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. The design of sparse quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels is addressed. Particular emphasis is placed on the approximation of the main eigenpairs of the initial operator. A special attention is drawn to the design of sparse quadratures with support included in fixed finite sets of points (that is, quadrature-sparsification), this framework in particular encompassing the approximation of kernel matrices. For a given kernel, the accuracy of a quadrature approximation is assessed through the squared Hilbert-Schmidt norm (for operators acting on the underlying reproducing kernel Hilbert space) of the difference between the integral operators related to the initial and approximate measures; by analogy with the notion of kernel discrepancy, the underlying criterion is referred to as the squared-kernel discrepancy between the two measures. In the quadrature-sparsification framework, sparsity of the approximate quadrature is promoted through the introduction of an ℓ^1-type penalisation, and the computation of a penalised squared-kernel-discrepancy-optimal approximation then consists in a convex quadratic minimisation problem; such quadratic programs can in particular be interpreted as the Lagrange dual formulations of distorted one-class support-vector machines related to the squared kernel. Error bounds on the induced spectral approximations are derived, and the connection between penalisation and accuracy of the approximation is investigated. Numerical strategies for solving large-scale penalised squared-kernel-discrepancy minimisation problems are discussed, and the efficiency of the approach is illustrated by a series of examples. In particular, the ability of the proposed methodology to lead to accurate approximations of the main eigenpairs of kernel matrices related to large-scale datasets is demonstrated.

Key words. sparse Nyström approximation, RKHS, squared-kernel discrepancy, L1-type penalisation, convex quadratic programming, one-class SVM, approximate eigendecomposition.

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

1. Introduction.

This work addresses the problem of designing sparse quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels. We more specifically aim at obtaining quadratures leading to an accurate approximation of the main eigendirections of a given initial operator (i.e., the directions related to the largest eigenvalues). A particular attention is paid to the *quadrature-sparsification* problem, which consists in designing sparse quadratures with support included in a fixed finite set of points; this framework in particular encompasses the landmark-selection, or column-sampling, problem for the approximation of large-scale kernel matrices, see for instance [6, 9, 1] and references therein.

In practice, dealing with sparse quadratures is specially important when one aims, for instance, at computing the spectral decomposition of the approximate operator (in order to approximate the eigendecomposition of the initial operator, as considered in this work). Indeed, the spectral decomposition of an operator defined from a discrete measure supported by n points involves the diagonalisation of a $n \times n$ matrix; in the general case, the amount of computations required to perform this task scales as $\mathcal{O}(n^3)$ and becomes intractable for large values of n (not to mention storage issues).

The choice of the quadrature can strongly impact the quality of the induced approximation (especially since we consider sparse quadratures), raising question relative to the characterisation and the construction of "effective" quadratures. Following for instance [21, 22], we assess the approximation error through the squared Hilbert-Schmidt norm of the difference between the initial and approximate operators, both operators being interpreted as operators acting on the reproducing kernel Hilbert space (RKHS, see for instance [3]) associated with the kernel; we refer to the underlying criterion as the *squared-kernel discrepancy* between the initial and approximate measures (i.e., the measures defining, in combination with the kernel, the initial and approximate operators). The squared-kernel discrepancy can in addition be interpreted as a "weighted spectral sum-of-squared-errors-type criterion", further highlighting the interest of low squared-kernel-discrepancy configurations for the

[*]gauthierb@cardiff.ac.uk (corresponding author)
[‡]johan.suykens@esat.kuleuven.be
[§]Cardiff University, School of Mathematics
Senghennydd Road, Cardiff, CF24 4AG, United Kingdown
[†]KU Leuven, ESAT-STADIUS Centre for Dynamical Systems, Signal Processing and Data Analytics
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

computation of accurate spectral approximations.

For approximate measures with support included in a fixed finite set of points, the squared-kernel discrepancy can be expressed as a convex quadratic function; sparsity of the approximate measure can then be promoted through the introduction of an ℓ^1*-type penalisation*, and the induced penalised squared-kernel-discrepancy minimisation problems then consist in convex quadratic minimisation problems. The so obtained quadratic programs (QP) can be interpreted as the Lagrange duals of *distorted one-class support-vector machines* (SVM, see, e.g., [23]) defined from the squared kernel, the initial measure and the penalisation term; the points selected through penalised squared-kernel-discrepancy minimisation thus correspond to the support vectors of these SVMs. In the matrix-approximation framework, penalised squared-kernel-discrepancy minimisation defines a deterministic, QP-based, weighted column-sampling scheme, and offers a complement to the existing column-sampling-based methodology for kernel-matrix approximation; see, e.g., [6, 25, 13, 24, 4, 9] for an overview.

We present a careful analysis of the approach and describe numerical strategies to address large-scale penalised squared-kernel-discrepancy minimisation problems. We pay a special attention to the approximation of the main eigenpairs of an initial operator induced by the eigendecomposition of an approximate operator; to assess the accuracy of an approximate eigendirection while estimating the associated approximate eigenvalue, we in particular rely on the notion of *geometric approximate eigenvalues*. Motivated by the invariance property of the spectral approximations induced by proportional approximate measures, and in order to derive bounds on the overall accuracy of these spectral approximations, we also introduce the notion of *conic squared-kernel discrepancy*.

The paper is organised as follows. Section 2 introduces the theoretical framework considered in this work, and Section 3 discusses the approximate eigendecomposition of an operator. Section 4 focuses on approximate measures with support included in a fixed finite set of points (i.e., quadrature-sparsification) and on kernel-matrix approximation. The penalised squared-kernel-discrepancy-minimisation problems are introduced in Section 5, and the underlying SVM models are described in Section 6. Numerical strategies to solve large-scale penalised problems are investigated in Sections 7 and 8. Section 9 is devoted to a discussion relative to the selection of relevant penalisation directions. Some numerical experiments are carried out in Sections 10 and 11, and Section 12 concludes.

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

2. Notations, recalls and theoretical background. We consider a general space $\mathscr{X}$ and a symmetric and positive-semidefinite kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$; we denote by $\mathcal{H}$ the underlying RKHS of real-valued functions on $\mathscr{X}$ (see for instance [3]). We assume that $\mathcal{H}$ is a separable Hilbert space.

2.1. Integral operators. We assume that $\mathscr{X}$ is a measurable space and we denote by $\mathcal{A}$ the underlying σ-algebra. We suppose that the 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 [23, Chap. 4]), so that $\mathcal{H}$ 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

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

For $\mu \in \mathscr{T}(K)$, we have $K(\cdot,\cdot) \in L^2(\mu \otimes \mu)$ since in particular (from the reproducing property of $K(\cdot,\cdot)$ and the Cauchy-Schwarz inequality for the inner product of $\mathcal{H}$)

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

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 can thus define the symmetric and positive-semidefinite integral operator T_μ on $L^2(\mu)$, given by, for $f \in L^2(\mu)$ and $x \in \mathscr{X}$,

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

In particular, for all $f \in L^2(\mu)$, we have $T_\mu[f] \in \mathcal{H} \subset L^2(\mu)$, and for all $h \in \mathcal{H}$,

$$\big(h\big|T_\mu[f]\big)_{\mathcal{H}} = \big(h\big|f\big)_{L^2(\mu)}, \tag{2.1}$$

where $(\cdot|\cdot)_{\mathcal{H}}$ and $(\cdot|\cdot)_{L^2(\mu)}$ stand for the inner products of $\mathcal{H}$ and $L^2(\mu)$, respectively; see for instance [7, 8] for more details.

We introduce the closed linear subspaces $\mathcal{H}_{0\mu} = \big\{h \in \mathcal{H} \big| \|h\|_{L^2(\mu)} = 0\big\}$ 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 $\{\widetilde{\varphi}_k\}_{k\in\mathbb{I}_\mu^+}$ be a set of associated eigenfunctions, chosen to be orthonormal in $L^2(\mu)$, i.e., $\widetilde{\varphi}_k \in L^2(\mu)$, $T_\mu[\widetilde{\varphi}_k] = \lambda_k\widetilde{\varphi}_k$ in $L^2(\mu)$, and $(\widetilde{\varphi}_k|\widetilde{\varphi}_{k'})_{L^2(\mu)} = \delta_{k,k'}$ (Kronecker delta). 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 $\widetilde{\varphi}_k$ (the eigenfunctions $\widetilde{\varphi}_k$ are indeed only defined μ-almost everywhere, while the extensions φ_k are defined for all $x \in \mathcal{X}$). From (2.1), we obtain 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 $\mathcal{H}$, and the reproducing kernel $K_\mu(\cdot,\cdot)$ of $\mathcal{H}_\mu$ is thus given by, for all x and $t \in \mathcal{X}$,

$$K_\mu(x,t) = \textstyle\sum_{k\in\mathbb{I}_\mu^+} \lambda_k\varphi_k(x)\varphi_k(t). \tag{2.2}$$

We also recall that $\tau_\mu = \sum_{k\in\mathbb{I}_\mu^+}\lambda_k$ is the trace of the integral operator T_μ on $L^2(\mu)$.

Remark 2.1. Consider any measure $\mu \in \mathscr{T}(K)$; for $c > 0$, the strictly positive eigenvalues of the operator $T_{c\mu} = cT_\mu$ (i.e., the operator defined by the kernel $K(\cdot,\cdot)$ and the measure $c\mu$) are $c\lambda_k$, with $k \in \mathbb{I}_\mu^+$, and the associated (canonically extended) eigenfunctions, orthonormalised in $L^2(c\mu)$, are $\varphi_k/\sqrt{c}$. In particular, we have $\mathcal{H}_\mu = \mathcal{H}_{c\mu}$, and $K_\mu(\cdot,\cdot) = K_{c\mu}(\cdot,\cdot)$. ◁

2.2. 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 $\mathcal{H}$ (see, e.g., [21, 22]); 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.

We denote by $\mathrm{HS}(\mathcal{H})$ the Hilbert space of all Hilbert-Schmidt operators on $\mathcal{H}$. Let μ and $\nu \in \mathscr{T}(K)$; for an o.n.b. $\{h_j\}_{j\in\mathbb{I}}$ 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

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

and we recall that the value of $\big(T_\mu\big|T_\nu\big)_{\mathrm{HS}(\mathcal{H})}$ does not depend on the choice of the o.n.b. of $\mathcal{H}$, see, e.g., [19]. The underlying Hilbert-Schmidt norm (for operators on $\mathcal{H}$) is given by

$$\big\|T_\mu\big\|^2_{\mathrm{HS}(\mathcal{H})} = \big(T_\mu\big|T_\mu\big)_{\mathrm{HS}(\mathcal{H})} = \textstyle\sum_{j\in\mathbb{I}} \big\|T_\mu[h_j]\big\|^2_{\mathcal{H}}.$$

Definition 2.1. *The squared-kernel discrepancy $D_{K^2}(\mu,\nu)$ between μ and $\nu \in \mathscr{T}(K)$ is defined as*

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

Proposition 2.1. *For μ and $\nu \in \mathscr{T}(K)$, we have $(T_\mu|T_\nu)_{\mathrm{HS}(\mathcal{H})} = \|K\|^2_{L^2(\mu\otimes\nu)}$, so that*

$$D_{K^2}(\mu,\nu) = \|K\|^2_{L^2(\mu\otimes\mu)} + \|K\|^2_{L^2(\nu\otimes\nu)} - 2\|K\|^2_{L^2(\mu\otimes\nu)},$$

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

In particular, notice that $\|K\|^2_{L^2(\mu\otimes\nu)} \leqslant \tau_\mu\tau_\nu$, and that $\|T_\mu\|^2_{\mathrm{HS}(\mathcal{H})} = \sum_{k\in\mathbb{I}_\mu^+}\lambda_k^2$, where $\{\lambda_k\}_{k\in\mathbb{I}_\mu^+}$ is the set of all strictly positive eigenvalues of T_μ. By definition, we always have $D_{K^2}(\mu,\nu) \geqslant 0$, and $D_{K^2}(\mu,\mu) = 0$. We can 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 $\|K\|^2_{L^2(\mu\otimes\nu)} = 0$.

Lemma 2.1. *We denote by $\mathcal{G}$ the RKHS associated with the squared kernel $K^2(\cdot,\cdot) = \big(K(\cdot,\cdot)\big)^2$, and for all $\mu \in \mathscr{T}(K)$, we introduce the function $g_\mu(x) = \int_{\mathscr{X}} K^2(x,t)\mathrm{d}\mu(t)$, with $x \in \mathscr{X}$. For all μ and $\nu \in \mathscr{T}(K)$, we have g_μ and $g_\nu \in \mathcal{G}$, and*

$$(T_\mu|T_\nu)_{\mathrm{HS}(\mathcal{H})} = (g_\mu|g_\nu)_{\mathcal{G}} = \|K\|^2_{L^2(\mu\otimes\nu)} = \int_{\mathscr{X}} g_\mu(t)\mathrm{d}\nu(t) = \int_{\mathscr{X}} g_\nu(t)\mathrm{d}\mu(t),$$

so that, in particular, $D_{K^2}(\mu,\nu) = \|g_\mu - g_\nu\|^2_{\mathcal{G}}$.

The terminology *"squared-kernel discrepancy"* is motivated by the analogy with the notion of *"kernel discrepancy"* discussed for instance in [5, 20] (see Appendix A). Interestingly, the kernel discrepancy is related to approximate integration of functions in the RKHS $\mathcal{H}$, while the squared-kernel discrepancy is related to the approximation of integral operators defined from the reproducing kernel $K(\cdot,\cdot)$ of $\mathcal{H}$; by definition, the squared-kernel discrepancy is thus also related to approximate integration of functions in the RKHS $\mathcal{G}$ associated with the squared kernel $K^2(\cdot,\cdot)$.

Lemma 2.2. *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 $\{\sqrt{\lambda_k}\varphi_k\}_{k\in\mathbb{I}^+_\mu}$ be an o.n.b. of $\mathcal{H}_\mu$ defined by the spectral decomposition of T_μ, we have*

$$D_{K^2}(\mu,\nu) = \textstyle\sum_{k\in\mathbb{I}^+_\mu} \lambda_k \big\|T_\mu[\varphi_k] - T_\nu[\varphi_k]\big\|^2_{\mathcal{H}}, \tag{2.3}$$

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

In the framework of Lemma 2.2, and assuming that one aims at approximating T_μ (the *initial operator*) by T_ν (the *approximate operator*), the squared-kernel discrepancy can, in view of (2.3), be interpreted as a "weighted spectral sum-of-squared-errors-type criterion", the eigenvalues λ_k playing the rule of penalisation weights. When $D_{K^2}(\mu,\nu)$ is small, we can therefore expect the main eigendirections of T_ν to be accurate approximations of the main eigendirections of T_μ (and reciprocally), see in particular the forthcoming Theorem 3.2.

Remark 2.2. In Lemma 2.2, if the condition $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ is omitted, then the term $\sum_{m\in\mathbb{J}} \|T_\nu[h_m]\|^2_{\mathcal{H}} = (K|K_{0\mu})_{L^2(\nu\otimes\nu)} = (K_\nu|K_{0\mu})_{L^2(\nu\otimes\nu)} \geqslant 0$ needs to be added to the right-hand side of (2.3), where $\{h_m\}_{m\in\mathbb{J}}$ is an o.n.b. of the subspace $\mathcal{H}_{0\mu}$ of $\mathcal{H}$, and $K_{0\mu}(\cdot,\cdot)$ is the kernel of $\mathcal{H}_{0\mu}$, and $K_\nu(\cdot,\cdot)$ is the kernel of the subspace $\mathcal{H}_\nu$ related to T_ν (see Section 3). Nevertheless, the conclusions drawn from Lemma 2.2 still hold. Also notice that, in Lemma 2.2, we express the squared-kernel discrepancy as function of the eigenpairs of T_μ, but we might as well have used the eigenpairs of T_ν. ◁

Since $D_{K^2}(\mu,\mu) = 0$ (i.e., "the best approximation of T_μ is T_μ itself"), the unconstrained minimisation of $\nu \mapsto D_{K^2}(\mu,\nu)$ on $\mathscr{T}(K)$ is of no interest. Furthermore, in the framework of sparse pointwise quadrature approximation, we aim at obtaining a discrete measure ν supported by a relatively small number of points (in order to be able to compute the eigendecomposition of T_ν) and related to an as low as possible value of $D_{K^2}(\mu,\nu)$. However, for a given $n \in \mathbb{N}^*$, the search of an optimal discrete measure ν_n^* such that $D_{K^2}(\mu,\nu_n^*)$ is minimal among all measures ν_n supported by n points is in general a difficult (i.e., usually non-convex) optimisation problem on $(\mathscr{X}\times\mathbb{R}_+)^n$. To avoid this difficulty, we restrict the squared-kernel-discrepancy minimisation to measures ν with support included in a fixed finite set of points $\mathcal{S} = \{x_k\}_{k=1}^N$ (with, in practice, N large), see Section 4.2, and instead of fixing a priori the number n of support points, we promote sparsity through the introduction of an ℓ^1-type penalisation, as considered in Section 5.

3. Approximate eigendecomposition.

We consider two measures μ and $\nu \in \mathscr{T}(K)$, corresponding to an initial operator T_μ and an approximate operator T_ν.

3.1. Geometric approximate eigenvalues.

Following Section 2.1, we denote by $\{\sqrt{\lambda_k}\varphi_k\}_{k\in\mathbb{I}^+_\mu}$ an o.n.b. of $\mathcal{H}_\mu$ defined by the eigendecomposition of T_μ; in the same way, let $\{\sqrt{\vartheta_l}\psi_l\}_{l\in\mathbb{I}^+_\nu}$ be an o.n.b. of the subspace $\mathcal{H}_\nu$ of $\mathcal{H}$ related to T_ν, i.e., $T_\nu[\psi_l] = \vartheta_l\psi_l \in \mathcal{H}$, with $\vartheta_l > 0$ and $(\psi_l|\psi_{l'})_{L^2(\nu)} = \delta_{l,l'}$. We shall refer to the functions ψ_l as the *approximate eigendirections* of T_μ induced by T_ν. We recall that, from (2.1), we have

$$\|\psi_l\|^2_{L^2(\mu)} = \big(\psi_l\big|T_\mu[\psi_l]\big)_{\mathcal{H}} \text{ and } \|T_\mu[\psi_l]\|^2_{\mathcal{H}} = \big(\psi_l\big|T_\mu[\psi_l]\big)_{L^2(\mu)}.$$

Definition 3.1. *For all $l \in \mathbb{I}_\nu^+$ such that $\|\psi_l\|_{L^2(\mu)} > 0$, we introduce $\widehat{\varphi}_l = \psi_l / \|\psi_l\|_{L^2(\mu)}$, the normalised approximate eigenfunctions of T_μ induced by the spectral decomposition of T_ν.*

Notice that if $\mathcal{H}_\nu \subset \mathcal{H}_\mu$, then we necessarily have $\|\psi_l\|_{L^2(\mu)} > 0$ for all $l \in \mathbb{I}_\nu^+$. If $\|\psi_l\|_{L^2(\mu)} = 0$, then $\psi_l \in \mathcal{H}_{0\mu}$ and thus $T_\mu[\psi_l] = 0$; such directions are therefore of no use in approximating the eigendirections related to the strictly positive eigenvalues of T_μ.

Remark 3.1 (Orthogonality test)**.** The normalised approximate eigenfunctions $\widehat{\varphi}_l$ are by definition orthogonal in $L^2(\nu)$ and in $\mathcal{H}$, and verify $\|\widehat{\varphi}_l\|_{L^2(\mu)} = 1$. Controlling the orthogonality, in $L^2(\mu)$, between the approximations $\widehat{\varphi}_l$ appears as a relatively affordable way to assess their accuracy. Indeed, from (2.1) and due to their orthogonality in $\mathcal{H}$, accurate normalised approximate eigenfunctions $\widehat{\varphi}_l$ should be almost mutually orthogonal in $L^2(\mu)$; notice that this condition is however only a necessary condition. See Sections 10 and 11 for illustrations. ◁

It is very instructive to try to estimate the eigenvalue, for the operator T_μ, related to an approximate eigendirection ψ_l induced by T_ν, as discussed hereafter.

Definition 3.2. *For all $l \in \mathbb{I}_\nu^+$ such that $\|\psi_l\|_{L^2(\mu)} > 0$, we define*

$$\begin{aligned}
\widehat{\lambda}_l^{[1]} &= 1/\|\widehat{\varphi}_l\|_{\mathcal{H}}^2 = \vartheta_l \|\psi_l\|_{L^2(\mu)}^2 = \big(\sqrt{\vartheta_l}\psi_l \big| T_\mu[\sqrt{\vartheta_l}\psi_l]\big)_{\mathcal{H}} = \big(T_\nu[\psi_l] \big| T_\mu[\psi_l]\big)_{\mathcal{H}}, \\
\widehat{\lambda}_l^{[2]} &= \big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}}, \\
\widehat{\lambda}_l^{[3]} &= \big(\widehat{\varphi}_l \big| T_\mu[\widehat{\varphi}_l]\big)_{L^2(\mu)} = \big\|T_\mu[\widehat{\varphi}_l]\big\|_{\mathcal{H}}^2 = \big(\widehat{\lambda}_l^{[2]}\big)^2 / \widehat{\lambda}_l^{[1]}, \\
\widehat{\lambda}_l^{[4]} &= \big\|T_\mu[\widehat{\varphi}_l]\big\|_{L^2(\mu)} = \big\|T_\mu[\psi_l]\big\|_{L^2(\mu)} / \|\psi_l\|_{L^2(\mu)},
\end{aligned}$$

and if $\|\psi_l\|_{L^2(\mu)} = 0$, we set $\widehat{\lambda}_l^{[1]} = \widehat{\lambda}_l^{[2]} = \widehat{\lambda}_l^{[3]} = \widehat{\lambda}_l^{[4]} = 0$.

We refer to $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ as the four geometric approximate eigenvalues of T_μ related to the approximate eigendirection ψ_l induced by T_ν.

