymbol{\omega} = \operatorname{trace}(T_\mu) = 1$.

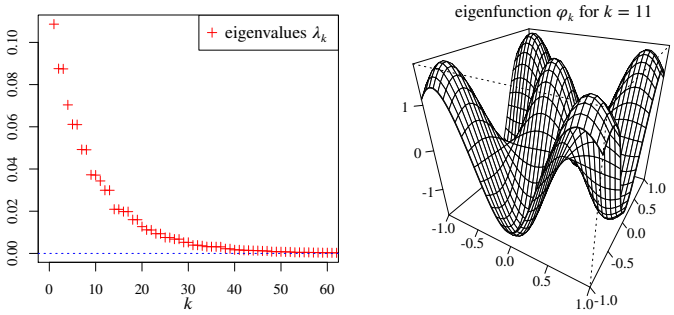

Fig. 8.1: For the two-dimensional example, eigenvalues λ_k of the integral operator T_μ (sorted in decreasing order, only the 62 largest eigenvalues are presented), and graph, on $[-1,1]^2$ of the canonically extended eigenfunction φ_k for $k = 11$.

8.1. First experiment. Figure 8.2 shows the (approximate) solution $\boldsymbol{v}^*$ to problem (3.2) with $\varkappa = 0.81$, or equivalently, to problem (3.1) with $\alpha \approx 8.354215 \times 10^{-3}$. The vector $\boldsymbol{v}^*$ has 160 strictly positive components, and the support of the related measure ν^* inherits an intersting "four-concentric-square" structure. We have $D(\boldsymbol{v}^*) = 7.631890 \times 10^{-4}$ (for comparison, notice that $D(0) = \frac{1}{2}\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\omega} = 2.661452 \times 10^{-2}$ and $D(\mathbf{e}_1) = 4.760566 \times 10^{-1}$, with $\mathbf{e}_1$ the first element of the canonical basis of $\mathbb{R}^N$).

The presented solution has been obtained using the regularisation-path strategy described in Section 6 (see Section 8.2 for more details). Considering the regularisation path for problem (3.1)

with decreasing values of α, the underlying value of $\alpha \approx 8.354215 \times 10^{-3}$ satisfies

$$\alpha_{p+1} = 8.352970 \times 10^{-3} \leqslant \alpha \leqslant \alpha_p = 8.355244 \times 10^{-3}, \text{ with } p = 4\,047.$$

Correspondingly, considering the regularisation path for problem (3.2) with increasing values of $\varkappa$, the underlying value $\varkappa = 0.81$ satisfies

$$\varkappa_p = 0.8099788 \leqslant \varkappa \leqslant \varkappa_{p+1} = 0.8100256, \text{ with } p = 4\,047.$$

In the framework of Section 7.1, the presented solution is related to a Frank-Wolfe error bound $\epsilon = 4.510281 \times 10^{-17}$.

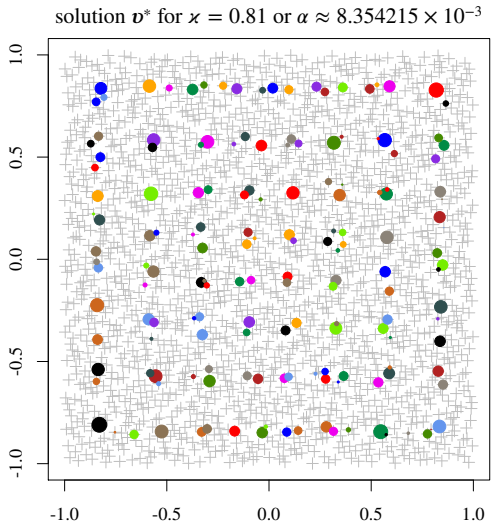

Fig. 8.2: Graphical representation (two-dimensional example) of the solution $\boldsymbol{\upsilon}^*$ to problem (3.2) with $\varkappa = 0.81$, or equivalently, to problem (3.1) with $\alpha \approx 8.354215 \times 10^{-3}$ (for $\mathbf{d} = \text{diag}(\mathbf{K})$ and $\boldsymbol{\omega} = \mathbb{1}/N$). The grey crosses represent the points in S and the filled dots are the strictly positive components of $\boldsymbol{\upsilon}^*$ (surface being proportional to υ_k^*).

The efficiency of the Nyström approximation induced by the solution $\boldsymbol{\upsilon}^*$ presented in Figure 8.2 is illustrated in Figure 8.3. In view of the values of Υ_k (see Section 4), we observe that the 21 main eigendirections of the operator T_{ν^*} leads to remarkably accurate approximations of the eigenpairs of T_μ related to the largest eigenvalues λ_k (for $k \in \{1, \cdots, 21\}$, we indeed have $0.9876064 \leqslant \Upsilon_k \leqslant 0.9999785$). The quality of the approximate eigenpairs $(\widehat{\lambda}_k, \widehat{\varphi}_k)$ starts to decrease for $k \geqslant 22$. For $k \in \{22, \cdots, 31\}$ the approximate eigenpairs are still relatively accurate; the approximation efficiency decreases significantly for $k \in \{32, \cdots, 43\}$ and becomes very poor for $k \geqslant 44$.

8.2. Regularisation path. We now investigate the impact of the parameters α and $\varkappa$ related to problems (3.1) and (3.2) respectively. We compute the 12 786 first events of the regularisation-path related to problem (3.1) with decreasing values of α (see Section 6), i.e., until we reach a precision issue; in particular, we have $\alpha_0 = 6.310163 \times 10^{-2}$ and $\alpha_{12785} = 1.514626 \times 10^{-5}$. Correspondingly, for the regularisation path related to problem (3.1) with increasing $\varkappa$, we have $\varkappa_0 = 0$ and $\varkappa_{12785} = 0.9995426$ (we recall that $\mathbf{d}^T \boldsymbol{\omega} = 1$).

Figure 8.4 shows that the number of strictly positive components of the solution $\boldsymbol{\upsilon}_\varkappa^*$ to problem (3.1) tends to increase when $\varkappa$ increases. As expected from Lemma 3.2-(vi), the functions $\varkappa \mapsto D(\boldsymbol{\upsilon}_\varkappa^*)$ is decreasing; in the same way, when $\varkappa$ increases, the corresponding value of the regularisation parameter α decreases (see Lemmas 3.2 and 3.3).

For 41 values of $\varkappa$ evenly spread between 0 and $\varkappa_{12785}$, Figure 8.5 shows the evolution of the Upsilon-test (4.2) related to the Nyström approximation induced by the various solutions $\boldsymbol{\upsilon}_\varkappa^*$. As

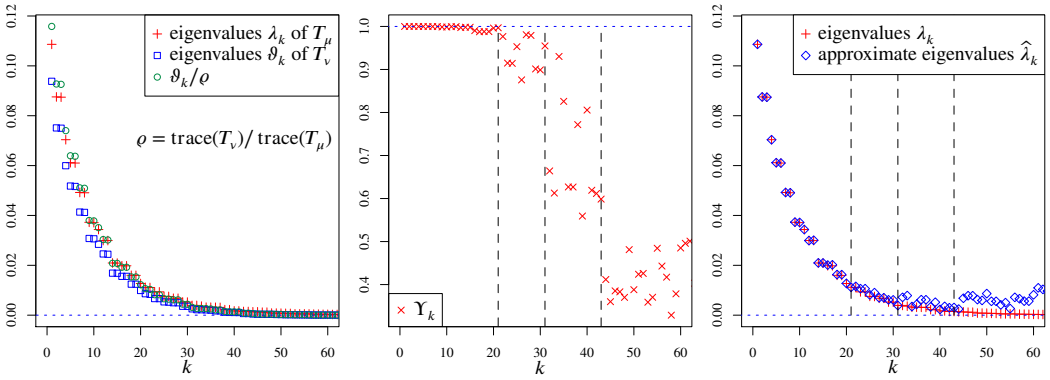

Fig. 8.3: Eigenpairs approximation induced by the solution $\boldsymbol{v}^*$ presented in Figure 8.2: graphical representations of the 62 largest eigenvalues λ_k, ϑ_k and rescaled eigenvalue ϑ_k/ϱ, sorted in decreasing ordered (left); Upsilon-test (4.2) for the 62 main eigendirections of T_{ν^*} (middle); and 62 first approximate eigenvalues $\widehat{\lambda}_k$ induced by $\boldsymbol{v}^*$ (right).

