Optimal quadrature-sparsification for integral operator approximation

Bertrand Gauthier, Johan a K Suykens

OPTIMAL QUADRATURE-SPARSIFICATION FOR INTEGRAL OPERATOR APPROXIMATION

BERTRAND GAUTHIER[*§†] AND JOHAN A.K. SUYKENS[‡†]

Abstract. We address the problem of designing sparse quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels. We more specifically focus on computing sparse quadratures with support included in fixed finite sets of points (quadrature-sparsification), this framework being of particular interest since it encompasses the Nyström approximation of kernel-matrices. For a given kernel, the accuracy of a quadrature approximation is assessed through the squared Hilbert-Schmidt norm (for operators 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, we refer to the underlying criterion as the squared-kernel discrepancy between two measures. Sparsity of the approximate quadrature is achieved through the introduction of an ℓ^1-type penalisation, and the computation of a penalised squared-kernel-discrepancy-optimal approximation thus consists in a convex quadratic minimisation problem. The penalisation can be introduced under the form of a regularisation term or of a constraint, both formulations being equivalent. The quadratic programs related to the regularised and constrained problems can in particular be interpreted as the Lagrange dual formulations of distorted one-class support-vector machines related to the squared kernel and the initial measure. Numerical strategies for solving large-scale squared-kernel discrepancy minimisation problems are investigated and the efficiency of the approach is illustrated on a series of examples. We in particular demonstrate the ability of the proposed methodology to lead to accurate sparse representations of the main eigenpairs of kernel-matrices related to large-scale datasets.

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

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

1. Introduction.

Computing the eigendecomposition of an integral operator defined from a symmetric and positive-semidefinite kernel and a discrete measure supported by N points is numerically equivalent to the diagonalisation of a $N \times N$ symmetric and positive-semidefinite matrix. Such a situation is for instance encountered each time a pointwise quadrature is used to approximate an integral operator,or more directly when one aims at computing the eigendecomposition of a kernel-matrix (or of any Gram matrix). In the non-sparse case and for the direct approach, the amount of computations required to perform the diagonalisation of such a matrix scales as $\mathcal{O}(N^3)$, and thus becomes quickly intractable when N is large (not to mention issues related to the storage of large-scale matrices).

In the framework of matrix approximation, a common alternative consists in carrying out the eigendecomposition of a (weighted) principal submatrix of the initial matrix, with size $n \ll N$, and then expanding the result back up to dimension N; by analogy with the integral operators framework (see, e.g., [9]), this approach is usually referred to as the Nyström method. Naturally, the choice of the considered submatrix has a strong impact on the quality of the induced approximation, raising questions relative to the selection of an appropriate submatrix, see for instance [4, 21, 12, 23, 2, 7] and references therein. Translated in the integral-operator framework, this operation consists in defining a sparse approximation of an initial discrete measure, the support of the approximate measure being included in the support of the initial one. More generally, we refer to the problem consisting in computing a sparse quadrature while enforcing the support of this quadrature to be included in a fixed finite set of points as *quadrature-sparsification*.

Following for instance [21], for a given kernel $K(\cdot,\cdot)$ and an initial measure μ (defining an integral operator T_μ, see Section 2 for a detailed discussion), we assess the accuracy of an approximate measure ν (defining an operator T_ν) by considering the squared Hilbert-Schmidt norm of the difference between the integral operators T_μ and T_ν viewed as operators on the underlying reproducing kernel Hilbert space (RKHS, see for instance [1]). By analogy with the notion of kernel discrepancy, see [3] and Appendix A, we refer to the underlying criterion as the *squared-kernel discrepancy* between 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

measures μ and ν, denoted by $D_{K^2}(\mu,\nu)$.

For a given n, the search of a measure ν_n^* supported by n points and such that $D_{K^2}(\mu,\nu_n^*)$ is minimal, is in general a difficult non-convex minimisation problem. As an alternative, we assume that the support of the quadrature ν is included in a fixed finite set of N points (quadrature-sparsification; this framework in particular encompasses the Nyström approximation of kernel-matrices). The function $\nu \mapsto D_{K^2}(\mu,\nu)$ then consists in a convex quadratic function on $\mathbb{R}^N$, and sparsity of the approximate quadrature can be induced by the introduction of an ℓ^1-type penalisation of the squared-kernel discrepancy (the sample size n is thus not fixed a priori). The penalisation can be introduced under the form of a regularisation term or of a constraint (both formulations being equivalent), and the computation of a penalised squared-kernel-discrepancy-optimal quadrature with support included in S then consists in a convex quadratic minimisation problem (see Section 3). The quadratic programs (QPs) related to the regularised and constrained squared-kernel-discrepancy minimisation problems can in particular be interpreted as the Lagrange dual formulations of distorted one-class support-vector machines related to the squared kernel and the initial measure, as discussed in Section 4.