The intuition behind these four approximate eigenvalues $\widehat{\lambda}_l^{[\cdot]}$ is further discussed in the proof of Theorem 3.1 (Appendix C); see Remark 3.2 for comments relative to their computation. Notice that the various expressions characterising $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ follow form (2.1) and Definition 3.1; in particular, if $\|\psi_l\|_{L^2(\mu)} > 0$, then $\sqrt{\widehat{\lambda}_l^{[1]}}\widehat{\varphi}_l = \sqrt{\vartheta_l}\psi_l$.

Theorem 3.1. *For all $l \in \mathbb{I}_\nu^+$, we have $\widehat{\lambda}_l^{[1]} \leqslant \widehat{\lambda}_l^{[2]} \leqslant \widehat{\lambda}_l^{[3]} \leqslant \widehat{\lambda}_l^{[4]}$, with equality when ψ_l is an eigendirection of T_μ; in case of equality, the approximation $\widehat{\lambda}_l^{[\cdot]}$ corresponds exactly to the eigenvalue of T_μ related to the eigendirection ψ_l (in particular, equality between the four geometric approximate eigenvalues occurs as soon as two of them are equal).*

In addition, for $\lambda \in \mathbb{R}$, the function

$$\lambda \mapsto \big\|\lambda\sqrt{\vartheta_l}\psi_l - T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}}^2 = \lambda^2 - 2\lambda\widehat{\lambda}_l^{[1]} + \big(\widehat{\lambda}_l^{[2]}\big)^2 \tag{3.1}$$

reaches its minimum at $\lambda = \widehat{\lambda}_l^{[1]}$. In the same way, if $\|\psi_l\|_{L^2(\mu)} > 0$ (so that the normalised approximate eigenfunction $\widehat{\varphi}_l$ is well-defined), the function

$$\lambda \mapsto \big\|\lambda\widehat{\varphi}_l - T_\mu[\widehat{\varphi}_l]\big\|_{L^2(\mu)}^2 = \lambda^2 - 2\lambda\widehat{\lambda}_l^{[3]} + \big(\widehat{\lambda}_l^{[4]}\big)^2 \tag{3.2}$$

reaches its minimum at $\lambda = \widehat{\lambda}_l^{[3]}$.

In view of Theorem 3.1, one may assess the accuracy of the approximate eigendirections ψ_l (as eigendirections for T_μ) by checking how close to each other are the approximations $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$. In particular, for $\widehat{\lambda}_l^{[1]} > 0$, we have $0 < \widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]} \leqslant 1$, and this ratio corresponds to the inner product, in $\mathcal{H}$, between the normalised functions $\sqrt{\vartheta_l}\psi_l$ and $T_\mu[\sqrt{\vartheta_l}\psi_l]/\big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}}$; in the same way, we have $0 < \widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]} \leqslant 1$, and this ratio stands for the inner product, in $L^2(\mu)$, between the normalised functions $\widehat{\varphi}_l$ and $T_\mu[\widehat{\varphi}_l]/\|T_\mu[\widehat{\varphi}_l]\|_{L^2(\mu)}$. From (3.1) and (3.2), one may also, for instance,

consider the quantities, for $\widehat{\lambda}_l^{[1]} > 0$,

$$\big\| \sqrt{\vartheta_l}\psi_l - T_\mu[\sqrt{\vartheta_l}\psi_l]/\widehat{\lambda}_l^{[1]} \big\|_{\mathcal{H}}^2 = \big(\widehat{\lambda}_l^{[2]}/\widehat{\lambda}_l^{[1]}\big)^2 - 1 \text{ and} \tag{3.3}$$

$$\big\| \widehat{\varphi}_l - T_\mu[\widehat{\varphi}_l]/\widehat{\lambda}_l^{[3]} \big\|_{L^2(\mu)}^2 = \big(\widehat{\lambda}_l^{[4]}/\widehat{\lambda}_l^{[3]}\big)^2 - 1, \tag{3.4}$$

so that the closer (3.3) and (3.4) are from zero, the more accurate is the approximate eigendirection ψ_l; see Sections 10 and 11 for illustrations.

Remark 3.2. Once ϑ_l and ψ_l are known, we obtain the normalised approximate eigenfunction $\widehat{\varphi}_l$ and the approximate eigenvalue $\widehat{\lambda}_l^{[1]}$ by simply evaluating $\|\psi_l\|_{L^2(\mu)}^2$. Computing the other approximate eigenvalues $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ requires the knowledge of $T_\mu[\psi_l]$. We then obtain $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ by evaluating an inner product in $L^2(\mu)$, and $\widehat{\lambda}_l^{[2]}$ follows from the relation $\widehat{\lambda}_l^{[2]} = \sqrt{\widehat{\lambda}_l^{[1]}\widehat{\lambda}_l^{[3]}}$.

Remark that we have to compute $T_\mu[\psi_l]$ (which may prove challenging) only when we are interested in assessing precisely the accuracy of an approximate eigendirection ψ_l of T_μ. Otherwise, we might simply consider the approximate eigenpairs $\{\widehat{\lambda}_l^{[1]}, \widehat{\varphi}_l\}_{l\in\mathbb{I}_\nu^+}$, while eventually checking the orthogonality, in $L^2(\mu)$, between the approximate eigendirections.

The computation of the geometric approximate eigenvalues when μ is a discrete measure with finite support is further discussed in Section 4.3. ◁

Notice that the kernel $K_\nu(\cdot,\cdot)$ of the subspace $\mathcal{H}_\nu$ of $\mathcal{H}$ is given by, for x and $t \in \mathscr{X}$,

$$K_\nu(x,t) = \textstyle\sum_{l\in\mathbb{I}_\nu^+} \vartheta_l \psi_l(x)\psi_l(t); \tag{3.5}$$

in addition, if $\|\psi_l\|_{L^2(\mu)} > 0$ for all $l \in \mathbb{I}_\nu^+$, we have $K_\nu(x,t) = \sum_{l\in\mathbb{I}_\nu^+} \widehat{\lambda}_l^{[1]}\widehat{\varphi}_l(x)\widehat{\varphi}_l(t)$.

Remark 3.3. Consider any measure $\nu \in \mathscr{T}(K)$; following Remark 2.1, the kernel $K_\nu(\cdot,\cdot)$ and the approximations $\widehat{\varphi}_l$, $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ (related to the initial operator T_μ) remain unchanged if we replaces ν by $c\nu$, for any $c > 0$. ◁

3.2. Conic squared-kernel discrepancy. In view of Remark 3.3, in the framework of Section 3.1, proportional (non-null) approximate measures lead to the same spectral approximation of T_μ and to the same approximate kernel $K_\nu(\cdot,\cdot)$. For a given measure $\nu \in \mathscr{T}(K)$, we can thus search the value of $c \geqslant 0$ for which $D_{K^2}(\mu, c\nu)$ is minimal.

Theorem 3.2. *Consider μ and $\nu \in \mathscr{T}(K)$, with ν such that $\|K\|_{L^2(\nu\otimes\nu)}^2 > 0$. We denote by c_ν the argument of the minimum of the function $\phi : c \mapsto \phi(c) = D_{K^2}(\mu, c\nu)$, with $c \in \mathbb{R}$; we have*

$$c_\nu = \frac{(T_\mu|T_\nu)_{\mathrm{HS}(\mathcal{H})}}{\|T_\nu\|_{\mathrm{HS}(\mathcal{H})}^2} = \frac{\|K\|_{L^2(\mu\otimes\nu)}^2}{\|K\|_{L^2(\nu\otimes\nu)}^2}, \text{ and } \phi\big(c_\nu\big) = \|K\|_{L^2(\mu\otimes\mu)}^2 - \frac{\|K\|_{L^2(\mu\otimes\nu)}^4}{\|K\|_{L^2(\nu\otimes\nu)}^2}.$$

In particular, $T_{c_\nu\nu} = c_\nu T_\nu$ is the orthogonal projection, in $\mathrm{HS}(\mathcal{H})$, of T_μ onto the linear subspace spanned by T_ν; in addition, $\|c_\nu T_\nu - \frac{1}{2}T_\mu\|_{\mathrm{HS}(\mathcal{H})}^2 = \frac{1}{4}\|K\|_{L^2(\mu\otimes\mu)}^2$, so that, in $\mathrm{HS}(\mathcal{H})$, the approximate operator $c_\nu T_\nu$ lies on a sphere centered at $\frac{1}{2}T_\mu$ and with radius $\frac{1}{2}\|T_\mu\|_{\mathrm{HS}(\mathcal{H})}$. Assuming that $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ (for simplicity and without loss of generality), we also have

$$\textstyle\sum_{l\in\mathbb{I}_\nu^+} \widehat{\lambda}_l^{[1]} \big\| T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l^{[1]}\widehat{\varphi}_l \big\|_{\mathcal{H}}^2 \leqslant \sum_{l\in\mathbb{I}_\nu^+} \widehat{\lambda}_l^{[1]} \big\| T_\mu[\widehat{\varphi}_l] - c_\nu\vartheta_l\widehat{\varphi}_l \big\|_{\mathcal{H}}^2 \leqslant D_{K^2}(\mu, c_\nu\nu), \text{ and} \tag{3.6}$$

$$\textstyle\sum_{l\in\mathbb{I}_\nu^+} \widehat{\lambda}_l^{[1]} \big\| T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l^{[3]}\widehat{\varphi}_l \big\|_{L^2(\mu)}^2 \leqslant \sum_{l\in\mathbb{I}_\nu^+} \widehat{\lambda}_l^{[1]} \big\| T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l^{[1]}\widehat{\varphi}_l \big\|_{L^2(\mu)}^2 \leqslant \tau_\mu D_{K^2}(\mu, c_\nu\nu). \tag{3.7}$$

In Theorem 3.2, we are exploiting the positive cone structure of $\mathscr{T}(K)$; we thus refer to $\phi\big(c_\nu\big) = D_{K^2}(\mu, c_\nu\nu)$ as the *conic squared-kernel discrepancy* between μ and ν (notice that the measure μ is fixed); to avoid confusion, we shall sometimes refer to $D_{K^2}(\mu,\nu)$ as the *raw squared-kernel discrepancy* between μ and ν. The operator $T_{c_\nu\nu}$ is the best approximation of T_μ (in terms of squared-kernel discrepancy) among all operators defined from measures proportional to ν, i.e., of the form $c\nu$, with $c \geqslant 0$. In view of (3.6) and (3.7), the conic squared-kernel discrepancy $D_{K^2}(\mu, c_\nu\nu)$ is directly

related to the accuracy of the spectral approximation of T_μ induced by the approximate operator T_ν. In the framework of Section 5, the conic squared-kernel discrepancy is used to compare the approximations obtained from different penalisation schemes.

Remark 3.4. In view of Theorem 3.2 and following Remark 2.1, in order to approximate the eigenvalues of the initial operator T_μ induced by the eigendecomposition of T_ν, we could also define the "globally rescaled" approximate eigenvalues $\{c_\nu \vartheta_l\}_{l \in \mathbb{I}_\nu^+}$; in comparison, the approximate eigenvalues $\{\widehat{\lambda}_l^{[1]}\}_{l \in \mathbb{I}_\nu^+}$ are "individually rescaled". ◁

4. The discrete case. We now investigate in more detail the case of discrete measures with finite support. We pay a particular attention to the situation where the initial measure μ is discrete and the support of ν is included in the support of μ.

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

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

with $\mathbf{W} = \operatorname{diag}(\boldsymbol{\omega})$, and $\mathbf{k}(x) = \big(K(x_1, x), \cdots, K(x_N, x)\big)^T$, 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 $\mathbf{KW}$, where $\mathbf{K} \in \mathbb{R}^{N \times N}$ is the kernel matrix with i, j entry $\mathbf{K}_{i,j} = K(x_i, x_j)$; in particular, we have $\mathbf{KWf} = \big(T_\mu[f](x_1), \cdots, T_\mu[f](x_N)\big)^T$.

We denote by $\lambda_1 \geqslant \cdots \geqslant \lambda_N \geqslant 0$ the eigenvalues of $\mathbf{KW}$ and by $\mathbf{v}_1, \cdots, \mathbf{v}_N$ a set of associated orthonormalised eigenvectors, i.e., $\mathbf{KW} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$, with $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_N)$ 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; since $\boldsymbol{\omega} > 0$, we also have

$$\mathbf{P}\mathbf{P}^T = \mathbf{W}^{-1} \text{ and } \mathbf{K} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T. \tag{4.1}$$

For $\lambda_k > 0$, the canonically extended eigenfunctions of T_μ are given by $\varphi_k(x) = \frac{1}{\lambda_k}\mathbf{k}^T(x)\mathbf{W}\mathbf{v}_k$, and we in particular have $\mathbf{v}_k = (\varphi_k(x_1), \cdots, \varphi_k(x_N))^T$.

For a general $\boldsymbol{\omega} > 0$, the matrix $\mathbf{KW}$ is non-symmetric; however, 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 the classical Euclidean space $\{\mathbb{R}^N, (\cdot|\cdot)_{\mathrm{Id}_N}\}$, with eigenvalues λ_k and orthonormalised eigenvectors $\mathbf{W}^{1/2}\mathbf{v}_k$. We can thus easily deduce the eigendecomposition of the matrix $\mathbf{KW}$ viewed as an operator on $\{\mathbb{R}^N, (\cdot|\cdot)_{\mathbf{W}}\}$ from the eigendecomposition of the symmetric matrix $\mathbf{W}^{1/2}\mathbf{K}\mathbf{W}^{1/2}$.

Remark 4.1. Let $\mu = \sum_{k=1}^N \omega_k \delta_{x_k}$, with $\boldsymbol{\omega} > 0$, and consider the kernel $K_\mu(\cdot, \cdot)$ of the subspace $\mathcal{H}_\mu$ of $\mathcal{H}$, see (2.2); also, introduce the $N \times N$ kernel matrix $\mathbf{K}_\mu$, with i, j entry $[\mathbf{K}_\mu]_{i,j} = K_\mu(x_i, x_j)$. From (4.1) and the definition of the eigenfunctions φ_k, we have $\mathbf{K}_\mu = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T = \mathbf{K}$. ◁

4.2. Restricting the support of the approximate measure. We consider a general measure $\mu \in \mathscr{T}(K)$ and a fixed set $\mathcal{S} = \{x_k\}_{k=1}^N$ of N points in $\mathscr{X}$. For a measure ν with support included in $\mathcal{S}$, i.e., $\nu = \sum_{k=1}^N \upsilon_k \delta_{x_k}$, with $\boldsymbol{\upsilon} = (\upsilon_1, \cdots, \upsilon_N)^T \geqslant 0$ (that is, $\upsilon_k \geqslant 0$ for all k), we have

$$\|K\|^2_{L^2(\nu \otimes \nu)} = \boldsymbol{\upsilon}^T \mathbf{S} \boldsymbol{\upsilon} \text{ and } \|K\|^2_{L^2(\mu \otimes \nu)} = \mathbf{g}_\mu^T \boldsymbol{\upsilon},$$

where $\mathbf{S}$ is the matrix defined by the squared kernel $K^2(\cdot, \cdot)$ and the set of points $\mathcal{S}$, i.e., with i, j entry $\mathbf{S}_{i,j} = K^2(x_i, x_j) \geqslant 0$ (the kernel matrix $\mathbf{S}$ is therefore non-negative and symmetric positive-semidefinite), and where $\mathbf{g}_\mu = (g_\mu(x_1), \cdots, g_\mu(x_N))^T \in \mathbb{R}^N$, with $g_\mu(x_k) = \int_{\mathscr{X}} K^2(x_k, t) \mathrm{d}\mu(t) \geqslant 0$.

Notice in particular that $\mathbf{S} = \mathbf{K} * \mathbf{K}$ (Hadamard product), where we recall that $\mathbf{K}$ is the kernel matrix defined by $K(\cdot,\cdot)$ and $\mathcal{S}$, i.e., $\mathbf{K}_{i,j} = K(x_i, x_j)$.

For such a discrete measure ν, we obtain

$$D_{K^2}(\mu,\nu) = \|K\|^2_{L^2(\mu\otimes\mu)} + \boldsymbol{\upsilon}^T\mathbf{S}\boldsymbol{\upsilon} - 2\mathbf{g}_\mu^T\boldsymbol{\upsilon}, \tag{4.2}$$

and $\nu \mapsto D_{K^2}(\mu,\nu)$ can in this way be interpreted as a quadratic function of $\boldsymbol{\upsilon} \in \mathbb{R}^N$ (i.e., the vector of the weights characterising ν). We shall refer to $\mathbf{g}_\mu$ as the *(dual) distortion term*.

Minimising $\boldsymbol{\upsilon} \mapsto \boldsymbol{\upsilon}^T\mathbf{S}\boldsymbol{\upsilon} - 2\mathbf{g}_\mu^T\boldsymbol{\upsilon}$ under the constraint $\boldsymbol{\upsilon} \geqslant 0$ leads to the best approximation of μ, in terms of squared-kernel discrepancy, among all discrete measures supported by $\mathcal{S}$. In practice, this minimisation however requires the knowledge of the vector $\mathbf{g}_\mu \in \mathbb{R}^N$, which might be problematic for general measures μ (although it can always be approximated). In this work, we nevertheless more specifically aim at computing approximate measures supported by a number of points significantly smaller than N, so that we do not consider such a minimisation; instead, we add an ℓ^1-type penalisation term to the squared-kernel-discrepancy, as detailed in Section 5.

4.3. The discrete-operator framework. Hereafter, we only consider measures with support included in a fixed set $\mathcal{S} = \{x_k\}_{k=1}^N$. More precisely, we assume that $\mu = \sum_{k=1}^N \omega_k\delta_{x_k}$, with $\boldsymbol{\omega} > 0$, and that $\nu = \sum_{k=1}^N \upsilon_k\delta_{x_k}$, with $\boldsymbol{\upsilon} \geqslant 0$, so that $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ for all $\boldsymbol{\upsilon} \geqslant 0$, and $\mathbf{g}_\mu = \mathbf{S}\boldsymbol{\omega}$, and $\|K\|^2_{L^2(\mu\otimes\mu)} = \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\omega}$. In the framework of Section 4.1, the operator T_μ thus corresponds to the matrix $\mathbf{KW}$, with $\mathbf{W} = \mathrm{diag}(\boldsymbol{\omega})$, and the operator T_ν corresponds to the matrix $\mathbf{KV}$, with $\mathbf{V} = \mathrm{diag}(\boldsymbol{\upsilon})$.

For such measures μ and ν (related to vectors $\boldsymbol{\omega} > 0$ and $\boldsymbol{\upsilon} \geqslant 0$, respectively), we have

$$D_{K^2}(\mu,\nu) = (\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}), \tag{4.3}$$

where we recall that $\mathbf{S} = \mathbf{K} * \mathbf{K}$, see Section 4.2.

Remark 4.2. Considering equation (4.3), we have, for instance,

$$\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon} = \sum_{i,j=1}^N \left(\sqrt{\omega_i}\mathbf{K}_{i,j}\sqrt{\upsilon_j}\right)^2 = \left\|\mathbf{W}^{1/2}\mathbf{K}\mathbf{V}^{1/2}\right\|_F^2,$$

where $\|\cdot\|_F$ stands for the Frobenius norm.

In particular, in the $\{0,1\}$-*sampling case*, i.e., assuming that $\boldsymbol{\omega} = \mathbb{1}$ and that the components of $\boldsymbol{\upsilon}$ are either 0 or 1 (so that the components of $\boldsymbol{\omega}-\boldsymbol{\upsilon}$ are also either 0 or 1), and introducing the index sets $I = \{i|\upsilon_i > 0\}$ and $I^c = \{1,\cdots,N\}\backslash I = \{i|\upsilon_i = 0\}$, we can remark that

$$(\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}) = \left\|(\mathrm{Id}_N - \mathbf{V})\mathbf{K}(\mathrm{Id}_N - \mathbf{V})\right\|_F^2 = \left\|\mathbf{K}_{I^c,I^c}\right\|_F^2,$$

where $\mathbf{K}_{I^c,I^c}$ stands for the principal submatrix of $\mathbf{K}$ defined by the index set I^c. In this framework, if we fix to $n < N$ the number of landmarks (i.e., the number of components of $\boldsymbol{\upsilon}$ equal to 1), minimising the squared-kernel discrepancy thus amounts in searching for the $(N-n)\times(N-n)$ principal submatrix of $\mathbf{K}$ with the smallest Frobenius norm (the principal submatrix $\mathbf{K}_{I^c,I^c}$ is indeed "omitted" by the approximation process). ◁

Following Section 3, we now illustrate how to compute the approximate eigendecomposition of the matrix $\mathbf{KW}$ related to T_μ induced by the matrix $\mathbf{KV}$ related to T_ν.

We assume that $\boldsymbol{\upsilon} \neq 0$ and we introduce the index set $I = \{i|\upsilon_i > 0\}$; 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 discard the points x_k such that $\upsilon_k = 0$); following Section 4.1, the strictly positive eigenvalues $\{\vartheta_l\}_{l\in\mathbb{I}_\nu^+}$ of T_ν and the associated canonically extended eigenfunctions $\psi_l \in \mathcal{H}$, orthonormalised for $L^2(\nu)$, can be obtained from the eigendecomposition of the $n\times n$ (symmetric and positive-semidefinite) principal submatrix $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I}$, i.e., the principal submatrix of $\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}$ defined by the index set I. Notice that since $\mathbf{V}$ is diagonal, we have $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I} = \mathbf{V}_{I,I}^{1/2}\mathbf{K}_{I,I}\mathbf{V}_{I,I}^{1/2}$. Let $\mathbf{a}_l \in \mathbb{R}^n$, with $l \in \mathbb{I}_\nu^+$, be a set of eigenvectors, orthonormalised in $\{\mathbb{R}^n, (\cdot|\cdot)_{\mathrm{Id}_n}\}$, associated with the strictly positive eigenvalues $\{\vartheta_l\}_{l\in\mathbb{I}_\nu^+}$ of $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I}$. Introducing the $N\times n$ matrix $\mathbf{K}_{\bullet,I}$ defined by the n columns of $\mathbf{K}$ with index in I, the canonically extended eigenvectors $\mathbf{u}_l$ of $\mathbf{KV}$ are given by

$$\mathbf{u}_l = \left(\psi_l(x_1),\cdots,\psi_l(x_N)\right)^T = \tfrac{1}{\vartheta_l}\mathbf{K}_{\bullet,I}\mathbf{V}_{I,I}(\mathbf{V}_{I,I})^{-1/2}\mathbf{a}_l = \tfrac{1}{\vartheta_l}\mathbf{K}_{\bullet,I}\mathbf{V}_{I,I}^{1/2}\mathbf{a}_l;$$

they satisfy $\mathbf{K}\mathbf{V}\mathbf{u}_l = \vartheta_l \mathbf{u}_l$ and $\mathbf{u}_l^T \mathbf{V} \mathbf{u}_{l'} = \delta_{l,l'}$. Notice that $[\mathbf{u}_l]_I = (\mathbf{V}_{I,I})^{-1/2}\mathbf{a}_l$ (where $[\mathbf{u}_l]_I \in \mathbb{R}^n$ consists in the components of $\mathbf{u}_l$ with index in I).