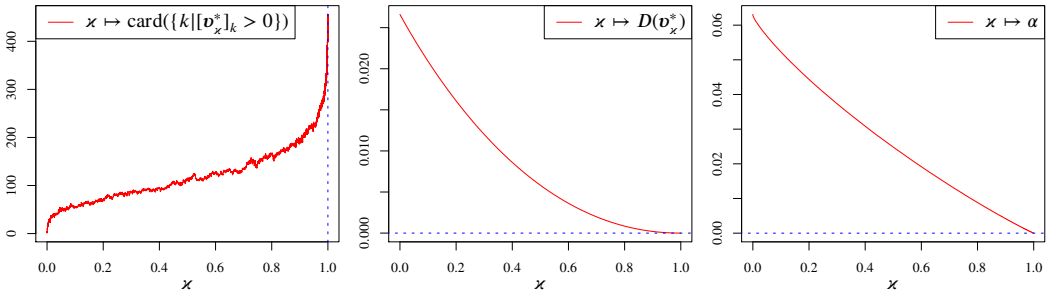

Fig. 8.4: Graphical representation of the 12 786 first events of the regularisation path related to problem (3.1) for increasing $\varkappa$ (two-dimensional example): number of strictly positive components in $\boldsymbol{v}^*_\varkappa$ as function of $\varkappa$ (left); graph of $\varkappa \mapsto D(\boldsymbol{v}^*_\varkappa)$ (middle), and relation between $\varkappa$ and the regularisation parameter α of problem (3.1) (see Lemma 3.3).

expected, the number of accurately approximate eigendirections increases with $\varkappa$. Remarkably, for each of the considered values of $\varkappa$, the number of eigendirections approximated with a high accuracy appears to be in close relation with the decay of the spectrum of T_μ; we recall that, since $\mathbf{d} = \mathrm{diag}(\mathbf{K})$, we have $\mathrm{trace}(T_{\nu^*_\varkappa}) = \varkappa$.

8.3. Components merging. We now perform the strong-pairwise-merging (see Section 7.2) of the solution $\boldsymbol{v}^*$ presented in Figure 8.2 (i.e., problem (3.2) with $\varkappa = 0.81$). As illustrated in Figure 8.6, for the first merging iterations, $D(\boldsymbol{v}_{[k]})$ stays very close to $D(\boldsymbol{v}^*) = 7.631890 \times 10^{-4}$. After 90 iterations, we have $D(\boldsymbol{v}_{[90]}) - D(\boldsymbol{v}^*) = 3.494809 \times 10^{-5}$ (i.e., increasing of 4.58%), and $\boldsymbol{v}_{[90]}$ is supported by 70 points (instead of 160 for $\boldsymbol{v}^*$); a graphical representation of $\boldsymbol{v}_{[90]}$ is given in the left-hand part of the figure. The efficiency of the Nyström approximation induced by $\boldsymbol{v}_{[90]}$ is presented in the right-hand part of Figure 8.6 (Upsilon-test). We observe that although being slightly less accurate than the approximate eigendecomposition induced by $\boldsymbol{v}^*$, the approximation induced by $\boldsymbol{v}_{[90]}$ remains very satisfactory while being related to a vector more than two times sparser.

8.4. Comparison with random sampling. For comparison purpose, we compute the approximate eigendecompositions induced by random uniform samples (without repetition) of size $n_{rand} = 300$, 600, 900 and 1200 (i.e., we randomly select n_{rand} distinct points among the $N = 2016$ points in $\mathcal{S}$, and we consider the uniform probability measure supported by the points selected); for each sample

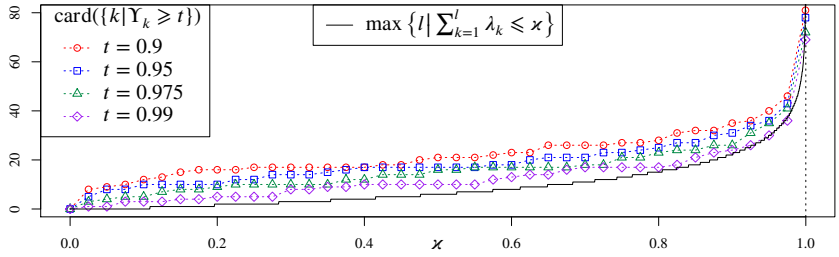

Fig. 8.5: Evolution of the accuracy of the approximate eigendecomposition of T_μ induced by $\upsilon^*_\varkappa$ for 41 values of $\varkappa$ between $\varkappa_0 = 0$ and $\varkappa_{12785} = 0.9995426$; the accuracy of the approximate eigendirections is measured trough the Upsilon-test, and for illustration purpose, the map $\varkappa \mapsto \max\{l | \sum_{k=1}^{l} \lambda_k \leqslant \varkappa\}$ is also presented (two-dimensional example).

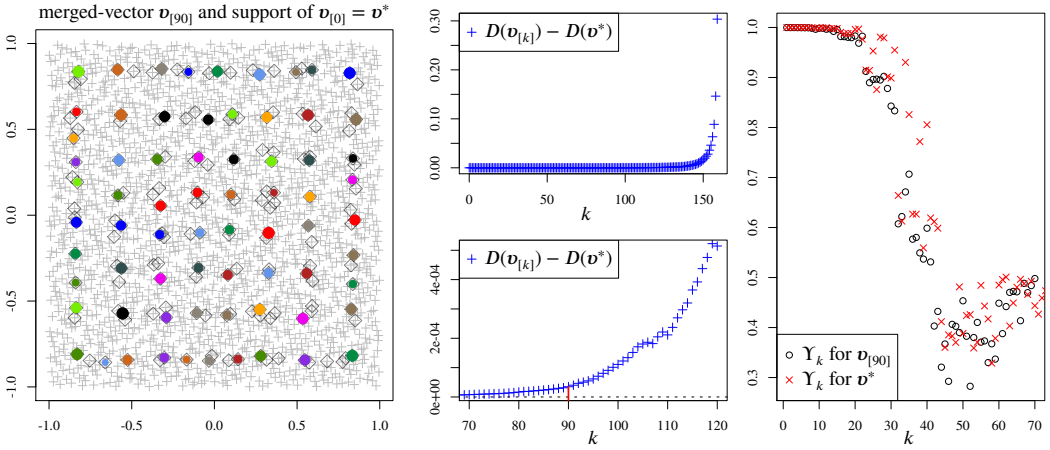

Fig. 8.6: Graphical representation of the merged-solution $\upsilon_{[90]}$ (two-dimensional example) obtained after 90 iterations of the strong-pairwise-merging strategy applied to the solution υ^* presented in Figure 8.2; the grey diamonds indicate the support of υ^* (left). Increase of the cost $D(\cdot)$ induced by each merging iteration, for the whole 159 iterations (top-middle), and zoom around the 90-th iteration (bottom-middle). Representation of the Upsilon-test obtained from the merged-vector $\upsilon_{[90]}$ and comparison with the Upsilon-test related to the solution υ^* (right).

size, we perform 100 repetitions. Figure 8.7 illustrates the accuracy of the resulting approximate eigendirections (measured through the Upsilon-test). As we could expect, the accuracy of the approximations increases with the size of the sample; however, in terms of trade-off between sparsity and accuracy, the results are far behind the one obtained using squared-kernel-discrepancy minimisation.