We present a careful analysis of the approach and present two different numerical strategies for solving penalised squared-kernel-discrepancy minimisation problems. The first strategy, described in Section 5, is a sequential direct method based on the notion of regularisation path (see for instance [16]); this method appears as an efficient way to explore the range of very sparse solutions to the regularised or constrained problems. The second strategy, see Section 6, applies more specifically to the constrained formulation and consists in using a vertex-exchange algorithm (see, e.g., [17, Chap. 9]) to build an approximate solution to the underlying QP. Importantly, the proposed strategies do not require the storage of any large objects, so that large-scale problems can be considered (more particularly since we are interested in the range of sparse solutions). In addition, to enhance the sparsity of a given discrete measure ν while trying to keep the squared-kernel discrepancy as low as possible, we describe two greedy heuristics based on iterative pairwise-component merging (Section 6.2).

We also further investigate the problem related to the computation of an approximate eigen-decomposition of the initial operator T_μ from the eigendecomposition of an approximate operator T_ν. Based on geometrical considerations, we introduce four different approximations (referred to as *approximate eigenvalues*) of the eigenvalue related to a given approximate eigendirection of T_μ (i.e, the eigendirections of the approximate operator T_ν), and the accuracy of a given approximate eigendirection can then in particular be assessed through the comparison of these different approximate eigenvalues, as detailed in Section 7. A discussion relative to the choice of relevant penalisation weights (since a weighted-ℓ^1-type penalisation is considered) is proposed Section 8. Finally, Sections 9 (two-dimension toy example) and 10 (large-scale problems) are devoted to numerical experiments, and Section 11 concludes.

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

2. Notations, recalls and theoretical motivations. This section is devoted to recalls related to integral operators defined from a symmetric positive-semidefinite kernel. 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 reproducing RKHS of real-valued functions on $\mathscr{X}$ (see for instance [1]). We assume that $\mathcal{H}$ is a separable Hilbert space.

2.1. Integral operators. We assume $\mathscr{X}$ is a measurable space and we denote by $\mathcal{A}$ the underlying σ-algebra. We suppose that the (real-valued) kernel $K(\cdot,\cdot)$ is measurable on $\mathscr{X} \times \mathscr{X}$ for the product σ-algebra $\mathcal{A} \otimes \mathcal{A}$ (see for instance [22, Chap. 4]), so that, in particular, the RKHS $\mathcal{H}$ consists of measurable functions on $\mathscr{X}$. We also assume that the diagonal of $K(\cdot,\cdot)$, i.e., the function $x \mapsto K(x,x)$ is measurable on $(\mathscr{X},\mathcal{A})$. We denote by $\mathcal{M}$ the set of all measures on $(\mathscr{X},\mathcal{A})$ and we introduce

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

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 $\mathcal{H}$)

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

In addition, for all $h \in \mathcal{H}$, we have $h \in L^2(\mu)$ and $\|h\|^2_{L^2(\mu)} \leqslant \tau_\mu \|h\|^2_{\mathcal{H}}$, i.e., $\mathcal{H}$ is continuously included in $L^2(\mu)$. We 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 $x \in \mathscr{X}$,