For all $l \in \mathbb{I}_\nu^+$, we have $\|\psi_l\|^2_{L^2(\mu)} = \|\mathbf{u}_l\|^2_{\mathbf{W}} = \mathbf{u}_l^T \mathbf{W} \mathbf{u}_l$, and the induced *normalised approximate eigenvectors* of $\mathbf{K}\mathbf{W}$ are given by (we have $\|\psi_l\|^2_{L^2(\mu)} > 0$, since $\mathcal{H}_\nu \subset \mathcal{H}_\mu$)

$$\widehat{\mathbf{v}}_l = \big(\widehat{\varphi}_l(x_1), \cdots, \widehat{\varphi}_l(x_N)\big)^T = \mathbf{u}_l / \|\mathbf{u}_l\|_{\mathbf{W}}.$$

Starting from a pair $\{\vartheta_l, (\mathbf{V}_{I,I})^{-1/2}\mathbf{a}_l\}$, the amount of computations required to obtain the extended components of the eigenvector $\mathbf{u}_l$ scales as $\mathcal{O}(n(N-n))$. The measure μ being supported by N points, computing an inner product in $L^2(\mu)$ requires $\mathcal{O}(N)$ operations. Following Remark 3.2, obtaining the normalised approximate eigenvector $\widehat{\mathbf{v}}_l$ and the approximate eigenvalue $\widehat{\lambda}_l^{[1]}$ is therefore relatively inexpensive. In order to obtain $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ or $\widehat{\lambda}_l^{[4]}$, we need to compute

$$\mathbf{K}\mathbf{W}\mathbf{u}_l = \big(T_\mu[\psi_l](x_1), \cdots, T_\mu[\psi_l](x_N)\big)^T,$$

the complexity of the underlying matrix-vector product thus scales as $\mathcal{O}(N^2)$ and is therefore costly; this operation can nevertheless be easily parallelised.

4.4. Kernel-matrix approximation. In the framework of Section 4.3 (we use the notations introduced in this section), the approximate operator T_ν is related to the matrix $\mathbf{K}\mathbf{V}$ (and thus also to $\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}$, as discussed in Section 4.1); notice that since $\mathbf{V}$ is diagonal, $\mathbf{K}\mathbf{V}$ can be interpreted as a weighted sample of columns of $\mathbf{K}$.

Considering the reproducing kernel $K_\nu(\cdot,\cdot)$ of the subspace $\mathcal{H}_\nu$ of $\mathcal{H}$, see (3.5), and following Remark 4.1, we introduce the $N \times N$ kernel matrix $\mathbf{K}_\nu$ defined by $K_\nu(\cdot,\cdot)$ and $\mathcal{S}$, i.e., with i,j entry $[\mathbf{K}_\nu]_{i,j} = K_\nu(x_i, x_j)$. From the eigendecomposition $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I} = \mathbf{A}\mathbf{\Theta}\mathbf{A}^T$ (with $\mathbf{A}$ a $n \times n$ orthogonal matrix), we deduce that

$$\mathbf{K}_\nu = \textstyle\sum_{l \in \mathbb{I}_\nu^+} \vartheta_l \mathbf{u}_l \mathbf{u}_l^T = \sum_{l \in \mathbb{I}_\nu^+} \hat{\lambda}_l^{[1]} \hat{\mathbf{v}}_l \hat{\mathbf{v}}_l^T = \mathbf{K}_{\bullet,I} \mathbf{V}_{I,I}^{1/2} \mathbf{A} \mathbf{\Theta}^\dagger \mathbf{A}^T \mathbf{V}_{I,I}^{1/2} \mathbf{K}_{I,\bullet},$$

with $\mathbf{K}_{I,\bullet} = \mathbf{K}_{\bullet,I}^T$, and where $\mathbf{\Theta}^\dagger$ is the Moore-Penrose generalised inverse of the diagonal matrix $\mathbf{\Theta}$, see for instance [2] (i.e., $\mathbf{\Theta}^\dagger$ is the diagonal matrix whose diagonal entries are the generalised inverses of the eigenvalues of $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I}$; that is $1/\vartheta_m$ if $\vartheta_m > 0$, and 0 if $\vartheta_m = 0$). The matrix $\mathbf{A}\mathbf{\Theta}^\dagger\mathbf{A}^T$ is therefore the Moore-Penrose generalised inverse of $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I}$; since the matrix $\mathbf{V}$ is diagonal and by definition of the index set I, we also obtain

$$\mathbf{K}_\nu = \mathbf{K}_{\bullet,I} \mathbf{V}_{I,I}^{1/2} \big([\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I}\big)^\dagger \mathbf{V}_{I,I}^{1/2} \mathbf{K}_{I,\bullet} = \mathbf{K}\mathbf{V}^{1/2}\big(\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}\big)^\dagger \mathbf{V}^{1/2}\mathbf{K},$$

and in particular, $\mathbf{V}^{1/2}\mathbf{K}_\nu\mathbf{V}^{1/2} = \mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}$. Following for instance [6, 9], the matrix $\mathbf{K}_\nu$ corresponds to the *Nyström approximation* of the kenel matrix $\mathbf{K}$ induced by the approximate operator T_ν (i.e., induced by the weighed column-sample defined by $\boldsymbol{\upsilon}$). Low-rank approximations of $\mathbf{K}_\nu$ can classically be obtained by spectral truncation, i.e., by considering a subset $\mathbb{I}_{\nu,trc}^+$ of $\mathbb{I}_\nu^+$ (the truncation subset usually corresponds to the largest eigenvalues of T_ν), and by defining $\mathbf{K}_{\nu,trc} = \sum_{l \in \mathbb{I}_{\nu,trc}^+} \vartheta_l \mathbf{u}_l \mathbf{u}_l^T$; in practice, in view of Section 3, one should in this case favour a truncation subset corresponding to accurately approximate eigendirections.

Notice that, for $\boldsymbol{\omega} = \mathbb{1}$, the approximate eigenpairs $\{\hat{\lambda}_l^{[1]}, \hat{\mathbf{v}}_l\}$ correspond to approximations of the eigenpairs of $\mathbf{K}\mathbf{W} = \mathbf{K}$; in this case, the matrix $\mathbf{K}_{\nu,trc}$ therefore consists in an approximation of the low-row rank approximation of $\mathbf{K}$ obtained by spectral truncation (i.e., obtained by truncating the spectrum of $\mathbf{K}$, see, e.g., [6, 9]).

5. Optimal quadrature-sparsification as quadratic programming. We consider the framework of Section 4.3; notice that we could as well consider the framework of Section 4.2 (and in this case, in particular, $\mathbf{S}\boldsymbol{\omega}$ is replaced by $\mathbf{g}_\mu$). Following (4.3), for a fixed discrete measure μ supported by $\mathcal{S}$ (i.e., $\boldsymbol{\omega} > 0$ is fixed), we define, for $\boldsymbol{\upsilon} \in \mathbb{R}^N$ (and in practice $\boldsymbol{\upsilon} \geqslant 0$),

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

the factor 1/2 being added for simplification purpose. To promote sparsity of the approximate measure and discard the trivial minimun at $\boldsymbol{\upsilon} = \boldsymbol{\omega}$, we now introduce squared-kernel-discrepancy-minimisation problems involving an ℓ^1-type penalisation.

5.1. Regularised squared-kernel-discrepancy minimisation. For a given *penalisation direction* $\mathbf{d} = (d_1, \cdots, d_N)^T \in \mathbb{R}^N$, with $\mathbf{d} > 0$ (see Section 9 for a discussion on the choice of relevant penalisation directions), and $\alpha \geqslant 0$, we introduce the minimisation problem, for $\boldsymbol{\upsilon} \in \mathbb{R}^N$,

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D_\alpha(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) + \alpha \mathbf{d}^T \boldsymbol{\upsilon} \text{ subject to } \boldsymbol{\upsilon} \geqslant 0. \tag{5.1}$$

A solution to (5.1) always exists (see for instance Section 5.2); we also recall that, for a given $\alpha \geqslant 0$, the set of all solutions is convex. The gradient of $D_\alpha(\cdot)$ at $\boldsymbol{\upsilon} \in \mathbb{R}^N$ is $\nabla D_\alpha(\boldsymbol{\upsilon}) = \mathbf{S}(\boldsymbol{\upsilon} - \boldsymbol{\omega}) + \alpha \mathbf{d}$.

Proposition 5.1. *Denote by $\boldsymbol{\upsilon}_\alpha^*$ a solution to* (5.1) *with $\alpha \geqslant 0$, we have:*

(a) for $\alpha = 0$, $\boldsymbol{\upsilon}_\alpha^* = \boldsymbol{\omega}$ is a solution to (5.1)*,*

(b) 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 $\mathbf{S}\boldsymbol{\omega}$),

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

(d) $\nabla D_\alpha(\boldsymbol{\upsilon}_\alpha^*) \geqslant 0$ and $(\boldsymbol{\upsilon}_\alpha^*)^T \nabla D_\alpha(\boldsymbol{\upsilon}_\alpha^*) = 0$,

(e) if $\widetilde{\boldsymbol{\upsilon}}_\alpha^*$ is another solution to (5.1)*, then $\mathbf{S}\widetilde{\boldsymbol{\upsilon}}_\alpha^* = \mathbf{S}\boldsymbol{\upsilon}_\alpha^*$ and $\mathbf{d}^T \widetilde{\boldsymbol{\upsilon}}_\alpha^* = \mathbf{d}^T \boldsymbol{\upsilon}_\alpha^*$,*

(f) if $[\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}]_k > 0$, or if $[\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}]_k = 0$ and $\mathbf{S}_{k,k} > 0$ (see Remark 5.1), then $[\boldsymbol{\upsilon}_\alpha^*]_k = 0$,

(g) the maps $\alpha \mapsto D(\boldsymbol{\upsilon}_\alpha^*)$, and $\alpha \mapsto D_\alpha(\boldsymbol{\upsilon}_\alpha^*)$ are increasing,

(h) the maps $\alpha \mapsto \mathbf{d}^T \boldsymbol{\upsilon}_\alpha^*$, and $\alpha \mapsto (\boldsymbol{\upsilon}_\alpha^*)^T \mathbf{S} \boldsymbol{\upsilon}_\alpha^*$, and $\alpha \mapsto \boldsymbol{\omega}^T \mathbf{S} \boldsymbol{\upsilon}_\alpha^*$ are decreasing.

Remark 5.1. Assuming $\mathbf{S}_{k,k} = K^2(x_k, x_k) > 0$ for all $k \in \{1, \cdots, N\}$ (what we shall 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 $K(x, x) = \|K(x, \cdot)\|_{\mathcal{H}}^2 \geqslant 0$. This assumption is thus non-restrictive: indeed, if $K(x_k, x_k) = 0$, then $h(x_k) = 0$ for all $h \in \mathcal{H}$; if μ and ν are supported by $\mathcal{S}$ (Section 4.3), then such a point x_k can be removed from $\mathcal{S}$ without inducing any modification of the operators T_μ and T_ν. ◁

Since $\boldsymbol{\upsilon} \geqslant 0$, the term $\mathbf{d}^T \boldsymbol{\upsilon}$ can be interpreted as a weighted ℓ^1-type regularisation (and α is the regularisation parameter). For appropriate $\mathbf{d}$ and α, we can therefore expect a solution $\boldsymbol{\upsilon}_\alpha^*$ to (5.1) to be sparse, and sparsity of the solutions should tend to increase with α (see, e.g., [12]). This intuition is confirmed by Proposition 5.1-(f), which shows that the number of strictly positive components of $\boldsymbol{\upsilon}_\alpha^*$ is bounded from above by the number of negative components of $\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}$ (this bound is however generally not tight).

5.2. Constrained squared-kernel-discrepancy minimisation. Instead of considering (5.1), for $\varkappa \geqslant 0$ (and, in practice, $\varkappa \leqslant \mathbf{d}^T \boldsymbol{\omega}$, see Proposition 5.2), we can equivalently introduce, for $\boldsymbol{\upsilon} \in \mathbb{R}^N$,

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) \text{ subject to } \boldsymbol{\upsilon} \geqslant 0 \text{ and } \mathbf{d}^T \boldsymbol{\upsilon} = \varkappa. \tag{5.2}$$

Notice that problem (5.2) consists in minimising a convex function on a convex compact domain.

Proposition 5.2. *Let $\boldsymbol{\upsilon}_\alpha^*$ be a solution to problem* (5.1) *with $\alpha \geqslant 0$; then $\boldsymbol{\upsilon}_\alpha^*$ is a solution to problem* (5.2) *with $\varkappa = \mathbf{d}^T \boldsymbol{\upsilon}_\alpha^*$. Reciprocally, assume that $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to problem* (5.2) *with $0 < \varkappa \leqslant \mathbf{d}^T \boldsymbol{\omega}$; then $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to problem* (5.1) *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* (5.1) *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 maps $\varkappa \mapsto D(\boldsymbol{\upsilon}_\varkappa^*)$ and $\varkappa \mapsto (\boldsymbol{\upsilon}_\varkappa^*)^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}_\varkappa^*)/\varkappa$ are decreasing.*

Remark that, in view of Proposition 5.2, if $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to problem (5.2) with $0 \leqslant \varkappa \leqslant \mathbf{d}^T \boldsymbol{\omega}$, then $\boldsymbol{\upsilon}_\varkappa^*$ is also solution to

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) \text{ subject to } \boldsymbol{\upsilon} \geqslant 0 \text{ and } \mathbf{d}^T \boldsymbol{\upsilon} \leqslant \varkappa. \tag{5.3}$$

Problem (5.2) can be efficiently solved thanks to a sparse-descent-direction QP solver (and without storing the matrix $\mathbf{S}$), like for instance the vertex-exchange strategy, see [17, Chap. 9] and Section 8.1. A sequential strategy (based on the notion of regularisation path) for solving problems (5.1) and (5.2) is discussed in Section 7

5.3. Penalisation and conic squared-kernel discrepancy. We now investigate the properties of the solutions to penalised squared-kernel-discrepancy minimisation problems in the light of Theorem 3.2 (i.e., in terms of conic squared-kernel discrepancy).

We consider the solutions $\boldsymbol{v}_\alpha^*$ to (5.1) for $\alpha \geqslant 0$; results related to the solutions to (5.2) for $0 \leqslant \varkappa \leqslant \mathbf{d}^T\boldsymbol{\omega}$ can be obtained readily through Proposition 5.2. Following Theorem 3.2, for $\boldsymbol{v}_\alpha^* \neq 0$, we denote by c_α the argument of the minimum of the function $c \mapsto D(c\boldsymbol{v}_\alpha^*)$. From Proposition 5.1-(e), we can remark that, even in case of non-uniqueness of the solution, c_α and $D(c_\alpha \boldsymbol{v}_\alpha^*)$ are unique.

Theorem 5.1. *For $0 \leqslant \alpha < \max_k \left\{[\mathbf{S}\boldsymbol{\omega}]_k/d_k\right\} = \alpha_0$, we have $\mathbf{d}^T\boldsymbol{v}_\alpha^* \leqslant c_\alpha \mathbf{d}^T\boldsymbol{v}_\alpha^* \leqslant \mathbf{d}^T\boldsymbol{\omega}$; in addition, if the map $\alpha \mapsto \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_\alpha^*/\mathbf{d}^T\boldsymbol{v}_\alpha^*$ is increasing on the interval $[0, \alpha_0)$, then the maps $\alpha \mapsto D(c_\alpha\boldsymbol{v}_\alpha^*)$ and $\alpha \mapsto c_\alpha\mathbf{d}^T\boldsymbol{v}_\alpha^*$ are respectively increasing and deceasing on this interval.*

Theorem 5.1 thus gives a sufficient condition for the conic squared-kernel discrepancy of the solutions to the regularised problem (5.1) to increase with the regularisation parameter α; in combination with Proposition 5.1-(f), this result therefore shows that increasing the amount of penalisation tends to increase the sparsity of the approximate measure, at the expense of reducing the overall accuracy of the induced spectral approximation; see Sections 10 and 11 for illustrations. This sufficient condition is further discussed in Section 7.1, notice for instance that it is always verified when the matrix $\mathbf{S}$ is non-singular.

6. Analogy with one-class SVM. Following for instance [18], problems (5.1) and (5.2) can be interpreted as the dual formulations of *one-class distorted SVMs* (or discrepancy-SVMs) related to the squared kernel, the initial discrete measure μ and the penalisation direction $\mathbf{d}$. Soft-margin-type extensions of the one-class SVMs introduced in Sections 6.1 and 6.2 are described in Appendix B.

We recall that we denote by $\mathcal{G}$ the RKHS associated with the squared kernel $K^2(\cdot,\cdot)$, and that for $\mu \in \mathscr{T}(K)$, the function $g_\mu \in \mathcal{G}$ is defined as $g_\mu(x) = \int_{\mathscr{X}} K^2(t,x)\mathrm{d}\mu(t)$, see Lemma 2.1.

6.1. One-class SVM related to the regularised problem. We first describe the SVM related to problem (5.1) with $\alpha \geqslant 0$. For $g \in \mathcal{G}$, we consider the convex minimisation problem

$$\begin{array}{ll} \underset{g}{\text{minimise}} & \frac{1}{2}\|g\|_{\mathcal{G}}^2 + (g|g_\mu)_{\mathcal{G}} \\ \text{subject to} & g(x_k) \geqslant -\alpha d_k \text{ for all } k \in \{1,\cdots,N\}. \end{array} \tag{6.1}$$

We shall refer to g_μ as the *primal distortion term*; we recall that, in (5.1), $\mu = \sum_{k=1}^N \omega_k \delta_{x_k}$. The application $g \mapsto \|g\|_{\mathcal{G}}^2$ being strictly convex, a solution to problem (6.1) is necessarily unique.

Proposition 6.1. *If $\boldsymbol{v}_\alpha^*$ is a solution to* (5.1) *with $\alpha \geqslant 0$, then $g_\alpha^*(x) = \sum_{k=1}^N [\boldsymbol{v}_\alpha^* - \boldsymbol{\omega}]_k K^2(x,x_k)$ is the solution to* (6.1)*. For all $k \in \{1,\cdots,N\}$ such that $[\boldsymbol{v}_\alpha^*]_k > 0$, we have $g_\alpha^*(x_k) = -\alpha d_k$.*

Notice that for all k, we have $g_\alpha^*(x_k) = [\mathbf{S}(\boldsymbol{v}_\alpha^* - \boldsymbol{\omega})]_k$. By introducing the change of variable $\check{g} = g + g_\mu \in \mathcal{G}$, problem (6.1) leads to (up to an additive constant)

$$\begin{array}{ll} \underset{\check{g}}{\text{minimise}} & \frac{1}{2}\|\check{g}\|_{\mathcal{G}}^2 \\ \text{subject to} & \check{g}(x_k) \geqslant g_\mu(x_k) - \alpha d_k \text{ for all } k \in \{1,\cdots,N\}, \end{array} \tag{6.2}$$

which is an equivalent formulation for (6.1), with solution $\check{g}_\alpha^*(x) = \sum_{k=1}^N [\boldsymbol{v}_\alpha^*]_k K^2(x,x_k)$. In view of Lemma 2.1, if we denote by ν_α^* the discrete measure supported by S related to a solution $\boldsymbol{v}_\alpha^*$ to problem (5.1), then $\check{g}_\alpha^* = g_{\nu_\alpha^*}$.

Remark 6.1. Problem (6.1) may also be considered in the framework of Section 4.2, i.e., for a general measure $\mu \in \mathscr{T}(K)$; however, the optimisation needs in this case to be restricted to the closed linear subspace $\mathcal{G}_S = \mathrm{span}\{K_{x_k}^2\}_{k=1}^N$ of $\mathcal{G}$. ◁

6.2. One-class SVM related to the constrained problem. We now describe the SVM related to problem (5.2) with $\varkappa > 0$. For $g \in \mathcal{G}$ and $\gamma \in \mathbb{R}$, we introduce the problem

$$\begin{array}{ll} \underset{g,\gamma}{\text{minimise}} & \frac{1}{2}\|g\|_{\mathcal{G}}^2 + (g|g_\mu)_{\mathcal{G}} - \gamma \\ \text{subject to} & g(x_k) \geqslant \gamma d_k/\varkappa \text{ for all } k \in \{1,\cdots,N\}. \end{array} \tag{6.3}$$

Again, a solution to problem (6.3) is necessarily unique.

Proposition 6.2. *If $\boldsymbol{v}^*_\varkappa$ is a solution to* (5.2)*, then $g^*_\varkappa(x) = \sum_{k=1}^N [\boldsymbol{v}^*_\varkappa - \boldsymbol{\omega}]_k K^2(x, x_k)$ and $\gamma^*_\varkappa = (\boldsymbol{v}^*_\varkappa)^T \mathbf{S}(\boldsymbol{v}^*_\varkappa - \boldsymbol{\omega})$ is the solution to* (6.3)*. For all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{v}^*_\varkappa]_k > 0$, we have $g^*_\varkappa(x_k) = \gamma^*_\varkappa d_k/\varkappa$.*

In view of Proposition 5.2, for $0 < \varkappa \leqslant \mathbf{d}^T\boldsymbol{\omega}$, we know that $\boldsymbol{v}^*_\varkappa$ is a solution to (5.1) for $\alpha = -\gamma^*_\varkappa/\varkappa$; since $\alpha \geqslant 0$, we therefore have $\gamma^*_\varkappa \leqslant 0$.