9. Application to medium/large-scale problems. The main motivation behind this section is to illustrate the ability of the proposed framework to tackle relatively large-scale problems. The datasets have been downloaded from the UCI Machine Learning Repository, see [12]. All the computations have been performed on a 2015 desktop endowed with an Intel Core i7-4790 processor with 16 Gb of RAM; the various methods have been entirely encoded in C (the code is available upon request).

9.1. MiniBooNE dataset. We consider the standardised entries of the MiniBooNE dataset, without labels; $\mathcal{S}$ thus consists of $N = 129\,596$ points in $\mathbb{R}^{50}$. We use a Gaussian kernel (same expression as in Section 8) with $\ell = 0.02$, and we set $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathrm{diag}(\mathbf{K}) = \mathbb{1}$ (in particular, $\mathbf{d}^T\boldsymbol{\omega} = 1$). Notice that $\ell = 0.02$ belongs to the range of "good parameters" for the SVM binary classification of this dataset.

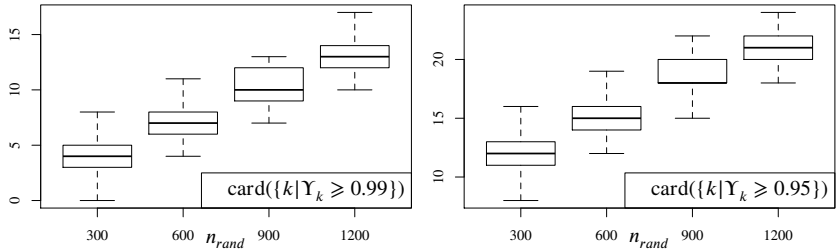

Fig. 8.7: For the two-dimensional example, accuracy of the approximate eigendecompositions induced by random samples of size n_{rand}; for each values of n_{rand}, boxplot (minimum, first quartile, median, third quartile and maximum), over 100 repetitions, of the number of approximate eigendirections such that $\Upsilon_k \geqslant 0.99$ (left) and $\Upsilon_k \geqslant 0.95$ (right).

We compute the 3 000 first events of the regularisation path related to problems (3.1) and (3.2). We have $\alpha_0 = 0.2188961$ and $\alpha_{2999} = 3.546703 \times 10^{-3}$, and correspondingly $\varkappa_0 = 0$ and $\varkappa_{2999} = 0.655808$. A graphical representation of the obtained results is proposed in Figure 9.1. We can observe that for $\varkappa \geqslant 0.5$, the number of strictly positive components of the solution $\boldsymbol{\upsilon}^*_\varkappa$ increases with a signifiant rate; as a consequence, the computation of the regularisation path quickly becomes numerically intractable (notice that the calculation of the 3 000 first events of the regularisation path has required around 4 hours on our aforementioned 2015 desktop).

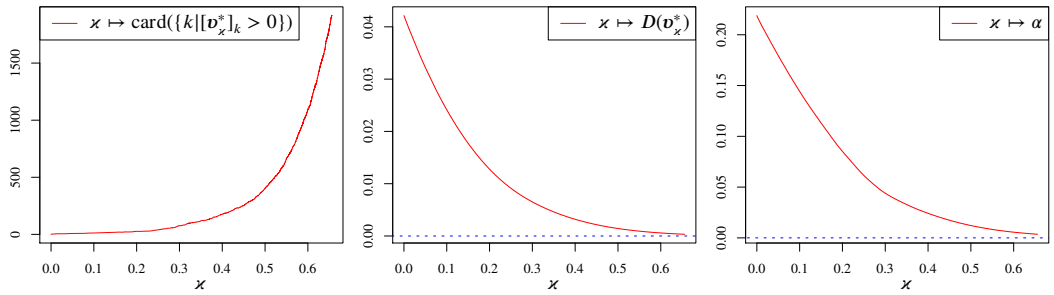

Fig. 9.1: For the MiniBooNE dataset, graphical representation of the 3 000 first events of the regularisation path related to problem (3.1) for increasing $\varkappa$: number of strictly positive components in $\boldsymbol{\upsilon}^*_\varkappa$ as function of $\varkappa$ (left); graph of $\varkappa \mapsto D(\boldsymbol{\upsilon}^*_\varkappa)$ (middle), and relation between $\varkappa$ and the regularisation parameter α of problem (3.1) (see Lemma 3.3).

We then compute the approximate eigendecompositions induced by the solutions to problem (3.1) for $\varkappa = 0.3$ and $\varkappa = 0.655$ obtained with the regularisation-path strategy (or equivalently, to problem (3.1) with $\alpha \approx 4.400276 \times 10^{-2}$ and $\alpha \approx 3.571413 \times 10^{-3}$, respectively). The underlying solutions have 76 and 1 902 strictly positive components, respectively. The efficiency of the induced approximate eigendecompositions is illustrated in Figure 9.2. For $\varkappa = 0.3$, we obtain an accurate approximation of the three main eigenpairs of T_μ while considering only 76 points (we recall that $N = 129\,596$); the approximation of the other eigendirections is relatively poor. For $\varkappa = 0.655$, the eight main eigendirections of T_μ are approximate with high accuracy, and the approximation becomes poor for $k \geqslant 30$. We in particular observe a relatively fast decay of the spectrum.

To explore the type of solutions obtained for larger values of $\varkappa$, we consider the vertex-exchange strategy described in Section 7.1. We compute an approximate solution for $\varkappa = 0.8$; the vertex-exchange algorithm is initialised at $\widetilde{\boldsymbol{\upsilon}} = \mathbf{e}_1$ and after 300 000 iterations, we obtain a Frank-Wolfe error bound of $\epsilon = 1.692408 \times 10^{-8}$. The obtained approximate solution $\widehat{\boldsymbol{\upsilon}}^*$ to problem (3.1) verifies $D(\widehat{\boldsymbol{\upsilon}}^*) = 4.934072 \times 10^{-5}$ and has 9544 strictly positive components. The efficiency of the