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

We have $T_\mu[f] \in \mathcal{H}$ for all $f \in L^2(\mu)$ and, for $h \in \mathcal{H}$, $(h|T_\mu[f])_{\mathcal{H}} = (h|f)_{L^2(\mu)}$, see for instance [5, 6] 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 $\widetilde{\varphi}_k \in L^2(\mu)$ be the associated eigenfunctions, orthonormalised for $L^2(\mu)$. For $k \in \mathbb{I}_\mu^+$, let $\varphi_k = \frac{1}{\lambda_k} T_\mu[\widetilde{\varphi}_k] \in \mathcal{H}$ be the canonical extension of $\widetilde{\varphi}_k$, so that $\{\sqrt{\lambda_k}\varphi_k | k \in \mathbb{I}_\mu^+\}$ is an orthonormal basis (o.n.b.) of the subspace $\mathcal{H}_\mu$ of $\mathcal{H}$. The reproducing kernel of $\mathcal{H}_\mu$ (for the Hilbert structure of $\mathcal{H}$) is $K_\mu(x,t) = \sum_{k\in\mathbb{I}_\mu^+} \lambda_k \varphi_k(x)\varphi_k(t)$, and $K_{0\mu}(\cdot,\cdot) = K(\cdot,\cdot) - K_\mu(\cdot,\cdot)$ is the reproducing kernel of $\mathcal{H}_{0\mu}$; We also recall that $\tau_\mu = \sum_{k\in\mathbb{I}_\mu^+} \lambda_k$ is the trace of the integral operator T_μ on $L^2(\mu)$.

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

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

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

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

2.3. Hilbert-Schmidt norm and squared-kernel discrepancy. In view of Section 2.1, for $\mu \in \mathscr{T}(K)$, the operator T_μ can also be interpreted as an operator on $\mathcal{H}$ (see, e.g., [21]); with a slight abuse of notation, we keep the same notation for "T_μ viewed as an operator on $L^2(\mu)$", and "T_μ viewed as an operator on $\mathcal{H}$". In both cases, T_μ is an Hilbert-Schmidt operator.

Let μ and $\nu \in \mathscr{T}(K)$; for an o.n.b. $\{h_j | j \in \mathbb{I}\}$ 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

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

and we recall that the value of $\left(T_\mu | T_\nu\right)_{\mathrm{HS}(\mathcal{H})}$ does not depend on the choice of the o.n.b. of $\mathcal{H}$, see, e.g., [20]. The underlying Hilbert-Schmidt norm (for operators on $\mathcal{H}$) is given by

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

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

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

Lemma 2.1. *For μ and $\nu \in \mathscr{T}(K)$, we have $(T_\mu | T_\nu)_{\mathrm{HS}(\mathcal{H})} = \|K\|^2_{L^2(\mu\otimes\nu)}$, 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 $\|K\|^2_{L^2(\mu\otimes\nu)} = \int_{\mathscr{X}\times\mathscr{X}} \left(K(x,t)\right)^2 \mathrm{d}\mu(x)\mathrm{d}\nu(t)$.

In particular, note 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 $\|K\|^2_{L^2(\mu\otimes\nu)} = 0$.

The terminology "squared-kernel discrepancy" is motivated by the analogy with the notion of "kernel discrepancy" discussed for instance in [3] (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 $\mathcal{H}$. The squared-kernel discrepancy is also related to approximate integration of functions in the RKHS associated with the *squared kernel*; in what follows we denote by $\mathcal{G}$ the RKHS associated with $K^2(\cdot,\cdot) = \left(K(\cdot,\cdot)\right)^2$. Following Appendix A, for all $\mu \in \mathscr{T}(K)$ (notice that $\mathscr{T}(K) = \mathscr{I}(K^2)$), we can introduce $g_\mu(x) = \int_{\mathscr{X}} K^2(x,t)\mathrm{d}\mu(t)$, with $x \in \mathscr{X}$. We have $g_\mu \in \mathcal{G}$ and finally

$$D_{K^2}(\mu,\nu) = \|g_\mu - g_\nu\|^2_{\mathcal{G}}. \tag{2.1}$$

The following Lemma 2.2 further highlights the connection between the squared-kernel discrepancy and the error induced by the approximation of T_μ (the initial integral operator) by T_ν (the approximate operator).

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 $\left\{\sqrt{\lambda_k}\varphi_k \middle| k \in \mathbb{I}^+_\mu\right\}$ be the o.n.b. basis of $\mathcal{H}_\mu$ defined by T_μ, we have*

$$D_{K^2}(\mu,\nu) = \sum_{k\in\mathbb{I}^+_\mu} \lambda_k \left\|T_\mu[\varphi_k] - T_\nu[\varphi_k]\right\|^2_{\mathcal{H}}, \tag{2.2}$$

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

In view of (2.2), for a fixed measure μ, by minimising the function $\nu \mapsto D_{K^2}(\mu,\nu)$ under the condition $\mathcal{H}_\nu \subset \mathcal{H}_\mu$, we minimise, for the RKHS norm, $\lambda_k\varphi_k - T_\nu[\varphi_k]$ for all $k \in \mathbb{I}^+_\mu$, while putting more emphasis on the approximation of the eigenpairs with largest eigenvalues (the eigenvalues λ_k playing the rule of penalisation weights). 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 (i.e., the ones related to the largest eigenvalues).