7. Regularisation path. In this section, we further discuss the properties the solutions to problem (5.1); following for instance [16, 11], we also describe the regularisation-path method (or homotopy method) for solving the regularised problem. Results related to the constrained problem (5.2) can be obtained from Proposition 5.2.

7.1. Generalities. Let $\boldsymbol{v}^*_\alpha$ be a solution to (5.1) 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, from Proposition 5.1, $[\nabla D_\alpha(\boldsymbol{v}^*_\alpha)]_k > 0$ for all $k \in J^c_\alpha$; in addition, the index set J_α is unique (i.e., for a given α, in case of non-uniqueness of the solution to (5.1), J_α does not depends on the solution considered). We shall refer to J_α as the *sparsity pattern* of the solutions to problem (5.1) for $\alpha \geqslant 0$; in particular, from Proposition 5.1-(d), if $[\boldsymbol{v}^*_\alpha]_k > 0$, then k necessarily belongs to J_α. Knowing J_α, a solution to (5.1) is characterised by the conditions

$$\boldsymbol{v}^*_\alpha \geqslant 0, \text{ and } [\boldsymbol{v}^*_\alpha]_{J^c_\alpha} = 0, \text{ and } \mathbf{S}_{J_\alpha, J_\alpha}[\boldsymbol{v}^*_\alpha]_{J_\alpha} = [\mathbf{S}\boldsymbol{\omega} - \alpha\mathbf{d}]_{J_\alpha}, \tag{7.1}$$

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.

Proposition 7.1. *Let $\boldsymbol{v}^*_{\alpha_1}$ and $\boldsymbol{v}^*_{\alpha_2}$ be solutions to problem* (5.1) *with α_1 and $\alpha_2 \geqslant 0$, respectively. Assume that $J_{\alpha_1} = J_{\alpha_2} = J$, 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* (5.1) *with $\alpha = \theta\alpha_1 + (1-\theta)\alpha_2$, and $J_\alpha = J$.*

Proposition 7.1 therefore shows that the set of all solutions $\boldsymbol{v}^*_\alpha$ related to a same sparsity pattern $J \subset \{1, \cdots, N\}$ is convex, and that the values of α such that $J_\alpha = J$ belongs to a convex interval. When α varies, we refer to a change in the sparsity pattern J_α as an *event*; in particular, since there cannot exist more than 2^N different subsets of $\{1, \cdots, N\}$, Proposition 7.1 implies that the number M_{ev} of events related to problem (5.1) necessarily satisfies $M_{ev} \leqslant 2^N - 1$. We also call *kinks* the values of α where an event occurs; more precisely, the kinks consist in the strictly positive infima and suprema of the intervals of α related to a same sparsity pattern (see Remark 7.1).

Remark 7.1. Assume that $\alpha > 0$ is a kink for (5.1); for all $\epsilon > 0$ such that $\alpha - \epsilon \geqslant 0$, we therefore have $J_{\alpha+\epsilon} \neq J_{\alpha-\epsilon}$. We refer to the limits when ϵ tends to 0 of $J_{\alpha+\epsilon}$ and $J_{\alpha-\epsilon}$ as the *right and left sparsity patterns* at α, denoted by R_α and L_α, respectively. The "true" sparsity pattern J_α at a kink α is either its left or right sparsity pattern (if α is not a kink, the left and right sparsity patterns are identical). In particular, since a change in the sparsity pattern only involves null components of $\boldsymbol{v}^*_\alpha$, if α is a kink for (5.1), then (7.1) is true for both the left and right sparsity patterns at α; in other words, if α is a kink, then in (7.1), we may replace J_α by R_α or L_α. ◁

We assume that the events occurs at the kinks $\alpha_0 > \alpha_1 > \cdots > \alpha_{M_{ev}-1} > 0$. From Proposition 5.1, for $\alpha \geqslant \max_k [\mathbf{S}\boldsymbol{\omega}]_k/d_k$, we have $\boldsymbol{v}^*_\alpha = 0$. We can therefore deduce that $\alpha_0 = \max_k [\mathbf{S}\boldsymbol{\omega}]_k/d_k$, and that $J_{\alpha_0} = \{k | [\mathbf{S}\boldsymbol{\omega}]_k/d_k = \alpha_0\}$; for $\alpha > \alpha_0$, the sparsity pattern J_α is the empty set, and J_{α_0} is thus also the left sparsity pattern at the kink α_0. Since $\nabla D(\boldsymbol{\omega}) = 0$, the kink $\alpha_{M_{ev}-1}$ is the supremum of the set of all α such that $J_\alpha = \{1, \cdots, N\}$, and $\{1, \cdots, N\}$ is thus also the left sparsity pattern at the kink $\alpha_{M_{ev}-1}$. More generally, for all $\alpha \in (\alpha_{p+1}, \alpha_p)$, with $p \in \{0, \cdots, M_{ev} - 2\}$, we have $J_\alpha = L_{\alpha_p} = R_{\alpha_{p+1}}$, where L_{α_p} stands for the left sparsity pattern at the kink α_p, and $R_{\alpha_{p+1}}$ is the right sparsity pattern at the kink α_{p+1}, see Remark 7.1.

We conclude this section with a result related to the sufficient condition appearing in Theorem 5.1.

Proposition 7.2. *For $\alpha \geqslant 0$, the maps $\alpha \mapsto \mathbf{d}^T\boldsymbol{v}^*_\alpha$ and $\alpha \mapsto \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}^*_\alpha$ are continuous and piecewise linear. In addition, if for all $\alpha \in [0, \alpha_0)$ such that α is not a kink for* (5.1)*, there exists a solution $\boldsymbol{v}^*_\alpha$ such that $[\boldsymbol{v}^*_\alpha]_{J_\alpha} = \mathbf{G}[\mathbf{S}\boldsymbol{\omega} - \alpha\mathbf{d}]_{J_\alpha}$, with $\mathbf{G}$ symmetric and positive-semidefinite, then the map $\alpha \mapsto \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}^*_\alpha/\mathbf{d}^T\boldsymbol{v}^*_\alpha$ is increasing on the interval $[0, \alpha_0)$.*

The sufficient condition in Proposition 7.2 is in particular always verified when the matrix $\mathbf{S}$ is non-singular. Indeed, any principal submatrix of a symmetric and positive-definite matrix is also symmetric and positive-definite; in addition, for any $\alpha \geqslant 0$, since the the solutions to (5.1) are in this case unique, we then necessarily have $\mathbf{G} = (\mathbf{S}_{J_\alpha,J_\alpha})^{-1}$.

7.2. Computing the path. The regularisation-path method consists in iteratively computing the kinks α_0, α_1, etc., while keeping track of the evolution of the sparsity pattern of the solutions to (5.1).

Hereafter, we consider a kink α_p, for $p \in \{0, \cdots, M_{ev} - 2\}$, with related left sparsity pattern L_{α_p} (see Remark 7.1). We describe how to compute the next kink $\alpha_{p+1} < \alpha_p$, and how to characterise the related left sparsity pattern $L_{\alpha_{p+1}}$ (notice that, by definition, $R_{\alpha_{p+1}} = L_{\alpha_p}$). For simplicity, we use the notation $J = L_{\alpha_p}$, and we assume that the submatrix $\mathbf{S}_{J,J}$ is invertible (numerical strategies to deal with singular submatrices exist, but they are out of the scope of this study).

From (7.1), we introduce the vector $\boldsymbol{\upsilon}_\alpha$ such that $[\boldsymbol{\upsilon}_\alpha]_{J^c} = 0$ and $[\boldsymbol{\upsilon}_\alpha]_J = (\mathbf{S}_{J,J})^{-1}[\mathbf{S}\boldsymbol{\omega} - \alpha\mathbf{d}]_J$. By definition, α_{p+1} corresponds to the smallest α such that $0 \leqslant \alpha < \alpha_p$ and

$$[\boldsymbol{\upsilon}_\alpha]_J \geqslant 0 \text{ and } [\mathbf{S}(\boldsymbol{\upsilon}_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{d}]_{J^c} \geqslant 0. \tag{7.2}$$

Lemma 7.1. *Consider a kink α_p with left sparsity pattern $L_{\alpha_p} = J \neq \{1, \cdots, N\}$, and assume that the submatrix $\mathbf{S}_{J,J}$ is invertible. We introduce the $(N - n_{\alpha_p}) \times n_{\alpha_p}$ matrix $\mathbf{M} = \mathbf{S}_{J^c,J}(\mathbf{S}_{J,J})^{-1}$, with $n_{\alpha_p} = \mathrm{card}(L_{\alpha_p})$, and we define*

$$\alpha_+ = \max_l \left\{ \left[\mathbf{M}[\mathbf{S}\boldsymbol{\omega}]_J - [\mathbf{S}\boldsymbol{\omega}]_{J^c}\right]_l / \left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l \middle| \left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l < 0 \right\}, \text{ and}$$
$$\alpha_- = \max_m \left\{ \left[(\mathbf{S}_{J,J})^{-1}[\mathbf{S}\boldsymbol{\omega}]_J\right]_m / \left[(\mathbf{S}_{J,J})^{-1}\mathbf{d}_J\right]_m \middle| \left[(\mathbf{S}_{J,J})^{-1}\mathbf{d}_J\right]_m < 0 \right\}.$$

The next event then occurs at $\alpha_{p+1} = \max\{\alpha_+, \alpha_-\}$. If $\alpha_{p+1} = \alpha_+$, then the indices in J^c corresponding to the maximum defining α_+ are transferred from J^c to $L_{\alpha_{p+1}}$; if $\alpha_{p+1} = \alpha_-$, then the indices in J corresponding to the maximum defining α_- are transferred from J to $L^c_{\alpha_{p+1}}$.

If $\mathbf{S}_{L_{\alpha_{p+1}},L_{\alpha_{p+1}}}$ is invertible, we can next compute α_{p+2} and $L_{\alpha_{p+2}}$ in exactly the same way, and we may potentially iterate like this until we reach the last event, or, at least, as far as we do not encounter numerical issues.

7.3. Computational complexity. The preliminary computation of the distortion term $\mathbf{S}\boldsymbol{\omega}$ is relatively challenging, with a worst-case complexity scaling as $\mathcal{O}(N^2)$; remark that the underlying matrix-vector product can nevertheless be very easily parallelised. Importantly, we shall not store the kernel matrix $\mathbf{S}$, but rather compute on the fly any required entry. Generally speaking, notice that obtaining the distortion term appears as the main bottleneck of the penalised squared-kernel-discrepancy-minimisation framework.

In view of Lemma 7.1, once a kink α_p and its left sparsity pattern $J = L_{\alpha_p}$ are known, defining the next event (i.e., computing α_{p+1} and $L_{\alpha_{p+1}}$) involves the calculation of $(\mathbf{S}_{J,J})^{-1}\mathbf{d}_J$ and $(\mathbf{S}_{J,J})^{-1}[\mathbf{S}\boldsymbol{\omega}]_J$ (i.e., solving 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 computational complexity of this task scales as $\mathcal{O}(n^3_{\alpha_p})$, with $n_{\alpha_p} = \mathrm{card}(L_{\alpha_p})$. Update formulae can nevertheless be used to reduce this complexity, by for instance iteratively updating the Cholesky decomposition of $\mathbf{S}_{J,J}$; in the favourable cases, the computational complexity may thus reduce to $\mathcal{O}(n^2_{\alpha_p})$; an alternative might also consist in using an indirect iterative approach, like for instance a conjugate gradient method (but numerical errors could then quickly lead to precision issues). Finally, the complexity of the two matrix-vector products involving the matrix $\mathbf{S}_{J^c,J}$ scales as $\mathcal{O}(n_{\alpha_p}(N - n_{\alpha_p}))$. As a result, the computation of the regularisation path becomes intractable once large values of n_{α_p} are reached. When N is large, the regularisation-path method may therefore only be used to explore the range of very sparse approximate measures ν. See Sections 10 and 11 for illustrations.

8. Numerical solver for the constrained problem. In this section, we discuss a strategy to compute approximate solutions to (5.2) (i.e., the constrained problem) for any value of the parameter $\varkappa > 0$. We also propose two greedy exchange-type strategies aiming at enhancing the sparsity of a given approximate measure while keeping the squared-kernel discrepancy as low as possible.

8.1. Vertex-exchange QP solver. Consider problem (5.2); for $\varkappa > 0$, we can define the change of variable $\widetilde{\boldsymbol{\upsilon}} = \mathbf{D}\boldsymbol{\upsilon}/\varkappa$, with $\mathbf{D} = \mathrm{diag}(\mathbf{d})$, (so that $\mathbf{d} = \mathbf{D}\mathbb{1}$). Problem (5.2) is thus turned into (up to an additive constant), for $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^N$,

$$\underset{\widetilde{\boldsymbol{\upsilon}}}{\text{minimise}}\ C(\widetilde{\boldsymbol{\upsilon}}) = \tfrac{1}{2}\widetilde{\boldsymbol{\upsilon}}^T\mathbf{A}\widetilde{\boldsymbol{\upsilon}} - \mathbf{b}^T\widetilde{\boldsymbol{\upsilon}} \text{ subject to } \widetilde{\boldsymbol{\upsilon}} \geqslant 0 \text{ and } \mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1, \tag{8.1}$$

with $\mathbf{A} = \varkappa^2\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$ and $\mathbf{b} = \varkappa\mathbf{D}^{-1}\mathbf{S}\boldsymbol{\omega}$. Since $\mathbf{A}_{i,j} = \varkappa^2 K^2(x_i, x_j)/(d_i d_j)$, any entry of $\mathbf{A}$ can be easily obtained from $\varkappa$, the squared kernel $K^2(\cdot,\cdot)$, the set S and the penalisation direction $\mathbf{d}$. Importantly, we shall therefore not store the matrix $\mathbf{A}$, but rather compute on the fly any required entry of $\mathbf{A}$; in this way, problems involving large N may be considered. Once $\mathbf{b}$ is known (requiring the knowledge of $\mathbf{S}\boldsymbol{\omega}$, see Section 7.3), the gradient $\nabla C(\widetilde{\boldsymbol{\upsilon}}) = \mathbf{A}\widetilde{\boldsymbol{\upsilon}} - \mathbf{b}$ can in particular be easily obtained for any sparse vector $\widetilde{\boldsymbol{\upsilon}}$.

The extreme points of the polytopes defined by the constraints in (8.1) 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 0). For a feasible $\widetilde{\boldsymbol{\upsilon}}$, let $I_{\widetilde{\boldsymbol{\upsilon}}} = \{k | \widetilde{\upsilon}_k > 0\}$ be the index set defined by the strictly positive components of $\widetilde{\boldsymbol{\upsilon}}$. An iteration of the vertex-exchange algorithm consists in searching

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

defining the sparse descent direction $\boldsymbol{\delta} = \mathbf{e}_{i^*} - \mathbf{e}_{j^*}$ (i.e., weight is transferred from the j^*-th to the i^*-th component of $\widetilde{\boldsymbol{\upsilon}}$); in case of non-uniqueness 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 ϱ being given by $\varrho = \min\left\{\widetilde{\upsilon}_{j^*}, -\left(\boldsymbol{\delta}^T\nabla C(\widetilde{\boldsymbol{\upsilon}})\right)/(\boldsymbol{\delta}^T\mathbf{A}\boldsymbol{\delta})\right\}$. In particular, since the descent direction $\boldsymbol{\delta}$ is sparse, the computation of the optimal step size is numerically affordable, and the same holds for the gradient update. Indeed, we have $\nabla C(\widetilde{\boldsymbol{\upsilon}} + \varrho\boldsymbol{\delta}) = \nabla C(\widetilde{\boldsymbol{\upsilon}}) + \varrho\mathbf{A}\boldsymbol{\delta}$, so that the gradient update involves only two columns of $\mathbf{A}$; the overall complexity therefore scales as $\mathcal{O}(N)$.

Denoting by $\widetilde{\boldsymbol{\upsilon}}^*$ a solution to (8.1), the convergence of the vertex-exchange algorithm can be easily verified (see, e.g., [10]) by simply remarking that since $\widetilde{\boldsymbol{\upsilon}} \geqslant 0$ and $\mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1$, by definition of j^*, we have $\widetilde{\boldsymbol{\upsilon}}^T\nabla C(\widetilde{\boldsymbol{\upsilon}}) \leqslant \mathbf{e}_{j^*}^T\nabla C(\widetilde{\boldsymbol{\upsilon}})$, and thus (distance from optimality)

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

In Sections 10 and 11, the accuracy of an approximate solution $\widetilde{\boldsymbol{\upsilon}}$ is indicated by the *Frank-Wolfe error bound* $\epsilon = (\widetilde{\boldsymbol{\upsilon}} - \mathbf{e}_{i^*})^T\nabla C(\widetilde{\boldsymbol{\upsilon}}) \geqslant 0$.

8.2. Enhancing sparsity through components merging. The formulation introduced in Section 8.1 offers a convenient framework to enhance the sparsity of an approximate measure ν while trying to keep its squared-kernel discrepancy as low as possible. Let $\widetilde{\boldsymbol{\upsilon}} \geqslant 0$ (with $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^N$) be such that $\mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1$. In practice, $\widetilde{\boldsymbol{\upsilon}}$ will be an exact or approximate solution to problem (8.1), or any vector related to an interesting low-squared-kernel-discrepancy configuration $\boldsymbol{\upsilon}$ through the change of variable $\widetilde{\boldsymbol{\upsilon}} = \mathbf{D}\boldsymbol{\upsilon}/\varkappa$, with $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ and $\varkappa = \mathbf{d}^T\boldsymbol{\upsilon}$, see Section 8.1. We assume that $\widetilde{\boldsymbol{\upsilon}}$ has $n = n_0$ strictly positive components and we introduce $I = \{i | \widetilde{\upsilon}_i > 0\}$. As illustrated in Sections 10 and 11, it is generally 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 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 $\nabla C(\widetilde{\boldsymbol{\upsilon}})$, the computation $C(\widetilde{\boldsymbol{\upsilon}}_{\{i,j\}})$ only involves four entries of the matrix $\mathbf{A}$ and two entries of $\nabla C(\widetilde{\boldsymbol{\upsilon}})$.

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

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

We thus obtain the "best" pairwise merging $\{i^*, j^*\}$ for $\widetilde{\boldsymbol{\upsilon}}$. We next update all the involved objects, i.e., $\widetilde{\boldsymbol{\upsilon}} \leftarrow \widetilde{\boldsymbol{\upsilon}}_{\{i^*,j^*\}}$, $I \leftarrow I\backslash\{j^*\}$, $n \leftarrow n-1$ and $\nabla C(\widetilde{\boldsymbol{\upsilon}}) \leftarrow \nabla C(\widetilde{\boldsymbol{\upsilon}}_{\{i^*,j^*\}})$, 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 tradeoff.

We thus obtain a sequence of merged vectors $\{\widetilde{\boldsymbol{\upsilon}}_{[0]}, \widetilde{\boldsymbol{\upsilon}}_{[1]}, \cdots\}$, where $\widetilde{\boldsymbol{\upsilon}}_0$ is our initial vector, $\widetilde{\boldsymbol{\upsilon}}_{[1]}$ results from the merging of two components of $\widetilde{\boldsymbol{\upsilon}}_{[0]}$, etc.; by construction, $\widetilde{\boldsymbol{\upsilon}}_{[m]} \geqslant 0$ and $\mathbb{1}^T\widetilde{\boldsymbol{\upsilon}}_{[m]} = 1$ for all m, and $\widetilde{\boldsymbol{\upsilon}}_{[m]}$ has $n_0 - m$ strictly positive components. Finally, instead of considering the approximation induced by $\boldsymbol{\upsilon} = \varkappa\mathbf{D}^{-1}\widetilde{\boldsymbol{\upsilon}}_{[0]}$, we may consider a sparser vector $\boldsymbol{\upsilon}_{[m]} = \varkappa\mathbf{D}^{-1}\widetilde{\boldsymbol{\upsilon}}_{[m]}$; see Sections 10 and 11 for illustrations.

9. Penalisation direction. In Section 5, the sparsity of the approximate measures is promoted through the introduction of an ℓ^1-type penalisation played by the term $\mathbf{d}^T\boldsymbol{\upsilon}$, for a given $\mathbf{d} \in \mathbb{R}^N$ with $\mathbf{d} > 0$. In practice, we aim at obtaining measures which are both as sparse as possible and with a low (conic) squared-kernel discrepancy, naturally raising questions related to the choice of the penalisation direction $\mathbf{d}$. The impact of the penalisation direction on the tradeoff between sparsity and (conic) squared-kernel discrepancy is illustrated in Sections 10.5 and 10.6.

Lemma 9.1 (Penalisation direction inducing no sparsity)**.** *If $\mathbf{d} = \theta\mathbf{S}\boldsymbol{\omega}$, with $\theta > 0$, then for $\alpha \leqslant 1/\theta$, $\boldsymbol{\upsilon}_\alpha^* = (1-\alpha\theta)\boldsymbol{\omega} \geqslant 0$ is a solution to* (5.1)*; for $\alpha > 1/\theta$, we have $\boldsymbol{\upsilon}_\alpha^* = 0$.*

Thus, for $\mathbf{d} = \theta\mathbf{S}\boldsymbol{\omega}$, the solutions to (5.1) are non-sparse, and such a choice for $\mathbf{d}$ is of no practical interest. More generally (and as a proof for Lemma 9.1), we can remark that if $\mathbf{d} = \mathbf{S}\boldsymbol{\eta} \geqslant 0$, with $\boldsymbol{\eta} \in \mathbb{R}^N$, then for all α such that $\boldsymbol{\omega} - \alpha\boldsymbol{\eta} \geqslant 0$, we have $\nabla D_\alpha(\boldsymbol{\omega} - \alpha\boldsymbol{\eta}) = 0$, and $\boldsymbol{\upsilon}_\alpha^* = \boldsymbol{\omega} - \alpha\boldsymbol{\eta}$ is therefore a solution to (5.1); notice that in the framework of Section 7.1, this situation corresponds to solutions with full sparsity parttern, i.e., $J_\alpha = \{1, \cdots, N\}$.