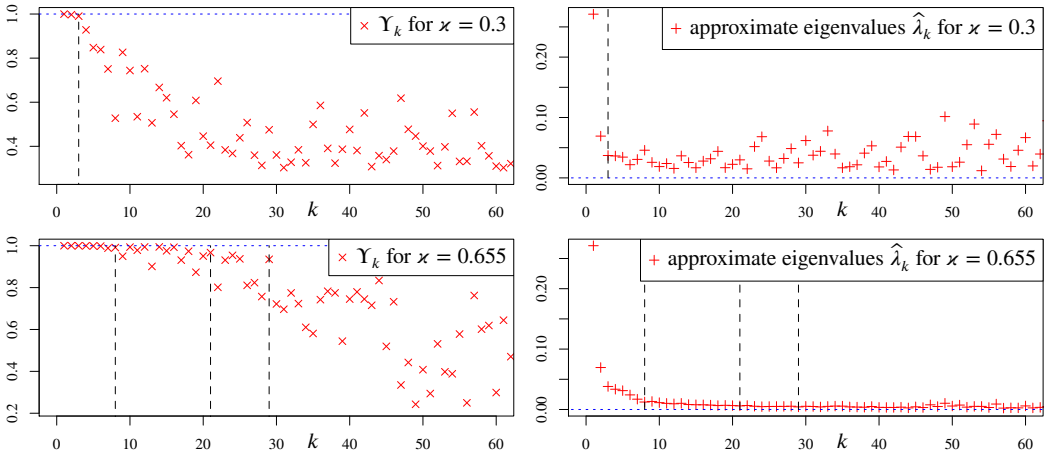

Fig. 9.2: For the MiniBooNE dataset, eigenpairs approximation induced by the solution to problem (3.1) with $\varkappa = 0.3$ (top) and $\varkappa = 0.655$ (bottom): Upsilon-test (4.2) for the 62 main eigendirections of the approximate operator T_{ν^*} (left); and 62 first approximate eigenvalues $\widehat{\lambda}_k$ induced by ν^* (right).

approximate eigendecomposition induced by $\widehat{\nu}^*$ is illustrated in Figure 9.3; we observe that the 21 main eigendirections of T_μ are approximated with high accuracy, and that the 43 main eigendirections are relatively well approximated.

In order to enhance the sparsity of the sampling, we perform a weak-pairwise merging of the solution $\widehat{\nu}^*$ (see Section 7.2); notice that performing a strong-pairwise merging is in this case numerically prohibitive. After 5044 iterations, the merged solution $\nu_{[5044]}$ is supported by 4500 points and $D(\nu_{[5044]}) = D(\widehat{\nu}^*) + 1.061787 \times 10^{-6}$ (i.e., increasing of 2.15%). We next compute the approximate eigendecomposition induced by $\nu_{[5044]}$, the result is presented in Figure 9.3. We observe that, in terms of accuracy, the approximation of the main eigendirections of T_μ induced by $\nu_{[5044]}$ is equivalent to the approximation induces by $\widehat{\nu}^*$ (and, remarkably, is even substantially better), while being more than two times sparser.

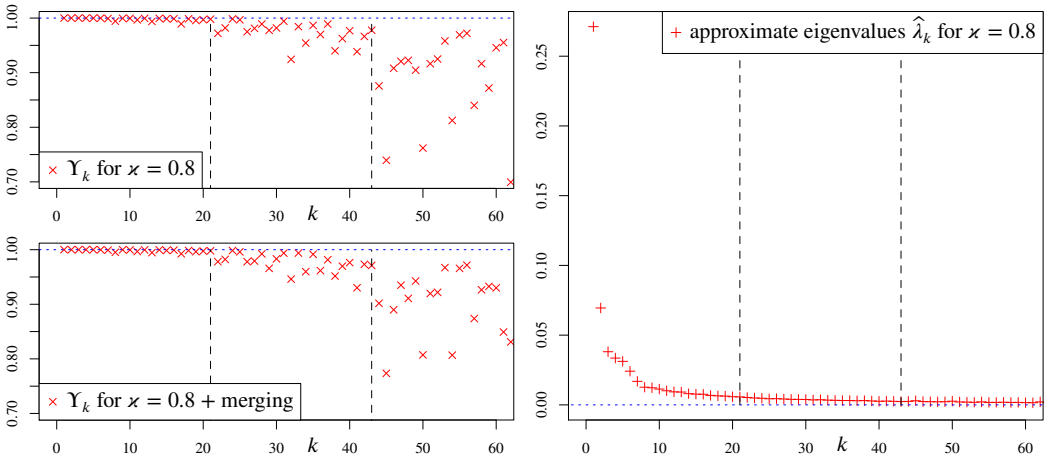

Fig. 9.3: For the MiniBooNE dataset, accuracy of the approximate eigendecompositions (Upsilon-test) induced by the solution $\widehat{\nu}^*$ to problem (3.1) with $\varkappa = 0.8$ obtained from the vertex-exchange algorithm (top-left) and from the merged solution $\nu_{[5044]}$ (bottom-left); graphical representation of the 62 first approximate eigenvalues $\widehat{\lambda}_k$ induced by $\widehat{\nu}^*$ (right).

To conclude this example, for the various approximations considered in Section 9.1, Table 9.1 presents the values taken by $\widehat{\tau}_m = \sum_{k=1}^{m} \widehat{\lambda}_k$, where, for each approximations, m is the number of accurately approximated eigendirections highlighted in Figures 9.2 and 9.3. We can in particular observe that the values of $\widehat{\tau}_m$ obtained from the various approximations are in agreement.

Table 9.1: For the various approximate eigendecompositions considered in Section 9.1 (MiniBooNE dataset), value of $\sum_{k=1}^{m} \widehat{\lambda}_k$ for specific values of m.

	$m = 3$	$m = 8$	$m = 21$	$m = 43$
$\varkappa = 0.3$	0.3773318			
$\varkappa = 0.655$	0.3787624	0.4971454	0.6085483	
$\widehat{\boldsymbol{\upsilon}}^*$ ($\varkappa = 0.8$)	0.3787608	0.4972059	0.6040054	0.6831240
$\boldsymbol{\upsilon}_{[5044]}$ ($\varkappa = 0.8$)	0.3787608	0.4972103	0.6039483	0.6829062

9.2. Test subsample of the SUSY dataset. We consider the standardised entries of the test subsample of the SUSY dataset (without labels), so that $\mathcal{S}$ consists of $N = 500\,000$ points in $\mathbb{R}^{18}$. We still use a Gaussian kernel (same expression as in Section 8) with $\ell = 0.4$, and we set $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathrm{diag}(\mathbf{K}) = \mathbb{1}$ (so that $\mathbf{d}^T\boldsymbol{\omega} = 1$). The computation of the distortion term $\mathbf{S}\boldsymbol{\omega}$ takes 5 665.6 seconds.

We compute an approximate solution for the constrained problem (3.2) with $\varkappa = 0.3$; we perform four consecutive batches of 50 000 iterations each, the solver being initialised at $\widetilde{\boldsymbol{\upsilon}} = \mathbf{e}_1$. After 200 000 iterations (i.e., at the end of the 4-th batch), the obtained approximate solution $\widehat{\boldsymbol{\upsilon}}^*$ verifies $D(\widehat{\boldsymbol{\upsilon}}^*) = 3.931629\times10^{-5}$ and has $n = 20\,664$ strictly positive components. Execution time, evolutions of the Frank-Wolfe error bound ϵ and of the sparsity of the approximate solution are reported in Table 9.2. We observe that a batch of 50 000 iterations of the vertex-exchange algorithm takes around 19 minutes, and that the approximate solution obtained at the end of the first batch is already relatively accurate.