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. 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, in the following Section 3, we restrict the problem 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), and instead of fixing a priori the number n of support points, we induce sparsity through the introduction of an ℓ^1-type penalisation. We refer to the underlying problem as *quadrature-sparsification*.

3. Optimal quadrature-sparsification as quadratic programming. We consider a fixed set of points $\mathcal{S}$, and we restrict our study to discrete measures ν with support included in $\mathcal{S}$. We more particularly focus on the case where $\mathcal{S}$ is the support of an initial discrete measure μ.

3.1. Preliminary discussion. Let $\nu = \sum_{k=1}^N \upsilon_k \delta_{x_k}$, with $\boldsymbol{\upsilon} = (\upsilon_1,\cdots,\upsilon_N)^T \geqslant 0$ (that is $\upsilon_k \geqslant 0$ for all k), and consider a general fixed measure $\mu \in \mathscr{T}(K)$; 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 kernel-matrix defined by the squared kernel $K^2(\cdot,\cdot)$ and the set of points $\mathcal{S}$, i.e., with i,j entry $\mathbf{S}_{i,j} = K^2(x_i,x_j) \geqslant 0$ ($\mathbf{S}$ is therefore a non-negative, positive-semidefinite, symmetric matrix), and where $\mathbf{g}_\mu = (g_\mu(x_1),\cdots,g_\mu(x_N))^T \in \mathbb{R}^N$, with $g_\mu(x_k) = \int_{\mathscr{X}} K^2(x_k,t)\mathrm{d}\mu(t)$ (so that $\mathbf{g}_\mu \geqslant 0$). For such a discrete measure ν, we thus obtain

$$D_{K^2}(\mu,\nu) = \|K\|^2_{L^2(\mu\otimes\mu)} + \boldsymbol{\upsilon}^T\mathbf{S}\boldsymbol{\upsilon} - 2\mathbf{g}_\mu^T\boldsymbol{\upsilon}, \tag{3.1}$$

and $\nu \mapsto D_{K^2}(\mu,\nu)$ can this way be interpreted as a quadratic function of $\boldsymbol{\upsilon}$ (i.e., the vector of weights characterising ν). By minimising $\boldsymbol{\upsilon} \mapsto \boldsymbol{\upsilon}^T\mathbf{S}\boldsymbol{\upsilon} - 2\mathbf{g}_\mu^T\boldsymbol{\upsilon}$ under the constraint $\boldsymbol{\upsilon} \geqslant 0$, we can obtain the best approximation of μ (in terms of squared-kernel discrepancy) among all discrete measures supported by $\mathcal{S}$ (note that this in practice requires the knowledge of the vector $\mathbf{g}_\mu \in \mathbb{R}^N$, which may be problematic for a general measure μ). We however not consider such a minimisation problem; instead, we introduce an ℓ^1-type penalisation of the squared-kernel discrepancy (to obtain approximate measures ν supported by a number of points significantly smaller than N).

In order to discard problems relative to the computation of the vector $\mathbf{g}_\mu$ and for the sake of simplicity, we assume that the measure μ is also discrete and supported by $\mathcal{S}$. More precisely, we consider that $\mu = \sum_{k=1}^N \omega_k\delta_{x_k}$, with $\boldsymbol{\omega} > 0$ (i.e., $\omega_k > 0$ for all k). We then in particular have $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ for all $\boldsymbol{\upsilon} \geqslant 0$ (see Lemma 2.2), and $\mathbf{g}_\mu = \mathbf{S}\boldsymbol{\omega}$; in addition, note that this framework is of particular interest since it corresponds to problems related to kernel-matrix low-rank approximation, see Section 2.2, Remark 3.1, and for instance [4].

Thus, for $\mu = \sum_{k=1}^N \omega_k\delta_{x_k}$ and $\nu = \sum_{k=1}^N \upsilon_k\delta_{x_k}$ be , with $\boldsymbol{\omega} > 0$ and $\boldsymbol{\upsilon} \geqslant 0$, we have

$$D_{K^2}(\mu,\nu) = (\boldsymbol{\omega}-\boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega}-\boldsymbol{\upsilon}). \tag{3.2}$$

For a fixed discrete measure μ supported by $\mathcal{S}$ (i.e., $\boldsymbol{\omega} > 0$ is fixed), we define

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

the factor ½ being added for simplification purpose.