As illustrated in Sections 10 and 11, considering $\mathbf{d} = \mathbb{1}$ generally leads to satisfactory results; it is however also possible to consider problem-dependent penalisation directions, leading to interesting interpretations. Following Remark 5.1, we can reasonably assume that $\operatorname{diag}(\mathbf{K}) > 0$, so that, in particular, $\mathbf{S}\boldsymbol{\omega} > 0$ (since $\boldsymbol{\omega} > 0$). In the following Remarks 9.1, 9.2 and 9.3, we discuss specific penalisation directions defined from the vectors $\mathbf{S}\boldsymbol{\omega}$ and $\operatorname{diag}(\mathbf{K})$. Notice that, in practice, one should always check that the considered penalisation direction does not correspond to the pathological case described in Lemma 9.1 (for instance, if $\mathbf{K}$ is a circulant matrix and if $\boldsymbol{\omega} \propto \mathbb{1}$, then $\mathbf{S}\boldsymbol{\omega} \propto \operatorname{diag}(\mathbf{K}) \propto \mathbb{1}$).

We first recall that from Section 6, if $\boldsymbol{\upsilon}_\alpha^*$ is a solution to the regularised problem (5.1) (with related measure ν_α^*), then $g_{\nu_\alpha^*} = \sum_{k=1}^N [\boldsymbol{\upsilon}_\alpha^*]_k K_{x_k}^2$ is the solution to, for $g \in \mathcal{G}$,

$$\underset{g}{\text{minimise}}\ \tfrac{1}{2}\|g\|_{\mathcal{G}}^2 \text{ subject to } g_\mu(x_k) - g(x_k) \leqslant \alpha d_k \ \text{ for all } k \in \{1, \cdots, N\}, \tag{9.1}$$

with $g_\mu(x_k) = \int_{\mathscr{X}} K^2(t, x_k)\mathrm{d}\mu(t) = [\mathbf{S}\boldsymbol{\omega}]_k$; in addition, if $[\boldsymbol{\upsilon}_\alpha^*]_k > 0$, then $g_\mu(x_k) - g_{\nu_\alpha^*}(x_k) = \alpha d_k$.

Remark 9.1 (Inverse-distortion-based penalisation)**.** In view of (9.1), considering a penalisation direction $\mathbf{d}$ such that $d_k = 1/[\mathbf{S}\boldsymbol{\omega}]_k^p$, with $p > 0$, results in a SVM where the upper bound on $g_\mu(x_k) - g_\nu(x_k)$ is inversely proportional to a positive power of $g_\mu(x_k)$ (i.e., the larger $g_\mu(x_k)$, the smaller the bound on $g_\mu(x_k) - g_\nu(x_k)$). In addition, from (5.1), the positive components of $\boldsymbol{\upsilon}_\alpha^*$ are more likely to correspond to large values of $g_\mu(x_k) = [\mathbf{S}\boldsymbol{\omega}]_k$. Since for any measure ν supported by S (with related weights υ_k), we have $(g_\mu|g_\nu)_{\mathcal{G}} = \sum_{k=1}^N \upsilon_k g_\mu(x_k)$ (see Lemma 2.1), such a penalisation tends to promote large values of the inner product between g_μ and $g_{\nu_\alpha^*}$ in $\mathcal{G}$. ◁

Remark 9.2 (Inverse-kernel-diagonal-based penalisation)**.** For all $x \in \mathscr{X}$, from the reproducing property in $\mathcal{G}$ and the Cauchy-Schwarz inequality, we have

$$\forall \mu \text{ and } \nu \in \mathscr{T}(K),\ \left|g_\mu(x) - g_\nu(x)\right| \leqslant \sqrt{D_{K^2}(\mu,\nu)}K(x,x). \tag{9.2}$$

In view of (9.1) and (9.2), by considering a vector $\mathbf{d}$ such that $d_k = 1/(K(x_k,x_k))^p$, with $p > 0$, we enforce the bound on the difference $g_\mu(x_k) - g_{\nu_\alpha^*}(x_k)$ to be small at the points x_k where this difference can potentially be large, so that we can thus expect $g_\mu(x_k) - g_{\nu_\alpha^*}(x_k)$ to be relatively small for all the points in S. ◁

Remark 9.3 (Penalising the trace)**.** For $\mathbf{d} = \operatorname{diag}(\mathbf{K})$, we have $\mathbf{d}^T\boldsymbol{v} = \operatorname{trace}(T_\nu)$; by analogy with spectral truncation, from Proposition 5.1-(c), a solution $\boldsymbol{v}_\alpha^*$ to the regularised problem (5.1) then satisfies $\operatorname{trace}(T_{\nu_\alpha^*}) = \mathbf{d}^T\boldsymbol{v}_\alpha^* \leqslant \mathbf{d}^T\boldsymbol{\omega} = \operatorname{trace}(T_\mu)$. ◁

10. 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 [15]), as illustrated in Figure 10.2. We set $\boldsymbol{\omega} = \mathbb{1}/N$, so that the measure $\mu = \sum_k \omega_k \delta_{x_k}$ 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 $\ell = 1/0.16$ (a different kernel is considered in Section 10.6). An overview of the spectrum of the operator T_μ (obtained from the eigendecomposition of the matrix $\mathbf{K}/N$) is given in Figure 10.1. We first consider the penalisation direction $\mathbf{d} = \mathbb{1}$.

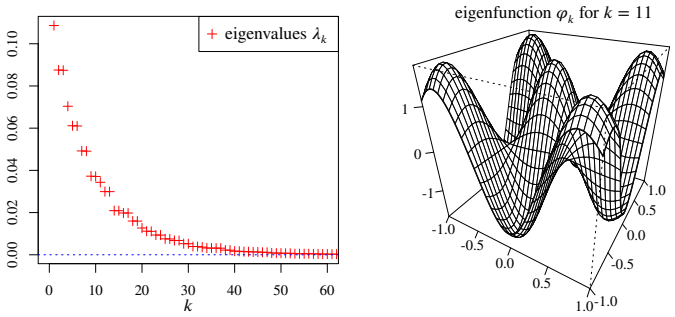

FIG. 10.1. *For the two-dimensional example (Gaussian kernel and $\boldsymbol{\omega} = \mathbb{1}/N$), eigenvalues λ_k of the integral operator T_μ (sorted in decreasing order, only the 62 largest eigenvalues are presented), and graph, on $[-1,1]^2$ of the canonically extended eigenfunction φ_k for $k = 11$.*

10.1. First experiment. Figure 10.2 shows the (approximate) solution $\boldsymbol{v}^*$ to problem (5.2) with $\varkappa = 0.81$, or equivalently, to problem (5.1) with $\alpha \approx 8.354214 \times 10^{-3}$ (with $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathbb{1}$). The vector $\boldsymbol{v}^*$ has 160 strictly positive components, and the support of the related measure ν^* inherits an interesting "four-concentric-squares" structure. We have $D(\boldsymbol{v}^*) = 7.631887 \times 10^{-4}$ (for comparison, notice that $D(\varkappa\mathbf{e}_1) = 3.041066 \times 10^{-1}$, with $\mathbf{e}_1$ the first element of the canonical basis of $\mathbb{R}^N$). In the framework of Section 8.1, the presented solution is related to a Frank-Wolfe error bound $\epsilon = 3.989864 \times 10^{-17}$.

The solution has been obtained using the regularisation-path strategy (see Section 10.2 for more details). Considering the regularisation path for problem (5.1) 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, for problem (5.2) (with increasing values of $\varkappa$), the value $\varkappa = 0.81$ satisfies

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

The accuracy of the approximate eigendecomposition of T_μ induced by the solution $\boldsymbol{v}^*$ presented in Figure 10.2 (i.e., $\varkappa = 0.81$) is illustrated in Figure 10.3. In view of the similarity between

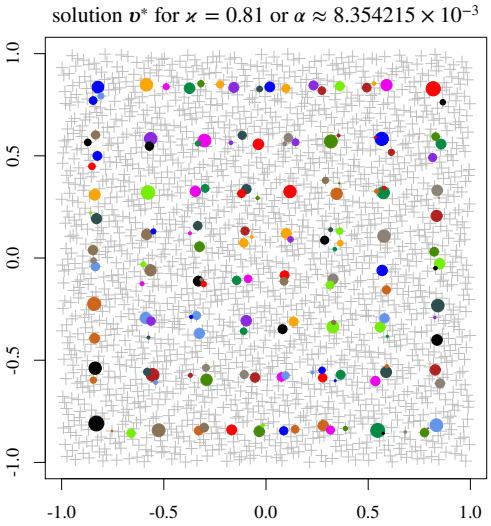

FIG. 10.2. *Graphical representation (two-dimensional example, Gaussian kernel, $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathbb{1}$) of the solution v^* to problem* (5.2) *with $\varkappa = 0.81$, or equivalently, to problem* (5.1) *with $\alpha \approx 8.354215 \times 10^{-3}$. The grey crosses represent the points in S and the filled dots are the strictly positive components of v^* (surface being proportional to v_k^*).*

the geometric approximate eigenvalues $\widehat{\lambda}_l^{[\cdot]}$, and more particularly of the ratios $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ and $\left(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]}\right)^2$ (see Section 3), we observe that the 21 main eigendirections of the operator T_{v^*} (i.e., for $l \in \{1, \cdots, 21\}$) leads to remarkably accurate approximations of the eigenpairs of T_μ related to the 21 largest eigenvalues λ_k. The accuracy of the approximate eigenpairs decreases for $l \in \{22, \cdots, 44\}$, and becomes very poor for $k > 44$. The orthogonality, in $L^2(\mu)$, between the normalised approximate eigenfunctions $\widehat{\varphi}_l$ is in perfect agreement with this observation, as illustrated in Figure 10.4.

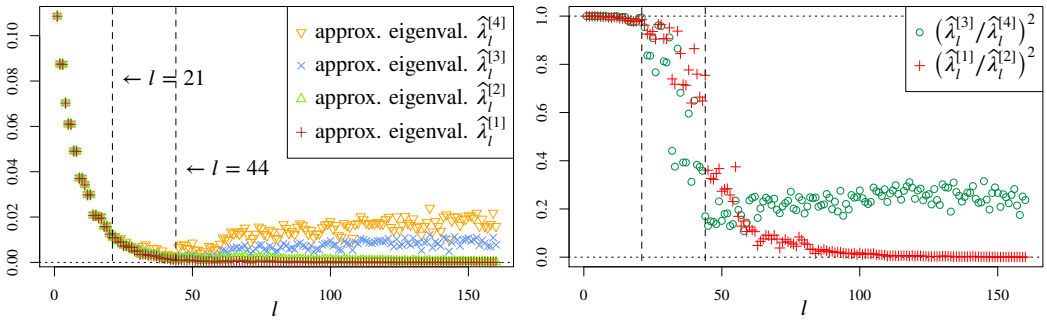

FIG. 10.3. *Approximate eigenvalues $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ induced by the solution v^* presented in Figure 10.2 (left); ratios $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ and $\left(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]}\right)^2$ highlighting the accuracy of the approximate eigendirections ψ_l of T_μ (right).*

A comparison between the true eigenvalues of T_μ and their approximations induced by the solution v^* of Figure 10.2 is presented in Figure 10.3; we for instance observe that for $1 \leqslant l \leqslant 8$, the approximate eigenvalues $\widehat{\lambda}_l^{[4]}$ are the most accurate. Table 10.1 gives the errors $\|\widehat{\varphi}_l - \varphi_l\|^2_{L^2(\mu)}$ for $1 \leqslant l \leqslant 20$; in accordance with the conclusions drawn from Figure 10.3, these approximations are remarkably accurate. Since orthonormalised sets of eigenfunctions are not unique, to perform this comparison, notice that we have when required replaced $\widehat{\varphi}_l$ by $-\widehat{\varphi}_l$, and applied a two-dimensional rotation to pairs of eigendirections related to the approximation of an eigensubspace of dimension two (i.e., when the operator is defined with respect to a uniform measure on $[-1, 1]^2$).

Following Theorem 5.1, we denote by $c_\varkappa$ and c_α the argument of the minimum of the functions $c \mapsto D(cv_\varkappa^*)$ and $c \mapsto D(cv_\alpha^*)$. For the solution presented in Figure (5.2) (i.e., $\varkappa = 0.81$), we obtain

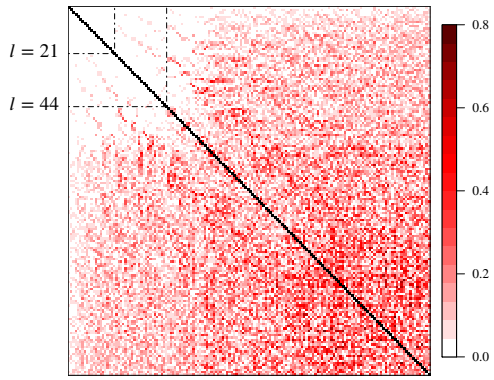

FIG. 10.4. *Graphical representation of the matrix with l, l' entry $\left|\left(\widehat{\varphi}_l \middle| \widehat{\varphi}_{l'}\right)_{L^2(\mu)}\right|$ for the 160 normalised approximate eigendirections induced by the solution $\boldsymbol{\upsilon}^*$ presented in Figure 10.2 (i.e., $\varkappa = 0.81$).*

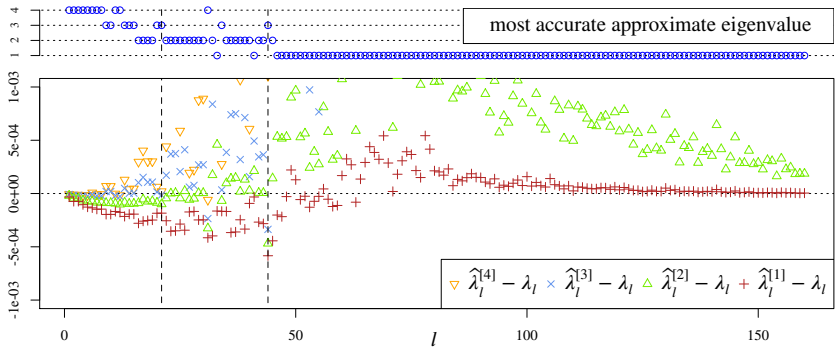

FIG. 10.5. *Errors $\widehat{\lambda}_l^{[\cdot]} - \lambda_l$ for the geometric approximate eigenvalues induced by the solution $\boldsymbol{\upsilon}^*$ presented in Figure 10.2 (bottom), and indication of the most accurate (smallest absolute error) approximation among $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ (top).*

$c_\varkappa = 1.177289$, and $D(c_\varkappa \boldsymbol{\upsilon}_\varkappa^*) = 1.633391 \times 10^{-4}$, and $c_\varkappa \mathbf{d}^T \boldsymbol{\upsilon}_\varkappa^* = 0.9536041$.

10.2. Regularisation path. Following Section 7, we compute the 12 818 first events (i.e., until we reach a precision issue) of the regularisation-path related to problem (5.1) with decreasing values of α; we have in particular $\alpha_0 = 6.310163 \times 10^{-2}$ and $\alpha_{12817} = 1.495359 \times 10^{-5}$. Correspondingly, for problem (5.2) and increasing $\varkappa$, we have $\varkappa_0 = 0$ and $\varkappa_{12817} = 0.9995482$ (and $\mathbf{d}^T \boldsymbol{\omega} = 1$).

Figure 10.6 shows that the number of strictly positive components of the solution $\boldsymbol{\upsilon}_\varkappa^*$ to problem (5.2) tends to increase when $\varkappa$ increases. As expected from Proposition 5.1-(g), the function $\varkappa \mapsto D(\boldsymbol{\upsilon}_\varkappa^*)$ is decreasing; in the same way, when $\varkappa$ increases, the corresponding value of the regularisation parameter α decreases (see Propositions 5.1 and 5.2). We also represent the evolution of conic squared-kernel discrepancy of the various solutions $\boldsymbol{\upsilon}_\varkappa^*$; in accordance with Theorem 5.1, the function $\varkappa \mapsto D(c_\varkappa \boldsymbol{\upsilon}_\varkappa^*)$ is decreasing.

For 51 values of $\varkappa$ evenly spread between $\varkappa_0$ and $\varkappa_{12817}$, Figure 10.7 shows the evolution of the ratio $\left(\widehat{\lambda}_l^{[1]} / \widehat{\lambda}_l^{[2]}\right)^2$ for the approximate eigendecompositions induced by the various solutions $\boldsymbol{\upsilon}_\varkappa^*$. As expected, the number of accurately approximate eigendirections increases with $\varkappa$. Remarkably, 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 we have $\mathrm{trace}(T_{\upsilon_\varkappa^*}) = \varkappa$, since $\mathrm{diag}(\mathbf{K}) = \mathbb{1}$ for the Gaussian kernel.

10.3. Components merging. We now perform the strong-pairwise-merging (see Section 8.2) of the solution $\boldsymbol{\upsilon}^*$ presented in Figure 10.2 (i.e., problem (5.2) with $\varkappa = 0.81$). As illustrated in Figure 10.8, for the first merging iterations, $D(\boldsymbol{\upsilon}_{[k]})$ stays very close to $D(\boldsymbol{\upsilon}^*) = 7.631890 \times 10^{-4}$.

TABLE 10.1

Approximation error $\|\widehat{\varphi}_l - \varphi_l\|^2_{L^2(\mu)}$, with $1 \leqslant l \leqslant 20$, for the normalised approximate eigendirections induced by the solution $\boldsymbol{\upsilon}^*$ presented in Figure 10.2 (i.e., $\varkappa = 0.81$); the values of l grouped together correspond to pairs of eigendirections related to the approximation of an eigensubspace of dimension two.

l	1	2 and 3		4	5 and 6		7 and 8		9 and 10	
$\widehat{\lambda}_l^{[1]}$	0.10861	0.08747	0.08737	0.07028	0.06103	0.06089	0.04907	0.04895	0.03706	0.03692
$\|\widehat{\varphi}_l - \varphi_l\|^2_{L^2(\mu)}$	0.00017	0.00035	0.00035	0.00056	0.00054	0.00120	0.00115	0.00117	0.00245	0.00243
l	11	12 and 13		14 and 15		16 and 17		18 and 19		20
$\widehat{\lambda}_l^{[1]}$	0.03418	0.02976	0.02971	0.02073	0.02070	0.01954	0.01954	0.01573	0.01571	0.01251
$\|\widehat{\varphi}_l - \varphi_l\|^2_{L^2(\mu)}$	0.00196	0.00128	0.00448	0.00438	0.00456	0.00773	0.00685	0.00843	0.00830	0.00711

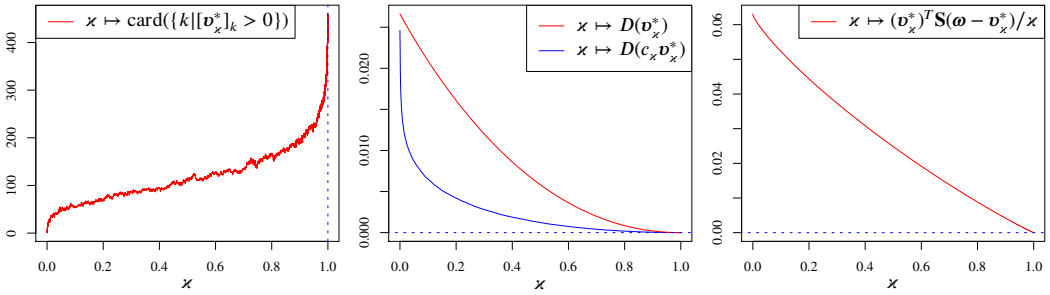

FIG. 10.6. *For the two-dimensional example (Gaussian kernel, $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathbb{1}$), graphical representation of the 12 818 first events of the regularisation path related to problem (5.2) for increasing $\varkappa$; number of strictly positive components of $\boldsymbol{\upsilon}^*_\varkappa$ as function of $\varkappa$ (left); graph of $\varkappa \mapsto D(\boldsymbol{\upsilon}^*_\varkappa)$ and $\varkappa \mapsto D(c_\varkappa \boldsymbol{\upsilon}^*_\varkappa)$ (middle), and relation between $\varkappa$ and the parameter α of problem (5.1) (right).*

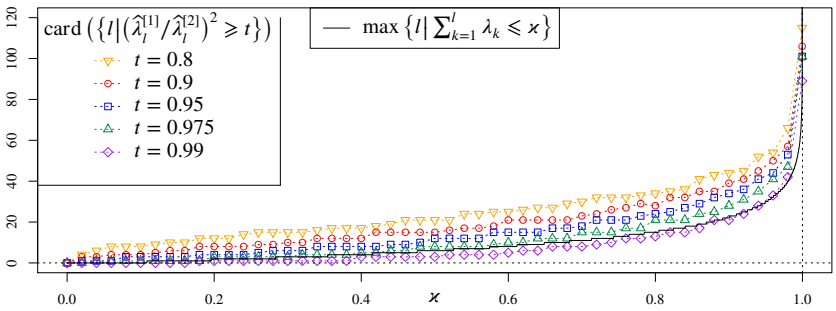

FIG. 10.7. *Evolution of the accuracy of the approximate eigendecomposition of T_μ induced by $\boldsymbol{\upsilon}^*_\varkappa$ for 51 values of $\varkappa$ between $\varkappa_0 = 0$ and $\varkappa_{12817} = 0.9995482$; the accuracy of the approximate eigendirections is measured trough the ratios $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$, and for illustration purpose, the map $\varkappa \mapsto \max\left\{l \middle| \sum_{k=1}^{l} \lambda_k \leqslant \varkappa\right\}$ is also presented (two-dimensional example, Gaussian kernel, $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathbb{1}$).*

After 90 iterations, we have $D(\boldsymbol{\upsilon}_{[90]}) - D(\boldsymbol{\upsilon}^*) = 3.494809 \times 10^{-5}$ (i.e., increase of 4.58%), and $\boldsymbol{\upsilon}_{[90]}$ is supported by 70 points (instead of 160 for $\boldsymbol{\upsilon}^*$); a graphical representation of $\boldsymbol{\upsilon}_{[90]}$ is given in the left-hand part of the figure. The accuracy of the approximate eigendecomposition induced by $\boldsymbol{\upsilon}_{[90]}$ is presented in the right-hand part of Figure 10.8. We observe that although being slightly less accurate than the approximate eigendecomposition induced by $\boldsymbol{\upsilon}^*$, the approximation induced by $\boldsymbol{\upsilon}_{[90]}$ remains very satisfactory while being related to a vector more than two times sparser (notice that the conic squared-kernel discrepancy of the merged solution is $D(c\boldsymbol{\upsilon}_{[90]}) = 2.091099 \times 10^{-4}$).