Table 9.2: For the test subsample of the SUSY dataset, information relative to approximate solution to problem (3.2) with $\varkappa = 0.3$ returned by the vertex-exchange algorithm for four consecutive batches of 50 000 iterations, the solver being initialised at $\widetilde{\boldsymbol{\upsilon}} = \mathbf{e}_1$: execution time, total number of iterations, Frank-Wolfe error bound ϵ and number n of strictly positive component of the approximate solution.

	batch 1	batch 2	batch 3	batch 4
time (in sec.)	1 148.7	1 158.3	1 158.5	1 159.1
total nb. of it.	50 000	100 000	150 000	200 000
ϵ	3.1413×10^{-7}	6.5477×10^{-8}	2.7049×10^{-8}	7.0928×10^{-9}
n	19 721	20 619	20 693	20 674

In order to enhance the sparsity of the sampling, we perform a weak-pairwise merging of the approximate solution $\widehat{\boldsymbol{\upsilon}}^*$; the computation of 20 673 merging iterations takes 78.86 seconds. The merged solution $\boldsymbol{\upsilon}_{[13674]}$ is supported by 7 000 points and $D(\boldsymbol{\upsilon}_{[13674]}) = D(\widehat{\boldsymbol{\upsilon}}^*) + 5.271960 \times 10^{-7}$ (i.e., increasing of only 1.34%). We next study the approximate eigendecomposition induced by $\boldsymbol{\upsilon}_{[13674]}$. Computing the 300 first normalised approximate eigenvectors of $\mathbf{KW}$ induced by $\boldsymbol{\upsilon}_{[13674]}$ (see Section 4) takes 3 313.6 seconds (time for canonical extension and rescaling); then, performing the Upsilon-test for this 300 approximate eigendirections takes 191 622.3 seconds (i.e., around 53 hours). The results are presented in Figure 9.4. As already observed, the accuracy of the approximate eigendirections $\widehat{\varphi}_k$ decreases when k increases (we recall that the eigenvalues of the approximate operator are stored in descending order). The obtained approximate eigendirections are remarkably

accurate: for instance, we have $\Upsilon_k \geqslant 0.99$ for 64 eigendirections, $\Upsilon_k \geqslant 0.95$ for 224 eigendirections, and $\Upsilon_k \geqslant 0.9$ for 299 eigendirections (among 300), and $\min_{k=1}^{300} \Upsilon_k = 0.8826657$. The approximate eigenvalues $\widehat{\lambda}_k$ and the related eigenvalues ϑ_k of the operator T_ν defined from $\boldsymbol{\upsilon}_{[13674]}$ are also presented.

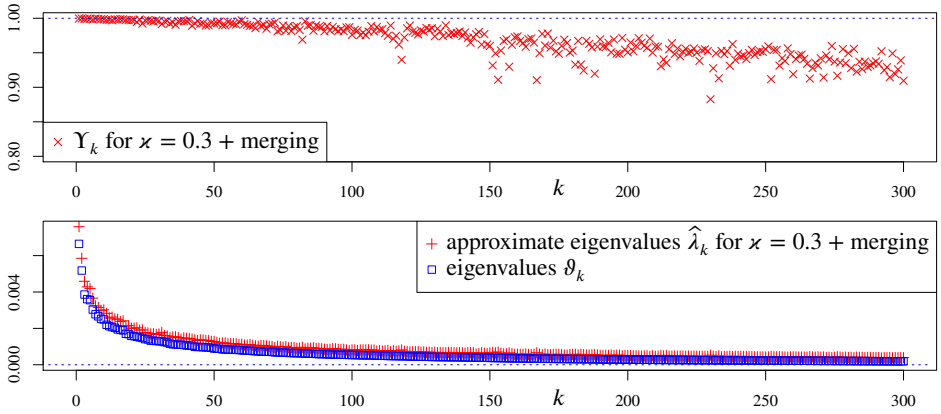

Fig. 9.4: For the test subsample of the SUSY dataset, accuracy of the 300 main approximate eigendirections (Upsilon-test) induced by the merged vector $\boldsymbol{\upsilon}_{[13674]}$ obtained from the approximate solution $\widehat{\boldsymbol{\upsilon}}^*$ to problem (3.2) with $\varkappa = 0.3$ (top); graphical representation of the underlying approximate eigenvalues $\widehat{\lambda}_k$ and of the corresponding eigenvalues ϑ_k of the operator related to $\boldsymbol{\upsilon}_{[13674]}$ (bottom).

10. Conclusion. We proposed a QP-based strategy for computing sparse Nyström approximations of integral operators defined from a symmetric positive-semidefinite kernel and a discrete measure with finite support. Our criterion is based on the notion of squared-kernel discrepancy (i.e., on the squared Hilbert-Schmidt norm for operators defined on the RKHS related to the underlying kernel). For a fixed kernel and starting from a given discrete measure μ, we compute an approximate measure ν which is as close as possible to μ in terms of squared-kernel discrepancy while being supported by a smaller number of points, the support of ν being included in the support of μ (quadrature-sparsification). From a methodological point of view, the obtained approximation scheme is deterministic, numerically scalable (i.e., large-scale problem can be tackled) and enjoys an optimality property.

In this work, the sparsity of ν is enforced through a weighed-ℓ_1-type penalisation. We more particularly focused on penalising the trace of the approximate operator T_ν; however, other type of penalisations may be considered and the study of "optimal" penalisation strategies may deserve further investigations: relations with the spectrum of the initial operator T_μ, impact on the sparsity-accuracy trade-off of the induced approximation, etc. The QPs related to the regularised or constrained squared-kernel-discrepancy minimisations problem can also be related to one-class SVMs with respect to the squared-kernel and the initial measure μ; a deeper analysis of this connection may potentially lead to interesting developments.

We investigated two numerical strategies for the resolutions of the regularised and constrained problems. The regularisation-path approach can be used to explore the range of very sparse solutions, with the interest of leading to exact solutions (up to precision errors); the vertex-exchange-based strategy permits the exploration of a wider range of solutions and offer a numerically efficient approach to build approximate solutions. We also proposed two greedy heuristics to enforce the sparsity of a measure obtained from squared-kernel-discrepancy minimisation; the strong and weak-pairwise merging algorithms appear as relatively efficient and very useful in practice. The study of efficient numerical strategies for solving problems related to squared-kernel-discrepancy minimisations is however still obviously largely open and is of great importance for applications.

As illustrated in Section 9, the proposed approach can be used to tackle relatively large-scale

problems while leading to accurate sparse approximations of the main eigenpairs of the considered operator T_μ in a reasonable amount of time (and without necessarily resorting to very powerful computing hardwares). In practice, the most computationally demanding steps are the computation of the distortion term $\mathbf{S}\boldsymbol{\omega}$ and of the Upsilon-test; the computation of $\mathbf{S}\boldsymbol{\omega}$ is mandatory and can not be avoided, while the Upsilon-test is facultative, although very useful (notice that performing the Upsilon-test for a large number of approximate eigendirections is significantly more demanding than the computation of the distortion term). When dealing with sparse measures, these operations are indeed the only ones with computational complexity scaling as $\mathcal{O}(N^2)$; however, as already mentioned, they can be easily parallelised since they consist in simple kernelised matrix-vector products.

In this work, we have mainly focused on the quadrature-sparsification problem (i.e., the initial measure μ is discrete, and the support of ν is included in the support of μ); however, as shown in Section 2, the notion of squared-kernel discrepancy appears as a natural tool in the general framework of Nyström approximation of integral operators defined from symmetric positive-semidefinite kernels. From a theoretical point of view, the study of "squared-kernel-discrepancy-optimal" quadrature for Nyström approximation of operators related to general measures μ may lead to interesting developments; for instance, the characterisation of discreteness-inducing penalisations (i.e., penalisations leading to discrete optimal measures ν^* with finite support) appears of great interest. More generally, the study of kernel-discrepancy-based strategies (not necessarily related to squared-kernels, see for instance [3]) for the optimal sparsification of other types of kernel-based models (like for instance SVM-classification models) may potentially be a fruitful research direction.