Remark 3.1 (relation with the classical Frobenius-norm-based criterion)**.** In the framework of equation (3.2), the squared-kernel discrepancy appears as a natural extension of the classical Frobenius-norm-based criterion for matrix low-rank approximation (see for instance [4]). Indeed, assume that $\boldsymbol{\omega} = \mathbb{1}$, with $\mathbb{1} = (1,\cdots,1)^T \in \mathbb{R}^N$, so that $\mathbf{W} = \mathrm{diag}(\boldsymbol{\omega}) = \mathrm{Id}_N$ is the $N\times N$ identity matrix, and thus $\mathbf{KW} = \mathbf{W}^{1/2}\mathbf{KW}^{1/2} = \mathbf{K}$ (notice the analogy with the notations used in Section 2.2).

Selecting columns of the matrix $\mathbf{K}$ can be interpreted as performing the product $\mathbf{KV}$, where $\mathbf{V} = \mathrm{diag}(\boldsymbol{\upsilon})$ is a $N\times N$ diagonal matrix with diagonal entries 0 or 1 (we are therefore assuming that

all the components of $\boldsymbol{\upsilon}$ are 0 or 1); in the same way, considering $\mathbf{VK}$ amounts to selecting rows of $\mathbf{K}$. Since such a sampling matrix $\mathbf{V}$ satisfies $\mathbf{V}^2 = \mathbf{V} = \mathbf{V}^T$, selecting a principal submatrix of the matrix $\mathbf{K}$ can be viewed as considering the matrix $\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2} = \mathbf{V}\mathbf{K}\mathbf{V}$.

In this case, i.e., if $\boldsymbol{\omega} = \mathbb{1}$ and if the components of $\boldsymbol{\upsilon}$ are all 0 or 1, we can easily verify that $(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) = \|\mathbf{K} - \mathbf{V}\mathbf{K}\mathbf{V}\|_F^2$, where $\|\cdot\|_F$ stands for the Frobenius norm. ◁

3.2. Regularised squared-kernel-discrepancy minimisation. We consider the framework of equation (3.2) (i.e., the measures μ and ν are discrete, with support included in S). For a given penalisation vector $\mathbf{f} = (f_1, \cdots, f_N)^T \in \mathbb{R}^N$, with $\mathbf{f} > 0$ (see Section 8), and $\alpha \geqslant 0$, we introduce the *regularised squared-kernel-discrepancy minimisation* problem

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D_\alpha(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) + \alpha\mathbf{f}^T\boldsymbol{\upsilon} \text{ subject to } \boldsymbol{\upsilon} \geqslant 0. \tag{3.3}$$

Notice that $D_\alpha(\boldsymbol{\upsilon}) = D(\boldsymbol{\upsilon}) + \alpha\mathbf{f}^T\boldsymbol{\upsilon}$. In particular, when $\mathbf{S}$ is invertible, $D_\alpha(\cdot)$ is strongly convex (and, in this case, a solution to (3.3) is thus necessarily unique). We also recall that, for a given α, the set of solutions to (3.3) is convex. The gradient of D_α at $\boldsymbol{\upsilon} \in \mathbb{R}^N$ is given by $\nabla D_\alpha(\boldsymbol{\upsilon}) = \mathbf{S}(\boldsymbol{\upsilon} - \boldsymbol{\omega}) + \alpha\mathbf{f}$. The following Lemma 3.1 recalls some simple properties verified by the solutions to problem (3.3).