10.4. Comparison with random sampling. We compute the approximate eigendecompositions induced by random uniform samples, without replacement, of size $n_{rand} = 300$, 600, 900 and 1200 (i.e., we randomly select n_{rand} distinct points among the $N = 2016$ points in S, and we consider the

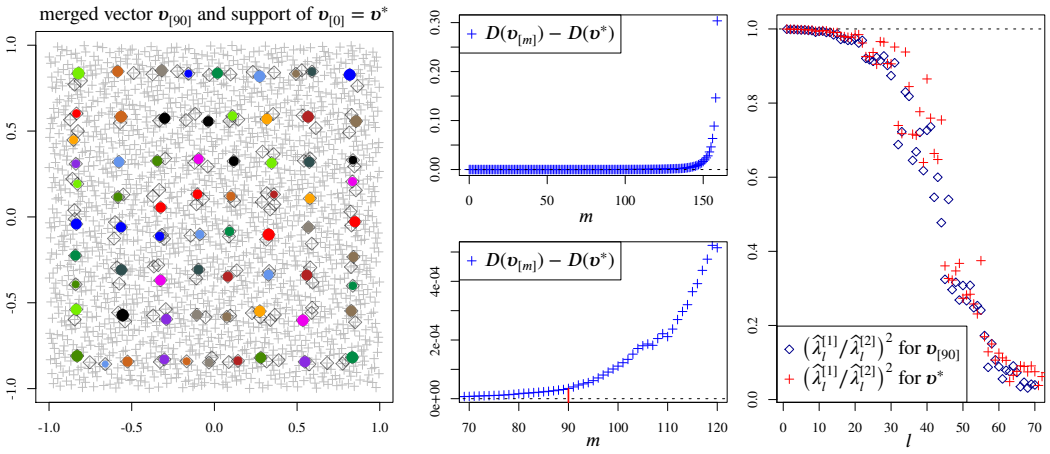

FIG. 10.8. *Graphical representation of the merged solution* $\boldsymbol{\upsilon}_{[90]}$ *(two-dimensional example with* $\boldsymbol{\omega} = \mathbb{1}/N$ *and* $\mathbf{d} = \mathbb{1}$*) obtained after* 90 *iterations of the strong-pairwise-merging strategy applied to the solution* $\boldsymbol{\upsilon}^*$ *presented in Figure 10.2; the grey diamonds indicate the support of* $\boldsymbol{\upsilon}^*$ *(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 ratios* $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ *obtained from the merged vector* $\boldsymbol{\upsilon}_{[90]}$ *and comparison with the same ratios for the solution* $\boldsymbol{\upsilon}^*$ *(right).*

uniform probability measure supported by the points selected); for each sample size, we perform 100 repetitions. Figure 10.9 illustrates the accuracy of the obtained approximate eigendirections, measured through the ratios $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$. As we could expect, the accuracy of the approximation increases with the size of the sample. In terms of trade-off between sparsity and number of eigendirections accurately approximated, the results are however far behind the ones obtained using penalised squared-kernel-discrepancy minimisation (see Figures 10.6 and 10.7). For instance and in comparison to Figure 10.9, penalised squared-kernel discrepancy minimisation leads to the following trade-offs:

- for $\varkappa = 0.81$, the solution $\boldsymbol{\upsilon}_{\varkappa}^*$ is supported by 160 points, and the numbers of approximate eigendirections such that $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2 \geqslant 0.8$, 0.95 and 0.99 are 34, 25 and 15, respectively;
- for $\varkappa = 0.98$, we have 276 support points, and for the same tresholds, the numbers of accurately approximate eigendirections are 66, 53 and 42;
- for $\varkappa = 0.999$, we have 407 support points, and again for the same tresholds, the numbers of accurately approximate eigendirections are 100, 89 and 82.

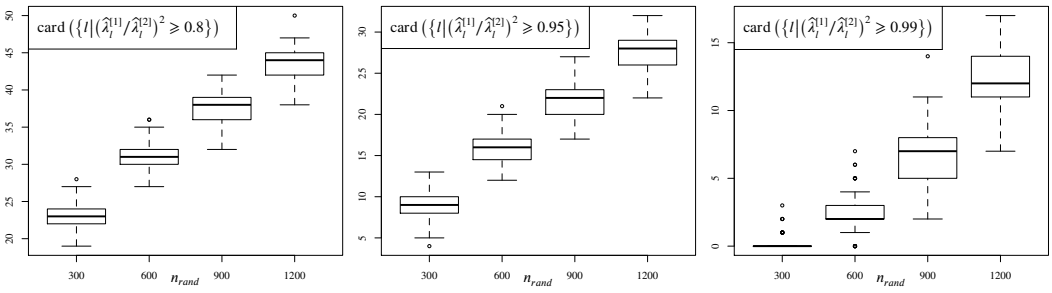

FIG. 10.9. *For the two-dimensional example, accuracy of the approximate eigendecompositions induced by random samples of size* n_{rand} *(without replacement); for each values of* n_{rand}*, Tukey's boxplot, over* 100 *repetitions, of the number of approximate eigendirections such that* $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2 \geqslant 0.8$ *(left),* 0.95 *(middle) and* 0.99 *(right).*

10.5. Impact of the penalisation direction. For the two-dimensional example (Gaussian kernel and $\boldsymbol{\omega} = \mathbb{1}/N$), we compute the regularisation path of problem (5.1) for seven different vectors $\mathbf{d} > 0$. We consider $\mathbf{d} = \mathbf{v}_{\max}(\mathbf{S})$ (i.e., the eigenvector related to the largest eigenvalue of the matrix $\mathbf{S}$, see the Perron–Frobenius theorem), $(\mathbf{S}\boldsymbol{\omega})^2$ (i.e., $d_k = [\mathbf{S}\boldsymbol{\omega}]_k^2$), $\sqrt{\mathbf{S}\boldsymbol{\omega}}$ (i.e., $d_k = \sqrt{[\mathbf{S}\boldsymbol{\omega}]_k}$), $\mathbb{1}$, $1/\sqrt{\mathbf{S}\boldsymbol{\omega}}$,

$1/(\mathbf{S}\boldsymbol{\omega})$ and $1/(\mathbf{S}\boldsymbol{\omega})^2$. In Figure 10.10, we compare the trade-offs between sparsity and (raw and conic) squared-kernel discrepancy yield by these penalisation directions. We recall that for the Gaussian kernel, we have $\mathbb{1} = \operatorname{diag}(\mathbf{K}) = 1/\operatorname{diag}(\mathbf{K})$.

In terms of conic squared-kernel discrepancy and in accordance with Section 9, the results obtained for $\mathbf{d} = \mathbb{1}$ and $\mathbf{d} = 1/(\mathbf{S}\boldsymbol{\omega})^p$ (with in this case $p = 1/2$, 1 and 2) appears as the more interesting; the trade-offs obtained for $\mathbf{d} = \sqrt{\mathbf{S}\boldsymbol{\omega}}$ are also very satisfactory; the performances for these 5 penalisation direction are very close. For this particular example, $\mathbf{d} = 1/(\mathbf{S}\boldsymbol{\omega})^2$ nevertheless appears as the best overall choice among the penalisations considered.

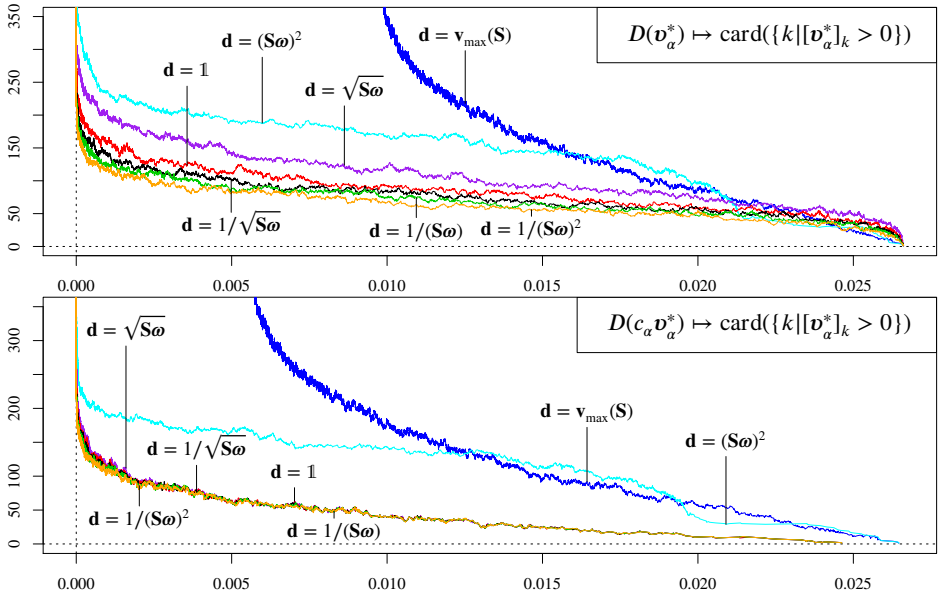

FIG. 10.10. *For the two-dimensional example (Gaussian kernel and $\boldsymbol{\omega} = \mathbb{1}/N$), number of strictly positive components of the solution υ_α^* to problem* (5.1) *as function of the squared-kernel discrepancy $D(\upsilon_\alpha^*)$ (top), and of the conic squared-kernel discrepancy $D(c_\alpha \upsilon_\alpha^*)$ (bottom) for various penalisation vectors* $\mathbf{d}$*; all the curves have been obtained thanks to the regularisation-path strategy.*

10.6. Modified kernel. We further illustrate the impact of the penalisation direction by now considering an alternative kernel (same set S as in the previous experiments, and $\boldsymbol{\omega} = \mathbb{1}/N$). We introduce the function, for $x \in [-1, 1]^2$, $s(x) = \sqrt{0.1 + \|x - c\|^2}$, with $c = (1, 1)$, and we define the kernel (modified Gaussian kernel)

$$K(x, y) = s(x)s(y)\exp(-\ell\|x - y\|^2); \tag{10.1}$$

we still consider $\ell = 1/0.16$. We then in particular have $K(x, x) = s^2(x)$. We make the same analysis as in Section 10.5, while considering $\mathbf{d} = \mathbb{1}$, $\operatorname{diag}(\mathbf{K})$, $1/\operatorname{diag}(\mathbf{K})$, $1/(\mathbf{S}\boldsymbol{\omega})$, $1/(\mathbf{S}\boldsymbol{\omega})^2$ and $(\mathbf{S}\boldsymbol{\omega})^2$. The results are presented in Figure 10.11. The overall tradeoff between sparsity and conic squared-kernel discrepancy obtained for $\mathbf{d} = (\mathbf{S}\boldsymbol{\omega})^2$ is very poor in comparison to the tradeoffs obtained for the five other penalisation directions, in accordance with the remarks of Section 9. The best overall tradeoff is obtained for $\mathbf{d} = \operatorname{diag}(\mathbf{K})$.

11. Application to medium/large-scale problems. This section aims at illustrating the ability of the proposed framework to tackle relatively large-scale problems. The datasets have been obtained from the UCI Machine Learning Repository, see [14]. 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 implemented in C.

11.1. MiniBooNE dataset. We consider the standardised entries of the MiniBooNE dataset (without labels); S thus consists of $N = 129\,596$ points in $\mathbb{R}^{50}$. We use a Gaussian kernel (same

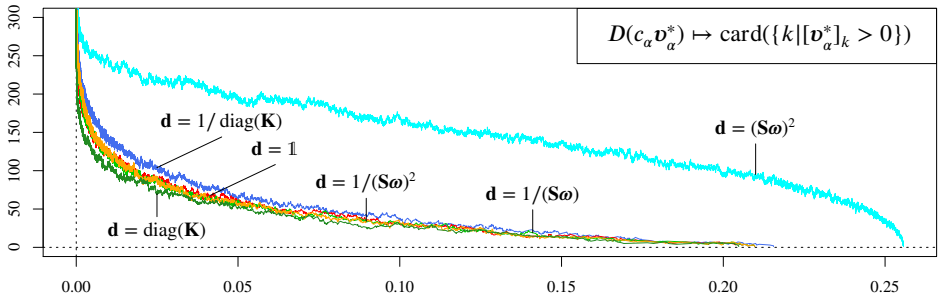

FIG. 10.11. *For the two-dimensional example (modified Gaussian kernel* (10.1) *and* $\boldsymbol{\omega} = \mathbb{1}/N$*), number of strictly positive components of the solution* $\boldsymbol{v}^*_\alpha$ *to problem* (5.1) *as function of the conic squared-kernel discrepancy* $D(c_\alpha \boldsymbol{v}^*_\alpha)$*, for various penalisation vectors* $\mathbf{d}$*; all the curves have been obtained thanks to the regularisation-path strategy.*

expression as in Section 10) with $\ell = 0.02$, and we set $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathbb{1}$ (notice that $\ell = 0.02$ belongs to the range of "good parameter values" for the SVM binary classification of this dataset).

We compute the 3 000 first events of the regularisation path related to problems (5.1) and (5.2). We have $\alpha_0 = 0.2188961$ and $\alpha_{2999} = 3.546703 \times 10^{-3}$, and correspondingly $\varkappa_0 = 0$ and $\varkappa_{2999} = 0.655808$ (notice that $\mathbf{d}^T\boldsymbol{\omega} = 1$); a graphical representation of the properties of these solutions is proposed in Figure 11.1. We can observe that for $\varkappa \geqslant 0.5$, the number of strictly positive components of $\boldsymbol{v}^*_\varkappa$ increases quickly with $\varkappa$; the computation of the regularisation path then becomes intractable (notice that the calculation of the 3 000 first events of the regularisation path took around 3 hours on our aforementioned 2015 desktop).

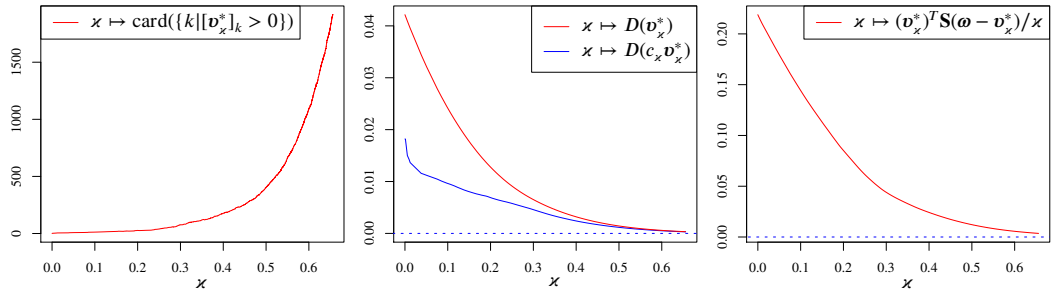

FIG. 11.1. *For the MiniBooNE dataset (Gaussian kernel,* $\boldsymbol{\omega} = \mathbb{1}/N$ *and* $\mathbf{d} = \mathbb{1}$*), graphical representation of the* 3 000 *first events of the regularisation path related to problem* (5.2) *for increasing* $\varkappa$*: number of strictly positive components of* $\boldsymbol{v}^*_\varkappa$ *as function of* $\varkappa$ *(left); graph of* $\varkappa \mapsto D(\boldsymbol{v}^*_\varkappa)$ *and* $\varkappa \mapsto D(c_\varkappa \boldsymbol{v}^*_\varkappa)$ *(middle), and relation between* $\varkappa$ *and the parameter* α *of the regularised problem* (5.1)*.*

From the regulation path, we build the solutions to problem (5.2) for $\varkappa = 0.3$ and $\varkappa = 0.655$ (i.e., for problem (5.1), $\alpha \approx 4.400276 \times 10^{-2}$ and $\alpha \approx 3.571413 \times 10^{-3}$); these solutions have 76 and 1 902 strictly positive components, respectively. The efficiency of the induced approximate eigendecompositions is illustrated in Figure 11.2. For $\varkappa = 0.3$, we obtain a relatively 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 (i.e., $1 \leqslant l \leqslant 8$), and the approximations remains relatively accurate until $l = 29$. Interestingly, we observe that contrary to the ratios $\left(\widehat{\lambda}^{[3]}_l / \widehat{\lambda}^{[4]}_l\right)^2$, the ratios $\left(\widehat{\lambda}^{[1]}_l / \widehat{\lambda}^{[2]}_l\right)^2$ remain relatively high for all the values of l presented in the graph (this behaviour could be a consequence of the 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 8.1. We compute an approximate solution for $\varkappa = 0.8$; the vertex-exchange algorithm is initialised at $\widetilde{\boldsymbol{v}} = \mathbf{e}_1$ and after 300 000 iterations, we obtain a Frank-Wolfe error bound of $\epsilon = 1.692408 \times 10^{-8}$; the obtained approximate solution $\widehat{\boldsymbol{v}}^*$ to problem (5.2) verifies $D(\widehat{\boldsymbol{v}}^*) = 4.934072 \times 10^{-5}$ and has 9544 strictly positive components (in terms of conic squared-kernel

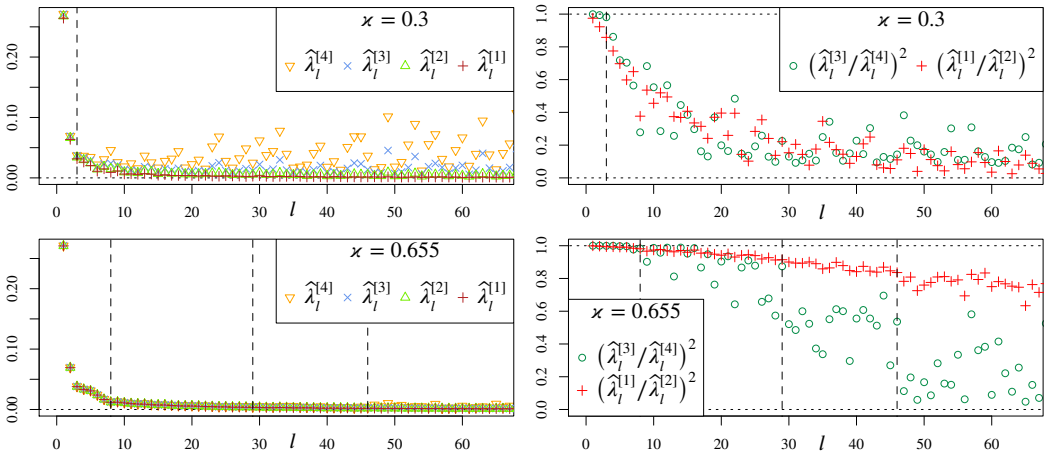

FIG. 11.2. *For the MiniBooNE dataset (Gaussian kernel, $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathbb{1}$), approximate eigenvalues $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ induced by the solution to problem* (5.1) *with $\varkappa = 0.3$ (top-left), and ratios $\big(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\big)^2$ and $\big(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\big)^2$ (top-right); same things for $\varkappa = 0.655$ (bottom-left) and (bottom-right).*

discrepancy, we obtain $D(c\widehat{\boldsymbol{\upsilon}}^*) = 4.672895 \times 10^{-5}$).

To enhance sparsity, we perform a weak-pairwise merging of the solution $\widehat{\boldsymbol{\upsilon}}^*$ for $\varkappa = 0.8$ (see Section 8.2). After 5044 iterations, the merged solution $\boldsymbol{\upsilon}_{[5044]}$ is supported by 4500 points and $D(\boldsymbol{\upsilon}_{[5044]}) = D(\widehat{\boldsymbol{\upsilon}}^*) + 1.061787 \times 10^{-6}$ (i.e., increase of 2.15%).

We next compute the approximate eigendecompositions induced by $\widehat{\boldsymbol{\upsilon}}^*$ and $\boldsymbol{\upsilon}_{[5044]}$; the results are presented in Figure 11.3. In particular, in both case, the 31 main eigendirections of $\boldsymbol{T}_\mu$ are approximated with high accuracy. We also observe that for all the values of l presented on the graph, the approximation induced by $\boldsymbol{\upsilon}_{[5044]}$ is equivalent, in terms of accuracy, to the approximation induced by $\widehat{\boldsymbol{\upsilon}}^*$, while being related to a solution more than two times sparser.

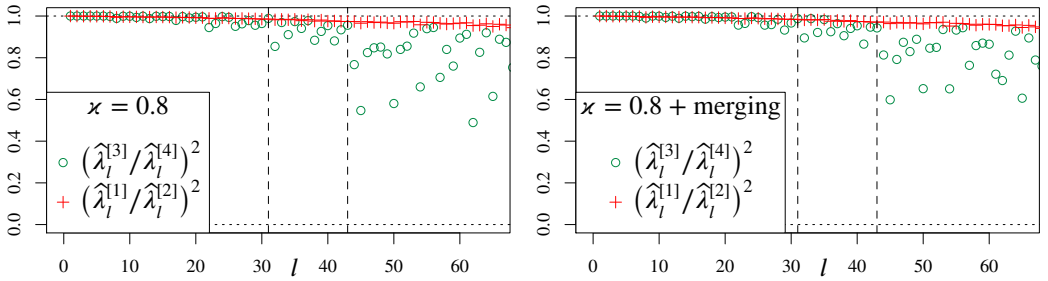

FIG. 11.3. *For the MiniBooNE dataset, accuracy of the approximate eigendecompositions induced by the solution $\widehat{\boldsymbol{\upsilon}}^*$ to problem* (5.1) *with $\varkappa = 0.8$ obtained from the vertex-exchange algorithm (left), and from the merged solution $\boldsymbol{\upsilon}_{[5044]}$ (left).*

11.2. 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 10) with $\ell = 0.4$, and we set $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{d} = \mathbb{1}$. The computation of the distortion term $\mathbf{S}\boldsymbol{\omega}$ took 5 665.6 seconds.