Acknowledgements. 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). 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: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T, PhD/Postdoc grants – Flemish Government: FWO: PhD/Postdoc grants, projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); IWT: PhD/Postdoc grants, projects: SBO POM (100031); iMinds Medical Information Technologies SBO 2014 – Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimisation, 2012-2017).

Appendix A. Proofs.

Proof of Lemma 2.1. Consider an o.n.b. $\{h_k | k \in \mathbb{I}\}$ of $\mathcal{H}$. By definition of T_μ and T_ν, for all $k \in \mathbb{I}$, we have

$$\begin{aligned}\big(T_\mu[h_k]\big|T_\nu[h_k]\big)_{\mathcal{H}} &= \big(h_k\big|T_\nu[h_k]\big)_{L^2(\mu)} = \big(T_\mu[h_k]\big|h_k\big)_{L^2(\nu)} \\ &= \textstyle\int_{\mathscr{X}\times\mathscr{X}} K(x,t)h_k(x)h_k(t)\mathrm{d}\mu(x)\mathrm{d}\nu(t).\end{aligned}$$

For x and $t \in \mathscr{X}$, we have $K(x,t) = \sum_{k\in\mathbb{I}} h_k(x)h_k(t)$, and we thus obtain

$$\big(T_\mu\big|T_\nu\big)_{\mathrm{HS},\mathcal{H}} = \textstyle\int_{\mathscr{X}\times\mathscr{X}} \big(K(x,t)\big)^2\mathrm{d}\mu(x)\mathrm{d}\nu(t) = \|K\|^2_{L^2(\mu\otimes\nu)}.$$

From Section 2.1, we also have $\|K\|^2_{L^2(\mu\otimes\mu)} = \|K_\mu\|^2_{L^2(\mu\otimes\mu)} = \sum_{k\in\mathbb{I}_\mu^+} \lambda_k^2$. □

Proof of Theorem 2.1. The proof directly follows from the properties discussed in Sections 2.1 and 2.3. In particular, (2.1) is obtained by considering the o.n.b. $\{\sqrt{\lambda_k}\varphi_k | k \in \mathbb{I}_\mu^+\}$ of $\mathcal{H}_\mu$ defined by T_μ while remarking that $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ implies $T_\mu[h] = T_\nu[h] = 0$ for all $h \in \mathcal{H}_{0,\mu}$. By expanding $T_\nu[\varphi_k]$ thanks to $P_{\mathcal{H}_\nu}[\varphi_k] = \sum_{l\in\mathbb{I}_\nu^+} \vartheta_l\psi_l(\psi_l|\varphi_k)_{\mathcal{H}}$ (orthogonal projection, in $\mathcal{H}$, of φ_k onto $\mathcal{H}_\nu$), we obtain (2.2). Finally, the last inequality directly follows from $\|h\|^2_{L^2(\mu)} \leqslant \tau_\mu\|h\|^2_{\mathcal{H}}$ for all $h \in \mathcal{H}$. □

Proof of Lemma 3.1. If $\boldsymbol{\omega} = \mathbb{1}$ and the components of $\boldsymbol{\upsilon}$ are all 0 or 1, then the components of $\boldsymbol{\omega} - \boldsymbol{\upsilon}$ are also all 0 or 1. Introducing the index set $I = \{k|\upsilon_k = 0\} = \{k|\omega_k - \upsilon_k = 1\}$, we obtain $(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) = \sum_{(i,j)\in I\times I} K^2(x_i, x_j)$, which is equal to $\|\mathbf{K} - \mathbf{V}\mathbf{K}\mathbf{V}\|_F^2$. □

Proof of Lemma 3.2. Assertion (i) follows directly from $D_{K^2}(\mu,\mu)=0$ (since $D_{K^2}(\mu,\nu)\geqslant 0$). From the first order optimality condition, for $\alpha\geqslant 0$, a feasible $\boldsymbol{v}_\alpha^*$ is solution to (3.1) if and only if, for any feasible $\boldsymbol{v}$, we have $(\boldsymbol{v}-\boldsymbol{v}_\alpha^*)^T\nabla D_\alpha(\boldsymbol{v}_\alpha^*)\geqslant 0$. Then, considering $\boldsymbol{v}_\alpha^*=0$ leads to assertion (ii), and since $\boldsymbol{\omega}$ is by assumption feasible for (3.1), assertion (iii) is obtained by taking $\boldsymbol{v}=\boldsymbol{\omega}$. To prove (iv), we first remark that for all $\theta\in[0,1]$, we have

$$D_\alpha(\theta\boldsymbol{v}_\alpha^*+(1-\theta)\widetilde{\boldsymbol{v}}_\alpha^*)=D_\alpha(\widetilde{\boldsymbol{v}}_\alpha^*)+\theta(\boldsymbol{v}_\alpha^*-\widetilde{\boldsymbol{v}}_\alpha^*)^T\nabla D_\alpha(\widetilde{\boldsymbol{v}}_\alpha^*)+\theta^2\tfrac{1}{2}(\boldsymbol{v}_\alpha^*-\widetilde{\boldsymbol{v}}_\alpha^*)^T\mathbf{S}(\boldsymbol{v}_\alpha^*-\widetilde{\boldsymbol{v}}_\alpha^*).$$

Since $D_\alpha(\widetilde{\boldsymbol{v}}_\alpha^*)=D_\alpha(\theta\boldsymbol{v}_\alpha^*+(1-\theta)\widetilde{\boldsymbol{v}}_\alpha^*)$, we necessarily have $(\boldsymbol{v}_\alpha^*-\widetilde{\boldsymbol{v}}_\alpha^*)^T\mathbf{S}(\boldsymbol{v}_\alpha^*-\widetilde{\boldsymbol{v}}_\alpha^*)=0$. From the Cauchy-Schwarz inequality for the bilinear form related to $\mathbf{S}$, we therefore have $(\boldsymbol{v}_\alpha^*-\widetilde{\boldsymbol{v}}_\alpha^*)^T\mathbf{S}(\widetilde{\boldsymbol{v}}_\alpha^*-\boldsymbol{\omega})=0$, and this completes the proof since we also have $(\boldsymbol{v}_\alpha^*-\widetilde{\boldsymbol{v}}_\alpha^*)^T\nabla D_\alpha(\widetilde{\boldsymbol{v}}_\alpha^*)=0$. Assertion (v) follows from the expansion, for all $\boldsymbol{v}\geqslant 0$,

$$D_\alpha(\boldsymbol{v})=D_\alpha(\boldsymbol{v}-v_k\mathbf{e}_k)+[\alpha\mathbf{d}-\mathbf{S}\boldsymbol{\omega}]_k v_k+\sum_{i\neq k}v_iv_k\mathbf{S}_{i,k}+\tfrac{1}{2}v_k^2\mathbf{S}_{k,k},$$