We compute an approximate solution (vertex-exchange strategy) for the constrained problem (5.2) 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 fourth batch), the obtained approximate solution $\widehat{\boldsymbol{\upsilon}}^*$ verifies $D(\widehat{\boldsymbol{\upsilon}}^*) = 3.931629 \times 10^{-5}$ 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 11.1. We observe that a batch of 50 000 iterations of the

vertex-exchange algorithm took around 19 minutes; the approximate solution obtained at the end of the first batch is already relatively accurate.

TABLE 11.1
For the test subsample of the SUSY dataset, information relative to the approximate solutions to problem (5.2) *with* $\varkappa = 0.3$ *returned by the vertex-exchange algorithm for four consecutive batches of* 50 000 *iterations, the solver being initialised at* $\widetilde{\boldsymbol{\upsilon}} = \mathbf{e}_1$*; for each batch, execution time, total number of iterations, Frank-Wolfe error bound* ϵ *and number* n *of strictly positive components 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

To enhance sparsity, we perform a weak-pairwise merging of the approximate solution $\widehat{\boldsymbol{\upsilon}}^*$; the computation of 20 673 merging iterations took 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., increase of only 1.34%). We then study the approximate eigendecomposition induced by $\boldsymbol{\upsilon}_{[13674]}$. Computing the 300 first normalised approximate eigenvectors $\widehat{\mathbf{v}}_l$ of $\mathbf{KW}$ induced by $\boldsymbol{\upsilon}_{[13674]}$ (i.e., $\widehat{\mathbf{v}}_l \in \mathbb{R}^N$ is the vector corresponding to $\widehat{\varphi}_l$, see Section 4.3) took 3 278.2 seconds (time for canonical extension and rescaling), and we thus also obtain the approximate eigenvalues $\widehat{\lambda}_l^{[1]}$. For l and $l' \in \{1, \cdots, 300\}$, we have $\max_{l \neq l'} |(\widehat{\varphi}_l | \widehat{\varphi}_{l'})_{L^2(\mu)}| \approx 0.003734$, so that we can expect the approximations $\widehat{\varphi}_l$ to be relatively accurate. To access precisely their accuracy, we compute $T_\mu[\widehat{\varphi}_l]$ (i.e., $\mathbf{KW}\widehat{\mathbf{v}}_l$) for these 300 first approximate eigendirections; this operation took 191 622.3 seconds (i.e., around 53 hours). The results are presented in Figure 11.4. As already observed, the accuracy of the approximate eigendirections decreases when l increases (we recall that the eigenvalues of the approximate operator are stored in descending order); all the obtained approximate eigenpairs are remarkably accurate (and we recall that we consider only 7 000 points among 500 000).

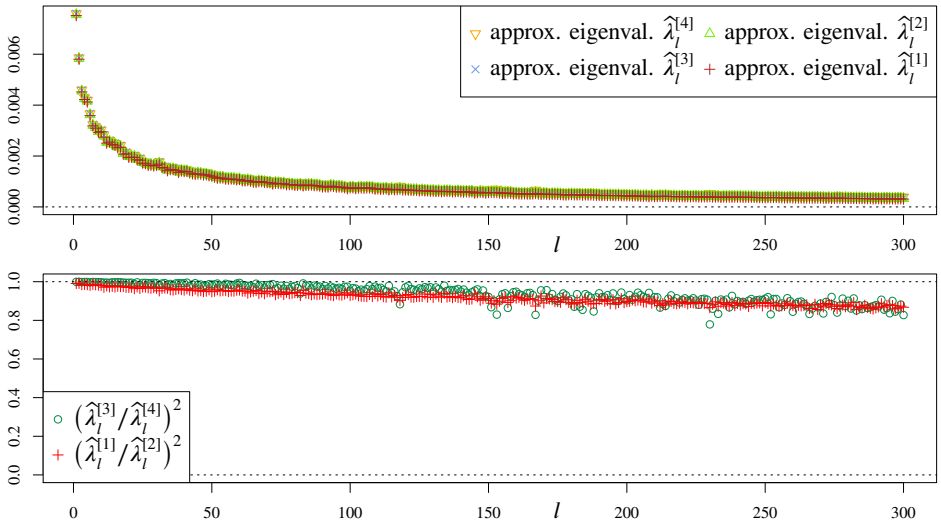

FIG. 11.4. *For the test subsample of the SUSY dataset, graphical representation of the* 300 *first approximate eigenvalues* $\widehat{\lambda}_l^{[\cdot]}$ *induced by the merged solution* $\boldsymbol{\upsilon}_{[13674]}$ *obtained from the approximate solution* $\widehat{\boldsymbol{\upsilon}}^*$ *to problem* (5.2) *with* $\varkappa = 0.3$ *(top); ratios* $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ *and* $\left(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]}\right)^2$ *measuring the accuracy of the underlying approximate eigendirections (bottom).*

12. Conclusion. We have studied a QP-based strategy to design sparse quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels in a quadrature-

sparsification framework, i.e., when only quadratures with support included in a fixed finite set of points are considered. The points selected through penalised squared-kernel-discrepancy minimisation can in particular be interpreted as the support vectors of one-class distorted SVMs defined from the squared kernel, the initial measure and the ℓ^1-type penalisation term.

A special attention has been drawn to the approximation of the main eigenpairs of an initial operator induced by the eigendecomposition of an approximate operator. To assess the accuracy these approximations, the notions of geometric approximate eigenvalue and of conic squared-kernel discrepancy have been introduced, and their properties have been investigated. We have in particular demonstrate that, for a given penalisation direction, increasing the impact of the penalisation generally tends to increase the sparsity of the approximate measure at the expense of reducing the overall accuracy of the induced spectral approximation.

Numerical strategies to solve large-scale penalised squared-kernel-discrepancy minimisation problems have been discussed. The regularisation-path approach can be used to explore the range of very sparse solutions, with the interest of leading to a set of exact solutions (up to precision errors); the vertex-exchange strategy permits the exploration of a wider range of solutions and offers a numerically efficient approach to build approximate solutions. Two greedy heuristics based on iterative pairwise-component merging have also been described, aiming at enhancing sparsity while keeping squared-kernel discrepancy as low as possible.

The main bottleneck of the penalised squared-kernel-discrepancy-minimisation framework is the preliminary computation of the dual distortion term $\mathbf{g}_\mu$ (i.e., in the discrete case, of $\mathbf{S}\boldsymbol{\omega}$); this operation can nevertheless be easily, and potentially massively, parallelised. Once $\mathbf{g}_\mu$ is known, sparse solutions can be obtained readily. Assessing the accuracy of an approximate eigendirection trough the computation of the four associated geometric approximate eigenvalues can also prove challenging (same complexity as the distortion term); this operation is nevertheless optional, and the more affordable orthogonality test might be performed to detect poorly approximated eigendirections.

We have observed that the penalisation direction can have a significant impact on the trade-off between sparsity and (conic) squared-kernel discrepancy, and specific problem-based penalisation directions have been discussed; the characterisation of efficient penalisation terms is however a widely open problem. Investigating in more detail the relations between sparsity, (conic) squared-kernel discrepancy, and accuracy of the induced spectral approximations also appears as an interesting perspective. In the matrix-approximation framework, the study of the properties of the low-rank approximations obtained by penalised squared-kernel-discrepancy minimisation should also deserve further attention.

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: CoE PFV/10/002 (OPTEC); Flemish Government: FWO projects G.0377.12 and G.088114N; Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO).

Appendix A. Kernel discrepancy and integration in RKHS. Consider the framework of Section 2 and introduce the subset $\mathscr{I}(K)$ of $\mathcal{M}$, defined as

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

notice that what follows may be extended to signed measures on $\mathscr{X}$.

From the reproducing property of $K(\cdot,\cdot)$ and the Cauchy-Schwarz inequality, we have

$$\forall h \in \mathcal{H} \text{ and } \forall \mu \in \mathscr{I}(K), \left|\int_{\mathscr{X}} h(x)\mathrm{d}\mu(x)\right| \leqslant \int_{\mathscr{X}} |h(x)|\mathrm{d}\mu(x) \leqslant \|h\|_{\mathcal{H}} \int_{\mathscr{X}} \sqrt{K(x,x)}\mathrm{d}\mu(x).$$

The linear functional I_μ on $\mathcal{H}$, defined as $I_\mu[h] = \int_{\mathscr{X}} h(x)\mathrm{d}\mu(x)$, is therefore continuous. Thus, from the Riesz representation theorem, there exists $h_\mu \in \mathcal{H}$ such that $I_\mu[h] = (h|h_\mu)_{\mathcal{H}}$, and for $x \in \mathscr{X}$, $h_\mu(x) = \int_{\mathscr{X}} K(x,t)\mathrm{d}\mu(t)$.

For μ and $\nu \in \mathscr{I}(K)$, we have $(h_\mu|h_\nu)_{\mathcal{H}} = \int_{\mathscr{X}\times\mathscr{X}} K(x,t)\mathrm{d}\mu(x)\mathrm{d}\nu(t)$. The *kernel discrepancy* between two measures μ and $\nu \in \mathscr{I}(K)$ is defined as

$$D_K(\mu,\nu) = \|h_\mu - h_\nu\|^2_{\mathcal{H}} = \|h_\mu\|^2_{\mathcal{H}} + \|h_\nu\|^2_{\mathcal{H}} - 2(h_\mu|h_\nu)_{\mathcal{H}},$$

and $E_K(\mu) = \|h_\mu\|^2_{\mathcal{H}}$ is sometimes referred to as the *energy* of the measure μ with respect to $K(\cdot,\cdot)$.

For μ and $\nu \in \mathscr{I}(K)$, from the Cauchy-Schwarz inequality, we have, for all $h \in \mathcal{H}$,

$$\left|\int_{\mathscr{X}} h(x)\mathrm{d}\mu(x) - \int_{\mathscr{X}} h(x)\mathrm{d}\nu(x)\right| = \left|(h|h_\mu - h_\nu)_{\mathcal{H}}\right| \leqslant \|h\|_{\mathcal{H}}\sqrt{D_K(\mu,\nu)}.$$

So, when the integrands belong to the RKHS $\mathcal{H}$, the error induced by approximating integrals with respect to μ by integrals with respect to ν has a tight bound in terms of kernel discrepancy; to approximate integrals with respect to μ, it is therefore of interest to deal with a measure ν such that $D_K(\nu,\mu)$ is small; see [5] for a further discussion.

Appendix B. Soft-margin-type extensions for the one-class SVM related to the regularised problem. By analogy with SVM models, we can define soft-margin-type extensions of problems (6.1) and (6.3), i.e, we can consider models where the inequalities appearing in the constraints can potentially be violated, the level of violation being penalised. We only discuss extensions related to problem (6.1), but a similar discussion holds for problem (6.3).

Let $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_N)^T \in \mathbb{R}^N$; 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 [23]. In what follows, we discuss the model obtained for two popular types of loss functions: the (weighted) hinge loss and the (generalised) square loss. Soft-margin extensions of problem (6.1) appear as way to further constrain or penalise the approximate measure ν (i.e., the vector $\boldsymbol{\upsilon}$).

B.1. Weighted hinge loss. Let $\mathbf{c} \in \mathbb{R}^N$, with $\mathbf{c} \geqslant 0$; the soft-margin extension of problem (6.1) corresponding to a weighted hinge loss consists in the problem, for $g \in \mathcal{G}$ and $\boldsymbol{\xi} \in \mathbb{R}^N$,

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

The Lagrange dual of problem (B.1) (written as a minimisation problem) is given by

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

Problem (B.2) only differs from problem (5.1) 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 $\boldsymbol{\upsilon}$.

B.2. Generalised square loss. Let $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ be a symmetric positive-definite matrix; the soft-margin extension of (6.1) corresponding to a generalised square loss is

$$\begin{array}{ll} \underset{g}{\text{minimise}} & \frac{1}{2}\|g\|^2_{\mathcal{G}} + (g|g_\mu)_{\mathcal{G}} + \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi} \\ \text{subject to} & g(x_k) \geqslant -\alpha d_k - \xi_k, \text{ for all } k \in \{1,\cdots,N\}. \end{array} \tag{B.3}$$

The Lagrange dual of problem (B.3) is given by

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D_{\alpha,\boldsymbol{\Sigma}}(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}) + \tfrac{1}{2}\boldsymbol{\upsilon}^T\boldsymbol{\Sigma}\boldsymbol{\upsilon} + \alpha\mathbf{d}^T\boldsymbol{\upsilon} \text{ subject to } \boldsymbol{\upsilon} \geqslant 0. \tag{B.4}$$

In comparison to problem (5.1), in (B.4), the term $\frac{1}{2}\boldsymbol{\upsilon}^T\boldsymbol{\Sigma}\boldsymbol{\upsilon}$ is added to the initial cost $D_\alpha(\boldsymbol{\upsilon})$; it tends to "harmonise" the components of the solution. In particular, $D_{\alpha,\boldsymbol{\Sigma}}$ is always strongly convex.

Appendix C. Proofs. This section groups together the proofs of the results stated in this work.

Proof of Proposition 2.1. Consider an o.n.b. $\{h_j\}_{j\in\mathbb{I}}$ of $\mathcal{H}$. From (2.1), for all $j \in \mathbb{I}$, we have

$$\begin{aligned}\big(T_\mu[h_j]\big|T_\nu[h_j]\big)_{\mathcal{H}} &= \big(h_j\big|T_\nu[h_j]\big)_{L^2(\mu)} = \big(T_\mu[h_j]\big|h_j\big)_{L^2(\nu)}\\ &= \int_{\mathscr{X}\times\mathscr{X}} K(x,t)h_j(x)h_j(t)\mathrm{d}\mu(x)\mathrm{d}\nu(t),\end{aligned} \tag{C.1}$$

so that $\big(T_\mu\big|T_\nu\big)_{\mathrm{HS}(\mathcal{H})} = \sum_{j\in\mathbb{I}}\int_{\mathscr{X}\times\mathscr{X}} K(x,t)h_j(x)h_j(t)\mathrm{d}\mu(x)\mathrm{d}\nu(t)$. For x and $t \in \mathscr{X}$, we have $K(x,t) = \sum_{j\in\mathbb{I}} h_j(x)h_j(t)$, and thus

$$\|K\|^2_{L^2(\mu\otimes\nu)} = \int_{\mathscr{X}\times\mathscr{X}} \sum_{j\in\mathbb{I}} K(x,t)h_j(x)h_j(t)\mathrm{d}\mu(x)\mathrm{d}\nu(t). \tag{C.2}$$

Equalities (C.1) and (C.2) hold for any o.n.b. of $\mathcal{H}$, so that we can in particular consider an o.n.b. which contains the o.n.b. $\{\sqrt{\lambda_k}\varphi_k\}_{k\in\mathbb{I}_\mu^+}$ of $\mathcal{H}_\mu$ defined by T_μ. From the linearity and continuity of T_μ, we then obtain

$$\big(T_\mu\big|T_\nu\big)_{\mathrm{HS}(\mathcal{H})} = \sum_{k\in\mathbb{I}_\mu^+}\int_{\mathscr{X}} \lambda_k^2\varphi_k^2(t)\mathrm{d}\nu(t) \text{ and } \|K\|^2_{L^2(\mu\otimes\nu)} = \int_{\mathscr{X}}\sum_{k\in\mathbb{I}_\mu^+}\lambda_k^2\varphi_k^2(t)\mathrm{d}\nu(t),$$

and we conclude by using the Tonelli theorem. □

Proof of Lemma 2.1. From the properties of $K(\cdot,\cdot)$, the squared kernel $K^2(\cdot,\cdot)$ is symmetric and positive-semidefinite (see in particular the Schur product theorem); in addition the squared kernel is non-negative, i.e., $K^2(x,t) \geqslant 0$ for all x and $t \in \mathscr{X}$. Considering the framework of Appendix A, we can remark that $\mathscr{T}(K) = \mathscr{I}(K^2)$, so that the result directly follows from the definition of g_μ and g_ν, and from Proposition 2.1. □

Proof of Lemma 2.2. The proof directly follows from the properties discussed in Sections 2.1 and 2.2. In particular, (2.3) 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 $h \in \mathcal{H}_{0\mu}$. The inequality involving τ_μ is consequence of the relation $\|h\|^2_{L^2(\mu)} \leqslant \tau_\mu\|h\|^2_{\mathcal{H}}$, for all $h \in \mathcal{H}$. □

Proof of Theorem 3.1. For all $k \in \mathbb{I}_\mu^+$, we have $\|\sqrt{\lambda_k}\varphi_k\|^2_{\mathcal{H}} = 1$. By analogy, for $l \in \mathbb{I}_\nu^+$ with $\|\psi_l\|_{L^2(\mu)} > 0$, we define $\widehat{\lambda}_l^{[1]}$ so that $\|\sqrt{\widehat{\lambda}_l^{[1]}}\widehat{\varphi}_l\|^2_{\mathcal{H}} = 1$. Assuming that ψ_l is a true eigendirection of T_μ and denoting by $\lambda > 0$ the associated eigenvalue; from (2.1), we have

$$1 = \widehat{\lambda}_l^{[1]}\|\widehat{\varphi}_l\|^2_{\mathcal{H}} = \widehat{\lambda}_l^{[1]}\big(\widehat{\varphi}_l\big|\widehat{\varphi}_l\big)_{\mathcal{H}} = \frac{\widehat{\lambda}_l^{[1]}}{\lambda}\big(T_\mu[\widehat{\varphi}_l]\big|\widehat{\varphi}_l\big)_{\mathcal{H}} = \frac{\widehat{\lambda}_l^{[1]}}{\lambda}\|\widehat{\varphi}_l\|^2_{L^2(\mu)} = \frac{\widehat{\lambda}_l^{[1]}}{\lambda},$$

so that $\widehat{\lambda}_l^{[1]} = \lambda$. From the Cauchy-Schwarz inequality, we have

$$\widehat{\lambda}_l^{[1]} = \big(\sqrt{\vartheta_l}\psi_l\big|T_\mu[\sqrt{\vartheta_l}\psi_l]\big)_{\mathcal{H}} \leqslant \|\sqrt{\vartheta_l}\psi_l\|_{\mathcal{H}}\big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}} = \big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}} = \widehat{\lambda}_l^{[2]},$$

with equality when ψ_l and $T_\mu[\psi_l]$ are collinear, i.e., when ψ_l is a true eigendirection of T_μ.

From $\widehat{\lambda}_l^{[1]} = \big(\sqrt{\vartheta_l}\psi_l\big|T_\mu[\sqrt{\vartheta_l}\psi_l]\big)_{\mathcal{H}}$, we obtain the definition of $\widehat{\lambda}_l^{[3]}$ by considering the Hilbert structure of $L^2(\mu)$ instead of the one of $\mathcal{H}$. The inequality $\widehat{\lambda}_l^{[2]} \leqslant \widehat{\lambda}_l^{[3]}$, with equality when ψ_l is an eigendirection of T_μ, directly follows from the relation $\widehat{\lambda}_l^{[3]} = \big(\widehat{\lambda}_l^{[2]}\big)^2/\widehat{\lambda}_l^{[1]}$. Finally, from the Cauchy-Schwarz inequality, we have

$$\widehat{\lambda}_l^{[3]} = \big(\widehat{\varphi}_l\big|T_\mu[\widehat{\varphi}_l]\big)_{L^2(\mu)} \leqslant \|\widehat{\varphi}_l\|_{L^2(\mu)}\big\|T_\mu[\widehat{\varphi}_l]\big\|_{L^2(\mu)} = \big\|T_\mu[\widehat{\varphi}_l]\big\|_{L^2(\mu)} = \widehat{\lambda}_l^{[4]},$$

with, again, equality when ψ_l is an eigendirection of T_μ.

The expansions (3.1) and (3.2) follow from the definition of the four geometric approximate eigenvalues related to an approximate eigendirection of T_μ induced by T_ν. □

Proof of Theorem 3.2. The expressions of c_ν and $\phi(c_\nu)$ follow from Proposition 2.1 and from the minimisation of the univariate convex quadratic function

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

The characterisation of $c_\nu T_\nu$ as an orthogonal projection and the fact that all such operators lie on a sphere in $\mathrm{HS}(\mathcal{H})$ is a direct consequence of the definition of c_ν. We then remark that $\left\{\sqrt{\widehat{\lambda}^{[1]}_l}\widehat{\varphi}_l\right\}_{l\in\mathbb{I}^+_\nu}$ is by definition an o.n.b. of $\mathcal{H}_\nu = \mathcal{H}_{c_\nu\nu}$. Following Lemma 2.2 and Remark 2.2, we have

$$D_{K^2}(\mu, c_\nu\nu) = \sum_{l\in\mathbb{I}^+_\nu} \left\|T_\mu[\sqrt{\widehat{\lambda}^{[1]}_l}\widehat{\varphi}_l] - c_\nu\vartheta_l\sqrt{\widehat{\lambda}^{[1]}_l}\widehat{\varphi}_l\right\|^2_{\mathcal{H}} + \left(K_{0\nu}\middle|K_\mu\right)_{L^2(\mu\otimes\mu)},$$

with, in particular, $\left(K_{0\nu}\middle|K_\mu\right)_{L^2(\mu\otimes\mu)} \geqslant 0$. We conclude by using the optimality properties of the approximate eigenvalues $\widehat{\lambda}^{[1]}_l$ and $\widehat{\lambda}^{[3]}_l$, see Theorem 3.1. □

Proof of Proposition 5.1. Assertion (a) follows from $D_{K^2}(\mu,\mu) = 0$ and $D_{K^2}(\mu,\nu) \geqslant 0$. From the first order optimality condition, for $\alpha \geqslant 0$, a feasible $\boldsymbol{\upsilon}^*_\alpha$ is solution to (5.1) if and only if, for any feasible $\boldsymbol{\upsilon}$, we have $(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) \geqslant 0$. Considering $\boldsymbol{\upsilon}^*_\alpha = 0$ gives $\alpha\mathbf{d} \geqslant \mathbf{S}\boldsymbol{\omega}$, leading to (b), in addition, since all the entries of $\mathbf{S}$ are positive, there can not exist a vector $\boldsymbol{\varepsilon} \geqslant 0$ such that $\mathbf{S}\boldsymbol{\varepsilon} = 0$ and $\boldsymbol{\varepsilon} \neq 0$, so that the solution is in this case unique; also, since $\boldsymbol{\omega}$ is feasible for (5.1), we obtain (c) by taking $\boldsymbol{\upsilon} = \boldsymbol{\omega}$. For assertion (d), we first remark that the first order optimality condition for $\boldsymbol{\upsilon} = 0$ gives $(\boldsymbol{\upsilon}^*_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) \leqslant 0$. Next, if we assume that there exists k such that $[\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha)]_k < 0$, then for all $\beta > (\boldsymbol{\upsilon}^*_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha)/[\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha)]_k \geqslant 0$, we obtain $(\beta\mathbf{e}_k - \boldsymbol{\upsilon}^*_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) < 0$, and the first order optimality condition would be violated for the feasible vector $\boldsymbol{\upsilon} = \beta\mathbf{e}_k$ (we recall that $\mathbf{e}_k$ stands for the k-th element of the canonical basis of $\mathbb{R}^N$, so that $\mathbf{e}_k^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) = [\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha)]_k$). We thus necessarily have $\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) \geqslant 0$ and $(\boldsymbol{\upsilon}^*_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) = 0$ (since $\boldsymbol{\upsilon}^*_\alpha \geqslant 0$). To prove (e), we first remark that

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

Since $D_\alpha(\widetilde{\boldsymbol{\upsilon}}^*_\alpha) = D_\alpha(\boldsymbol{\upsilon}^*_\alpha)$ and $(\widetilde{\boldsymbol{\upsilon}}^*_\alpha - \boldsymbol{\upsilon}^*_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) \geqslant 0$, we necessarily have $(\widetilde{\boldsymbol{\upsilon}}^*_\alpha - \boldsymbol{\upsilon}^*_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\alpha) = 0$ and $(\widetilde{\boldsymbol{\upsilon}}^*_\alpha - \boldsymbol{\upsilon}^*_\alpha)^T\mathbf{S}(\widetilde{\boldsymbol{\upsilon}}^*_\alpha - \boldsymbol{\upsilon}^*_\alpha) = 0$ (since the matrix $\mathbf{S}$ is symmetric and positive-semidefinite), and the result follows. Assertion (f) is a direct corollary of (d), since $\mathbf{S}\boldsymbol{\upsilon}^*_\alpha \geqslant 0$. To obtain (g) and (h), we consider $\alpha_1 < \alpha_2$, and we denote by $\boldsymbol{\upsilon}^*_{\alpha_1}$ and $\boldsymbol{\upsilon}^*_{\alpha_2}$ some corresponding solutions to (5.1). We have $D(\boldsymbol{\upsilon}^*_{\alpha_1}) - D(\boldsymbol{\upsilon}^*_{\alpha_2}) \leqslant \alpha_1\mathbf{d}^T(\boldsymbol{\upsilon}^*_{\alpha_2} - \boldsymbol{\upsilon}^*_{\alpha_1})$ and $D(\boldsymbol{\upsilon}^*_{\alpha_1}) - D(\boldsymbol{\upsilon}^*_{\alpha_2}) \geqslant \alpha_2\mathbf{d}^T(\boldsymbol{\upsilon}^*_{\alpha_2} - \boldsymbol{\upsilon}^*_{\alpha_1})$, so that, necessarily, $\mathbf{d}^T(\boldsymbol{\upsilon}^*_{\alpha_2} - \boldsymbol{\upsilon}^*_{\alpha_1}) \leqslant 0$, and therefore $D(\boldsymbol{\upsilon}^*_{\alpha_2}) - D(\boldsymbol{\upsilon}^*_{\alpha_1}) \geqslant 0$. Assuming that $\alpha_2 = \alpha_1 + \epsilon$, with $\epsilon > 0$, we can remark that $D_{\alpha_2}(\boldsymbol{\upsilon}^*_{\alpha_2}) = D_{\alpha_1}(\boldsymbol{\upsilon}^*_{\alpha_2}) + \epsilon\mathbf{d}^T\boldsymbol{\upsilon}^*_{\alpha_2} \geqslant D_{\alpha_1}(\boldsymbol{\upsilon}^*_{\alpha_1})$. In addition, from (d), we can deduce that $D_\alpha(\boldsymbol{\upsilon}^*_\alpha) = \frac{1}{2}\left(\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\omega} - (\boldsymbol{\upsilon}^*_\alpha)^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha\right)$, so that the map $\alpha \mapsto (\boldsymbol{\upsilon}^*_\alpha)^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha$ is decreasing (since $\alpha \mapsto D_\alpha(\boldsymbol{\upsilon}^*_\alpha)$ is increasing); finally, since $\alpha \mapsto 2D(\boldsymbol{\upsilon}^*_\alpha) = \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\omega} + (\boldsymbol{\upsilon}^*_\alpha)^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha - 2\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha$ is increasing and $\alpha \mapsto (\boldsymbol{\upsilon}^*_\alpha)^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha$ is decreasing, the function $\alpha \mapsto \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha$ is necessarily decreasing. □

Proof of Proposition 5.2. If $\boldsymbol{\upsilon}^*_\alpha$ is a solution to (5.1) with $\alpha \geqslant 0$, then by definition, $\boldsymbol{\upsilon}^*_\alpha$ minimises $D(\cdot)$ over the set $\{\boldsymbol{\upsilon} \geqslant 0 | \mathbf{d}^T\boldsymbol{\upsilon} = \mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha\}$, so that $\boldsymbol{\upsilon}^*_\alpha$ is a solution to (5.2) with $\varkappa = \mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha$.

The condition $\varkappa \leqslant \mathbf{d}^T\boldsymbol{\omega}$ follows directly from Proposition 5.1-(c): a solution $\boldsymbol{\upsilon}^*_\alpha$ to (5.1) indeed necessarily satisfies $\mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha \leqslant \mathbf{d}^T\boldsymbol{\omega}$. For $\varkappa = 0$, we have $\boldsymbol{\upsilon}^*_\varkappa = 0$, which, from Proposition 5.1-(b) is solution to (5.1) for $\alpha \geqslant \max_k\{[\mathbf{S}\boldsymbol{\omega}]_k/d_k\}$. For $0 < \varkappa \leqslant \mathbf{d}^T\boldsymbol{\omega}$, from Proposition 5.1-(d), if $\boldsymbol{\upsilon}^*_\varkappa$ is a solution to (5.1), then we necessarily have $(\boldsymbol{\upsilon}^*_\varkappa)^T\nabla D_\alpha(\boldsymbol{\upsilon}^*_\varkappa) = 0$, leading to the expected value for α. The last assertions follow directly from Proposition 5.1-(g) and (h), and the relations between the solutions to the problems (5.1) and (5.2). □

Proof of Theorem 5.1. For $0 \leqslant \alpha < \max_k\{[\mathbf{S}\boldsymbol{\omega}]_k/d_k\} = \alpha_0$, we have $\boldsymbol{\upsilon}^*_\alpha \neq 0$, see Proposition 5.1-(b); in addition, if α is such that $\mathbf{S}(\boldsymbol{\upsilon}^*_\alpha - \boldsymbol{\omega}) = 0$, then $c_\alpha = 1$. We now assume that $\mathbf{S}(\boldsymbol{\upsilon}^*_\alpha - \boldsymbol{\omega}) \neq 0$; from

Proposition 5.1-(d), we have $(\boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{v}_\alpha^* - \boldsymbol{\omega}) + \alpha\mathbf{d}^T\boldsymbol{v}_\alpha^* = 0$, leading to $c_\alpha = 1 + \frac{\alpha\mathbf{d}^T\boldsymbol{v}_\alpha^*}{(\boldsymbol{v}_\alpha^*)^T\mathbf{S}\boldsymbol{v}_\alpha^*} \geqslant 1$. By definition of c_α, we also have $c_\alpha(\boldsymbol{v}_\alpha^*)^T\mathbf{S}\boldsymbol{v}_\alpha^* = \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_\alpha^*$, so that

$$(\boldsymbol{\omega} - c_\alpha\boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{\omega} - c_\alpha\boldsymbol{v}_\alpha^*) = \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\omega} - c_\alpha\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_\alpha^* = (\boldsymbol{\omega} - c_\alpha\boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*) \geqslant 0,$$

and thus $c_\alpha(\boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*) \leqslant \boldsymbol{\omega}^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*)$, i.e.,

$$c_\alpha \leqslant \frac{\boldsymbol{\omega}^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*)}{(\boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*)} = 1 + \frac{(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*)}{(\boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*)}.$$

Using Proposition 5.1-(d), we obtain $\alpha c_\alpha\mathbf{d}^T\boldsymbol{v}_\alpha^* \leqslant \alpha\mathbf{d}^T\boldsymbol{v}_\alpha^* + (\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*)^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}_\alpha^*) \leqslant \alpha\mathbf{d}^T\boldsymbol{\omega}$, the last inequality being consequence of Proposition 5.1-(c).

Consider $0 \leqslant \alpha_1 < \alpha_2 < \alpha_0$; from Proposition 5.1-(d) and by definition of c_α, we have

$$\begin{aligned}(c_{\alpha_1}\boldsymbol{v}_{\alpha_2}^* - \boldsymbol{v}_{\alpha_1}^*)^T\left[\mathbf{S}(\boldsymbol{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \alpha_1\mathbf{d}\right] &= (c_{\alpha_1}\boldsymbol{v}_{\alpha_2}^*)^T\mathbf{S}(\boldsymbol{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \alpha_1 c_{\alpha_1}\mathbf{d}^T\boldsymbol{v}_{\alpha_2}^* \\ &= (\boldsymbol{v}_{\alpha_2}^*)^T\mathbf{S}(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \alpha_1 c_{\alpha_1}\mathbf{d}^T\boldsymbol{v}_{\alpha_2}^* - (c_{\alpha_1} - 1)\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_{\alpha_2}^* \\ &= (\boldsymbol{v}_{\alpha_2}^*)^T\mathbf{S}(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \tfrac{\alpha_1}{(\boldsymbol{v}_{\alpha_1}^*)^T\mathbf{S}\boldsymbol{v}_{\alpha_1}^*}\left[(\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_{\alpha_1}^*)\mathbf{d}^T\boldsymbol{v}_{\alpha_2}^* - (\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_{\alpha_2}^*)\mathbf{d}^T\boldsymbol{v}_{\alpha_1}^*\right] \geqslant 0. \end{aligned} \tag{C.3}$$

Since $(\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_{\alpha_2}^*)\mathbf{d}^T\boldsymbol{v}_{\alpha_1} \geqslant (\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_{\alpha_1}^*)\mathbf{d}^T\boldsymbol{v}_{\alpha_2}$ (we indeed assume that $\alpha \mapsto \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{v}_\alpha^*/\mathbf{d}^T\boldsymbol{v}_\alpha^*$ is increasing), inequality (C.3) entails that $(\boldsymbol{v}_{\alpha_2}^*)^T\mathbf{S}(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^* - \boldsymbol{\omega}) \geqslant 0$, so that $(c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^* - c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*)^T\mathbf{S}(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^* - \boldsymbol{\omega}) \geqslant 0$, and thus, by convexity, $D(c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^*) \geqslant D(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*)$; we recall that $\nabla D(\boldsymbol{v}) = \mathbf{S}(\boldsymbol{v} - \boldsymbol{\omega})$. From Proposition 5.1-(d), we also have, for all $\alpha \geqslant 0$, $c \geqslant 0$, and $\boldsymbol{v} \geqslant 0$,

$$(\boldsymbol{v} - c\boldsymbol{v}_\alpha^*)^T\left[\mathbf{S}(\boldsymbol{v}_\alpha^* - \boldsymbol{\omega}) + \alpha\mathbf{d}\right] \geqslant 0. \tag{C.4}$$

We introduce $\tau = \boldsymbol{\omega}^T\mathbf{S}(c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^* - c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*) = 2\big(D(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*) - D(c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^*)\big)$, and $\beta = \mathbf{d}^T(c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^* - c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*)$. From (C.4) and by definition of c_α, we deduce that

$$\begin{aligned}(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^* - c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^*)^T\mathbf{S}(\boldsymbol{v}_{\alpha_2}^* - \boldsymbol{\omega}) &= \tau + (\boldsymbol{v}_{\alpha_2}^*)^T\nabla D(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*) \\ &= \tau + \tfrac{1}{c_{\alpha_2}}(c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^* - c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*)^T\nabla D(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*) \geqslant \alpha_2\beta. \end{aligned} \tag{C.5}$$

From the Taylor expansion of $D(c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^*)$ at $c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*$, we can also deduce that

$$-\tfrac{1}{2}\tau \geqslant (c_{\alpha_2}\boldsymbol{v}_{\alpha_2}^* - c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*)^T\nabla D(c_{\alpha_1}\boldsymbol{v}_{\alpha_1}^*). \tag{C.6}$$

Since $0 < 1/c_{\alpha_2} \leqslant 1$ and $\tau \leqslant 0$, inequalities (C.5) and (C.6) imply $\beta \leqslant 0$, as expected. □

Proof of Proposition 6.1. Define the closed linear subspace $\mathcal{G}_S = \operatorname{span}\{K_{x_k}^2\}_{k=1}^N$ of $\mathcal{G}$, and let $\mathcal{G}_0 = \mathcal{G}_S^\perp$ be its orthogonal; by definition, $g_\mu \in \mathcal{G}_S$. For any $g_S \in \mathcal{G}_S$ and $g_0 \in \mathcal{G}_0$, we have

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

In addition, for any $k \in \{1, \cdots, N\}$, we have $g_0(x_k) = 0$, so that, necessarily, $g_\alpha^* \in \mathcal{G}_S$ (representer Theorem), i.e, there exists $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_N^*)^T \in \mathbb{R}^N$ such that $g_\alpha^* = \sum_{k=1}^N \beta_k^* K_{x_k}^2$. Restricting problem (6.1) to $\mathcal{G}_S$ then yields, for $\boldsymbol{\beta} \in \mathbb{R}^N$,

$$\underset{\boldsymbol{\beta}}{\text{minimise}}\ \tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\omega} \text{ subject to } \mathbf{S}\boldsymbol{\beta} \geqslant -\alpha\mathbf{d}. \tag{C.7}$$

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 condition gives $\mathbf{S}\boldsymbol{\beta} = \mathbf{S}(\boldsymbol{\upsilon} - \boldsymbol{\omega})$, leading to the Lagrange dual (5.1) (written as a minimisation problem). If $\boldsymbol{\upsilon}^*_\alpha$ is a solution to (5.2), then a solution $\boldsymbol{\beta}^*$ to (C.7) 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 $g^*_\alpha \in \mathcal{G}_S$ does not depend on such a $\boldsymbol{\varepsilon}$. 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 Proposition 6.2. We follow the same reasoning than in the proof of Proposition 6.1. By restricting problem (6.3) to $\mathcal{G}_S$, we obtain, for $\boldsymbol{\beta} \in \mathbb{R}^N$,

$$\underset{\boldsymbol{\beta},\gamma}{\text{minimise}}\ \tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\omega} - \gamma \text{ subject to } \mathbf{S}\boldsymbol{\beta} \geqslant \gamma\mathbf{d}/\varkappa. \tag{C.8}$$

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

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

The primal optimality condition gives $\mathbf{S}\boldsymbol{\beta} = \mathbf{S}(\boldsymbol{\upsilon} - \boldsymbol{\omega})$ and $\mathbf{d}^T\boldsymbol{\upsilon} = \varkappa$, leading to the Lagrange dual (5.2). If $\boldsymbol{\upsilon}^*_\varkappa$ is a solution to (5.2), then a solution $\boldsymbol{\beta}^*$ to (C.8) 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 $(\boldsymbol{\upsilon}^*_\varkappa)^T\left[\mathbf{S}(\boldsymbol{\upsilon}^*_\varkappa - \boldsymbol{\omega}) - \gamma^*_\varkappa\mathbf{d}/\varkappa\right] = 0$, as well as the equality $g^*_\varkappa(x_k) = \gamma^*_\varkappa d_k/\varkappa$ for all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{\upsilon}^*_\varkappa]_k > 0$. □

Proof of Proposition 7.1. 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 $\boldsymbol{\upsilon}^T_\alpha\left(\mathbf{S}(\boldsymbol{\upsilon}_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{d}\right) = 0$, 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 (5.1), and $J_\alpha = J$. □

Proof of Proposition 7.2. We first recall that, from Proposition 5.1-(e), for a given $\alpha \geqslant 0$, the terms $\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha$ and $\mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha$ are always unique. From (7.1), for any solution $\boldsymbol{\upsilon}^*_\alpha$ to (5.1), there exists a $n_\alpha \times n_\alpha$ matrix $\mathbf{G}$ such that

$$[\boldsymbol{\upsilon}^*_\alpha]_{J_\alpha} = \mathbf{G}\big([\mathbf{S}\boldsymbol{\omega}]_{J_\alpha} - \alpha\mathbf{d}_{J_\alpha}\big); \tag{C.9}$$

the matrix $\mathbf{G}$ is a generalised inverse of $\mathbf{S}_{J_\alpha,J_\alpha}$ (i.e., $\mathbf{S}_{J_\alpha,J_\alpha}\mathbf{G}\mathbf{S}_{J_\alpha,J_\alpha} = \mathbf{S}_{J_\alpha,J_\alpha}$), see for instance [2]. Combined with Proposition 7.1, the condition (7.1) thus implies that the maps $\alpha \mapsto \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha$ and $\alpha \mapsto \mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha$ are piecewise linear; in addition, since the indices of the strictly positive components of $\boldsymbol{\upsilon}^*_\alpha$ belongs to J_α, changes in the sparsity pattern always only involve null components of $\boldsymbol{\upsilon}^*_\alpha$, so that these two maps are also continuous. We introduce $\psi(\alpha) = \boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha/\mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha$; on the interval $[0, \alpha_0)$, the function $\psi(\cdot)$ is continuous (since we in this case have $\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha > 0$ and $\mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha > 0$). From (C.9), and since $[\boldsymbol{\upsilon}^*_\alpha]_{J^c_\alpha} = 0$, we have

$$\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\upsilon}^*_\alpha = [\mathbf{S}\boldsymbol{\omega}]^T_{J_\alpha}\mathbf{G}[\mathbf{S}\boldsymbol{\omega}]_{J_\alpha} - \alpha[\mathbf{S}\boldsymbol{\omega}]^T_{J_\alpha}\mathbf{G}\mathbf{d}_{J_\alpha}, \text{ and } \mathbf{d}^T\boldsymbol{\upsilon}^*_\alpha = [\mathbf{S}\boldsymbol{\omega}]^T_{J_\alpha}\mathbf{G}\mathbf{d}_{J_\alpha} - \alpha\mathbf{d}^T_{J_\alpha}\mathbf{G}\mathbf{d}_{J_\alpha}.$$

Thus, if $\alpha \geqslant 0$ is not a kink for problem (5.1), we obtain that $\psi'(\alpha) \geqslant 0$ if and only if

$$\big([\mathbf{S}\boldsymbol{\omega}]^T_{J_\alpha}\mathbf{G}\mathbf{d}_{J_\alpha}\big)^2 \leqslant \big([\mathbf{S}\boldsymbol{\omega}]^T_{J_\alpha}\mathbf{G}[\mathbf{S}\boldsymbol{\omega}]_{J_\alpha}\big)\big(\mathbf{d}^T_{J_\alpha}\mathbf{G}\mathbf{d}_{J_\alpha}\big). \tag{C.10}$$

If $\mathbf{G}$ is symmetric and positive-semidefinite, then inequality (C.10) corresponds to the Cauchy-Schwarz inequality, and is therefore verified. Since the number of kinks is finite, we can then conclude that $\psi(\cdot)$ is increasing on $[0, \alpha_0)$. □

Proof of Lemma 7.1. Let $\boldsymbol{v}_\alpha$ be such that $[\boldsymbol{v}_\alpha]_{J^c} = 0$ and $[\boldsymbol{v}_\alpha]_J = (\mathbf{S}_{J,J})^{-1}([\mathbf{S}\boldsymbol{\omega}]_J - \alpha \mathbf{d}_J)$. Following (7.2), from the condition $[\mathbf{S}(\boldsymbol{v}_\alpha - \boldsymbol{\omega}) + \alpha \mathbf{d}]_{J^c} \geqslant 0$, we define α_+ as the smallest α satisfying the constraint $\alpha \left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l \leqslant \left[\mathbf{M}[\mathbf{S}\boldsymbol{\omega}]_J - [\mathbf{S}\boldsymbol{\omega}]_{J^c}\right]_l$, for all $l \in \{1, \cdots, \mathrm{card}(J^c)\}$. By definition (and in view of Remark 7.1), this constraint is satisfied by α_p; the components l such that $\left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l \geqslant 0$ therefore carry no information. The problem thus consists in searching for the smallest α such that

$$\alpha \geqslant \left[\mathbf{M}[\mathbf{S}\boldsymbol{\omega}]_J - [\mathbf{S}\boldsymbol{\omega}]_{J^c}\right]_l / \left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l, \text{ for all } l \text{ such that } \left[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}\right]_l < 0.$$

In the same way, we define α_- as the smallest α such that $\alpha(\mathbf{S}_{J,J})^{-1}\mathbf{d}_J \leqslant (\mathbf{S}_{J,J})^{-1}[\mathbf{S}\boldsymbol{\omega}]_J$. □

REFERENCES

[1] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

[2] Adi Ben-Israel and Thomas N.E. Greville. *Generalized Inverses: Theory and Applications*. Springer, 2003.

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

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

[5] 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.

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

[7] 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.

[8] Bertrand Gauthier and Luc Pronzato. Convex relaxation for IMSE optimal design in random-field models. *Computational Statistics and Data Analysis*, 113:375–394, 2017.

[9] 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.

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

[11] 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.

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

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

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

[15] Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, 1992.

[16] 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.

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

[18] 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.

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

[20] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[21] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

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

[23] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

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

[25] Kai Zhang and James T. Kwok. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576–1587, 2010.