where $\mathbf{e}_k$ stands for the k-th element of the canonical basis of $\mathbb{R}^N$ (so that $[\boldsymbol{v}-v_k\mathbf{e}_k]_k=0$). Since all the entries of $\mathbf{S}$ are non-negative, if $[\alpha\mathbf{d}-\mathbf{S}\boldsymbol{\omega}]_k\geqslant 0$ and $v_k>0$, since $S_{k,k}>0$, we therefore have $D(\boldsymbol{v}-v_k\mathbf{e}_k)<D_\alpha(\boldsymbol{v})$. To obtain assertion (vi), consider $\alpha_1<\alpha_2$ and let $\boldsymbol{v}_{\alpha_1}^*$ and $\boldsymbol{v}_{\alpha_2}^*$ be solutions to problem (3.1) with $\alpha=\alpha_1$ and $\alpha=\alpha_2$ respectively. We have $D(\boldsymbol{v}_{\alpha_1}^*)-D(\boldsymbol{v}_{\alpha_2}^*)\leqslant\alpha_1\mathbf{d}^T(\boldsymbol{v}_{\alpha_2}^*-\boldsymbol{v}_{\alpha_1}^*)$ and $D(\boldsymbol{v}_{\alpha_1}^*)-D(\boldsymbol{v}_{\alpha_2}^*)\geqslant\alpha_2\mathbf{d}^T(\boldsymbol{v}_{\alpha_2}^*-\boldsymbol{v}_{\alpha_1}^*)$, so that, necessarily, $\mathbf{d}^T(\boldsymbol{v}_{\alpha_2}^*-\boldsymbol{v}_{\alpha_1}^*)\leqslant 0$, and therefore $D(\boldsymbol{v}_{\alpha_1}^*)-D(\boldsymbol{v}_{\alpha_2}^*)\leqslant 0$. □

Proof of Lemma 3.3. If $\boldsymbol{v}_\alpha^*$ is a solution to problem (3.1) with $\alpha\geqslant 0$, then by definition, $\boldsymbol{v}_\alpha^*$ in particular minimises $D(\cdot)$ over the set of all $\boldsymbol{v}\geqslant 0$ such that $\mathbf{d}^T\boldsymbol{v}=\mathbf{d}^T\boldsymbol{v}_\alpha^*$, and $\boldsymbol{v}_\alpha^*$ is therefore a solution to problem (3.2) with $\varkappa=\mathbf{d}^T\boldsymbol{v}_\alpha^*$.

The condition $\varkappa\leqslant\mathbf{d}^T\boldsymbol{\omega}$ follows directly from Lemma 3.2-(iii): a solution $\boldsymbol{v}_\alpha^*$ to problem (3.1) indeed necessarily satisfies $\mathbf{d}^T\boldsymbol{v}_\alpha^*\leqslant\mathbf{d}^T\boldsymbol{\omega}$. For $\varkappa=0$, we have $\boldsymbol{v}_\varkappa^*=0$ and the result follows from Lemma 3.2-(ii). For $0<\varkappa\leqslant\mathbf{d}^T\boldsymbol{\omega}$, in order to be a solution to problem (3.1), a solution $\boldsymbol{v}_\varkappa^*$ to problem (3.2) must satisfy (first oder optimality condition), $(\boldsymbol{v}-\boldsymbol{v}_\varkappa^*)^T\nabla D_\alpha(\boldsymbol{v}_\varkappa^*)\geqslant 0$ for all $\boldsymbol{v}\geqslant 0$. In particular, for all $k\in\{1,\cdots,N\}$ such that $[\boldsymbol{v}_\varkappa^*]_k>0$, the constraint $v_k\geqslant 0$ is not active; therefore, for all $\theta\geqslant 0$, the parameter α must satisfy $(\theta\mathbf{e}_k-\boldsymbol{v}_\varkappa^*)^T\nabla D_\alpha(\boldsymbol{v}_\varkappa^*)=0$ where $\mathbf{e}_k$ stands for the k-th element of the canonical basis of $\mathbb{R}^N$ (so that $[\boldsymbol{v}_\varkappa^*]_k=\mathbf{e}_k^T\boldsymbol{v}_\varkappa^*$). Considering $\theta=0$ directly leads to the expected result (and one can also easily verify that the obtained value of α does not depend on $\theta\geqslant 0$). The last assertion follows directly from Lemma 3.2-(vi) and the correspondence between solutions to problems (3.1) and (3.2). □

Proof of Lemma 4.1. The proof follows directly from the expansion of the involved squared norms. We indeed have $\|T_\mu[\widehat{\varphi}_l]\|^2_{L^2(\mu)}=\widehat{\lambda}_l^2$, $\|\widehat{\lambda}_l\widehat{\varphi}_l\|^2_{L^2(\mu)}=\widehat{\lambda}_l^2$, and $(\widehat{\lambda}_l\widehat{\varphi}_l|T_\mu[\widehat{\varphi}_l])_{L^2(\mu)}=\widehat{\lambda}_l^2\Upsilon_l$; and $\|T_\mu[\widehat{\varphi}_l]\|^2_{\mathcal{H}}=\widehat{\lambda}_l\Upsilon_l$, $\|\widehat{\lambda}_l\widehat{\varphi}_l\|^2_{\mathcal{H}}=\widehat{\lambda}_l\|\widehat{\lambda}_l^{1/2}\widehat{\varphi}_l\|^2_{\mathcal{H}}$, and $(\widehat{\lambda}_l\widehat{\varphi}_l|T_\mu[\widehat{\varphi}_l])_{\mathcal{H}}=\widehat{\lambda}_l$. □

Proof of Lemma 5.1. Define the closed linear subspace $\mathcal{F}_S=\operatorname{span}\{K^2_{x_k}\}_{k=1}^N$ of $\mathcal{F}$ and let $\mathcal{F}_0=\mathcal{F}_S^\perp$ be its orthogonal; by definition, $g_{\boldsymbol{\omega}}\in\mathcal{F}_S$. For any $g_S\in\mathcal{F}_S$ and $g_0\in\mathcal{F}_0$, we have

$$\tfrac{1}{2}\|g_S\|^2_{\mathcal{F}}+(g_S|g_{\boldsymbol{\omega}})_{\mathcal{F}}\leqslant\tfrac{1}{2}\|g_S+g_0\|^2_{\mathcal{F}}+(g_S+g_0|g_{\boldsymbol{\omega}})_{\mathcal{F}}=\tfrac{1}{2}\|g_S\|^2_{\mathcal{F}}+(g_S|g_{\boldsymbol{\omega}})_{\mathcal{F}}+\tfrac{1}{2}\|g_0\|^2_{\mathcal{F}}.$$

In addition, for any $k\in\{1,\cdots,N\}$, we have $g_0(x_k)=0$, so that necessarily $g_\alpha^*\in\mathcal{F}_S$ (representer theorem), i.e, there exists $\boldsymbol{\beta}^*=(\beta_1^*,\cdots,\beta_N^*)^T\in\mathbb{R}^N$ such that $g_\alpha^*=\sum_{k=1}^N\beta_k^*K^2_{x_k}$. Restricting problem (5.1) to $\mathcal{F}_S$ then yields, for $\boldsymbol{\beta}\in\mathbb{R}^N$,

$$\underset{\boldsymbol{\beta}}{\text{minimise}}\ \tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\omega}\ \text{subject to}\ \mathbf{S}\boldsymbol{\beta}\geqslant-\alpha\mathbf{d}. \tag{A.1}$$

We then introduce the Lagrangian function, for $\boldsymbol{v}\in\mathbb{R}^N$ with $\boldsymbol{v}\geqslant 0$ (dual feasibility conditions),

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{v})=\tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\omega}-\boldsymbol{v}^T\left[\mathbf{S}\boldsymbol{\beta}+\alpha\mathbf{d}\right].$$

The primal optimality conditions give $\mathbf{S}\boldsymbol{\beta} = \mathbf{S}(\boldsymbol{\upsilon} - \boldsymbol{\omega})$, leading to the Lagrange dual problem (3.1) (written as a minimisation problem). If $\boldsymbol{\upsilon}_\alpha^*$ is a solution to (3.2), then a solution $\boldsymbol{\beta}^*$ to (A.1) needs to satisfy $\mathbf{S}\boldsymbol{\beta}^* = \mathbf{S}(\boldsymbol{\upsilon}_\alpha^* - \boldsymbol{\omega})$, so that we can in particular consider $\boldsymbol{\beta}^* = \boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega}$. Notice that when $\mathbf{S}$ is non-invertible, other choices for $\boldsymbol{\beta}^*$ exist since for any $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ such that $\mathbf{S}\boldsymbol{\varepsilon} = 0$, we have $\mathbf{S}(\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}) = \mathbf{S}\boldsymbol{\beta}^*$; but the obtained solution $g_\alpha^* \in \mathcal{F}_S$ does not depend on such a $\boldsymbol{\varepsilon}$ (see also Remark 3.2). The equality $g_\alpha^*(x_k) = -\alpha d_k$ for all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{\upsilon}_\alpha^*]_k > 0$ is consequence of the complementary slackness condition $(\boldsymbol{\upsilon}_\alpha^*)^T \left[\mathbf{S}(\boldsymbol{\upsilon}_\alpha^* - \boldsymbol{\omega}) + \alpha\mathbf{d}\right] = 0$. □

Proof of Lemma 5.2. We follow the same reasoning than in the proof of Lemma 5.1. By restricting problem (5.3) to $\mathcal{F}_S$, we obtain, for $\boldsymbol{\beta} \in \mathbb{R}^N$,

$$\underset{\boldsymbol{\beta},\gamma}{\text{minimise}}\ \tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\omega} - \gamma \text{ subject to } \mathbf{S}\boldsymbol{\beta} \geqslant \gamma\mathbf{d}/\varkappa. \tag{A.2}$$

The underlying Lagrangian function is then given by, for $\boldsymbol{\upsilon} \in \mathbb{R}^N$ with $\boldsymbol{\upsilon} \geqslant 0$ (dual feasibility conditions),

$$\mathcal{L}(\boldsymbol{\beta}, \gamma, \boldsymbol{\upsilon}) = \tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\omega} - \gamma - \boldsymbol{\upsilon}^T\left[\mathbf{S}\boldsymbol{\beta} - \gamma\mathbf{d}/\varkappa\right].$$

The primal optimality conditions give $\mathbf{S}\boldsymbol{\beta} = \mathbf{S}(\boldsymbol{\upsilon} - \boldsymbol{\omega})$ and $\mathbf{d}^T\boldsymbol{\upsilon} = \varkappa$, leading to the Lagrange dual problem (3.2). If $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to (3.2), then a solution $\boldsymbol{\beta}^*$ to (A.2) needs to satisfy $\mathbf{S}\boldsymbol{\beta}^* = \mathbf{S}(\boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega})$, so that we can in particular consider $\boldsymbol{\beta}^* = \boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega}$. The expression of $\gamma_\varkappa^*$ follows form the complementary slackness condition $(\boldsymbol{\upsilon}_\varkappa^*)^T\left[\mathbf{S}(\boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega}) - \gamma_\varkappa^*\mathbf{d}/\varkappa\right] = 0$. The equality $g_\varkappa^*(x_k) = \gamma_\varkappa^* d_k/\varkappa$ for all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{\upsilon}_\varkappa^*]_k > 0$ is also consequence of the complementary slackness condition. □

Proof of Lemma 6.1. Let $\boldsymbol{\upsilon}_\alpha = \theta\boldsymbol{\upsilon}_{\alpha_1}^* + (1-\theta)\boldsymbol{\upsilon}_{\alpha_2}^*$ and define $J = J_{\alpha_1} = J_{\alpha_2}$; we have

$$\mathbf{S}_{J,J}[\boldsymbol{\upsilon}_\alpha]_J = \mathbf{S}_{J,J}[\theta\boldsymbol{\upsilon}_{\alpha_1}^* + (1-\theta)\boldsymbol{\upsilon}_{\alpha_2}^*]_J = [\mathbf{S}\boldsymbol{\omega}]_J - \alpha\mathbf{d}_J,$$

so that $[\mathbf{S}(\boldsymbol{\upsilon}_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{d}]_J = 0$, and in the same way,

$$[\mathbf{S}(\boldsymbol{\upsilon}_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{d}]_{J^c} = \theta[\mathbf{S}(\boldsymbol{\upsilon}_{\alpha_1}^* - \boldsymbol{\omega}) + \alpha_1\mathbf{d}]_{J^c} + (1-\theta)[\mathbf{S}(\boldsymbol{\upsilon}_{\alpha_2}^* - \boldsymbol{\omega}) + \alpha_2\mathbf{d}]_{J^c} > 0.$$

By construction, $\boldsymbol{\upsilon}_\alpha \geqslant 0$ and in addition, if k is such that $[\boldsymbol{\upsilon}_\alpha]_k > 0$, then $k \in J$ (since these conditions are verified by both $\boldsymbol{\upsilon}_{\alpha_1}^*$ and $\boldsymbol{\upsilon}_{\alpha_2}^*$). We therefore have $\boldsymbol{\upsilon}_\alpha^T\left(\mathbf{S}(\boldsymbol{\upsilon}_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{d}\right) = 0$, so that for all $\boldsymbol{\upsilon} \geqslant 0$, the optimality condition $(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_\alpha)^T \nabla D_\alpha(\boldsymbol{\upsilon}_\alpha) \geqslant 0$ holds, i.e., $\boldsymbol{\upsilon}_\alpha$ is a solution to (3.1), and $J_\alpha = J$. □

REFERENCES

[1] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science, 2011.

[2] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

[3] Steve B. Damelin. A walk through energy, discrepancy, numerical integration and group invariant measures on measurable subsets of Euclidean space. *Numerical Algorithms*, 48(1-3):213–235, 2008.

[4] Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

[5] Bertrand Gauthier and Luc Pronzato. Spectral approximation of the IMSE criterion for optimal designs in kernel-based interpolation models. *SIAM/ASA Journal on Uncertainty Quantification*, 2:805–825, 2014.

[6] Alex Gittens and Michael W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 17:1–65, 2016.

[7] Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's "away step". *Mathematical Programming*, 35(1):110–119, 1986.

[8] Wolfgang Hackbusch. *Integral equations: theory and numerical treatment*, volume 120. Birkhäuser, 2012.

[9] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the Support Vector Machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

[10] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press, 2015.

[11] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.

[12] Moshe Lichman. UCI Machine Learning Repository, 2013.

[13] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[14] Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

[15] Luc Pronzato and Andrej Pázman. *Design of Experiments in Nonlinear Models*. Springer, 2013.

[16] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[17] Laurent Schwartz. *Analyse Hilbertienne*. Hermann, 1978.

[18] Steve Smale and Ding-Xuan Zhou. Geometry on probability spaces. *Constructive Approximation*, 30(3):311–323, 2009.

[19] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer, 2008.

[20] Shusen Wang and Zhihua Zhang. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *The Journal of Machine Learning Research*, 14:2729–2769, 2013.