Lemma 3.1. *Denote by $\boldsymbol{\upsilon}_\alpha^*$ a solution to* (3.3) *with $\alpha \geqslant 0$, we have:*

(i) for $\alpha = 0$, $\boldsymbol{\upsilon}_\alpha^* = \boldsymbol{\omega}$ is a solution to (3.3)*,*

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

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

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

(v) if $\widetilde{\boldsymbol{\upsilon}}_\alpha^*$ is another solution to (3.3)*, then $\mathbf{S}(\boldsymbol{\upsilon}_\alpha^* - \widetilde{\boldsymbol{\upsilon}}_\alpha^*) = 0$ and $\mathbf{f}^T\widetilde{\boldsymbol{\upsilon}}_\alpha^* = \mathbf{f}^T\boldsymbol{\upsilon}_\alpha^*$,*

(vi) if $[\alpha\mathbf{f} - \mathbf{S}\boldsymbol{\omega}]_k \geqslant 0$ and $\mathbf{S}_{k,k} > 0$ (see Remark 3.2), then $[\boldsymbol{\upsilon}_\alpha^*]_k = 0$,

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

Since $\boldsymbol{\upsilon} \geqslant 0$, the term $\mathbf{f}^T\boldsymbol{\upsilon}$ can be interpreted as a weighted ℓ^1-type regularisation (and α is the regularisation parameter). For appropriate $\mathbf{f}$ and α, we can thus expect a solution $\boldsymbol{\upsilon}_\alpha^*$ to (3.3) to be sparse (see, e.g., [11]). This intuition is confirmed by Lemma 3.1-(vi); indeed, when all the diagonal entries of $\mathbf{S}$ are strictly positive, $\mathrm{card}(\{k | [\alpha\mathbf{f} - \mathbf{S}\boldsymbol{\omega}]_k < 0\})$ gives an upper bound on the number of strictly positive components of $\boldsymbol{\upsilon}_\alpha^*$ (notice that this bound is generally not tight); remark that assertion (vi) is a direct corollary of assertion (iv) (since $\mathbf{S}\boldsymbol{\upsilon}_\alpha^* \geqslant 0$).

Remark 3.2. 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 (i.e., $\mathrm{diag}(\mathbf{K}) > 0$); we recall that for all $x \in \mathscr{X}$, we have $K(x, x) = \|K_x\|_{\mathcal{H}}^2 \geqslant 0$. This assumption is in practice not restrictive at all: indeed, if $K(x_k, x_k) = 0$, then $K_{x_k} = 0$ and thus $h(x_k) = 0$ for all $h \in \mathcal{H}$. In the framework of equation (3.2) (i.e., μ and ν are discrete measures supported by S), such a point x_k may thus be removed from the sample S without inducing any modification of the operators T_μ and T_ν. ◁

3.3. Constrained squared-kernel-discrepancy minimisation. Instead of considering problem (3.3), we can equivalently introduce, for $\varkappa \geqslant 0$ (and, in practice, $\varkappa \leqslant \mathbf{f}^T\boldsymbol{\omega}$, see Lemma 3.2)

$$\underset{\boldsymbol{\upsilon}}{\text{minimise}}\ D(\boldsymbol{\upsilon}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\upsilon})^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}) \text{ subject to } \boldsymbol{\upsilon} \geqslant 0 \text{ and } \mathbf{f}^T\boldsymbol{\upsilon} = \varkappa. \tag{3.4}$$

Notice that problem (3.4) consists in minimising a convex function on a convex compact domain, so that a solution $\boldsymbol{\upsilon}_\varkappa^*$ to problem (3.4) always exists; in particular, if $\mathbf{S}$ is non-singular, then $\boldsymbol{\upsilon}_\varkappa^*$ is always unique (since $D(\cdot)$ is in this case strongly convex).

Lemma 3.2. *Let $\boldsymbol{\upsilon}_\alpha^*$ be a solution to problem* (3.3) *with $\alpha \geqslant 0$; then $\boldsymbol{\upsilon}_\alpha^*$ is a solution to problem* (3.4) *with $\varkappa = \mathbf{f}^T\boldsymbol{\upsilon}_\alpha^*$. Reciprocally, assume that $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to problem* (3.4) *with $0 < \varkappa \leqslant \mathbf{f}^T\boldsymbol{\omega}$; then $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to problem* (3.3) *with $\alpha = (\boldsymbol{\upsilon}_\varkappa^*)^T\mathbf{S}(\boldsymbol{\omega} - \boldsymbol{\upsilon}_\varkappa^*)/\varkappa$. For $\varkappa = 0$, we have $\boldsymbol{\upsilon}_\varkappa^* = 0$, which is solution to problem* (3.3) *with $\alpha \geqslant \max_k \left\{[\mathbf{S}\boldsymbol{\omega}]_k / f_k\right\}$. For $0 \leqslant \varkappa \leqslant \mathbf{f}^T\boldsymbol{\omega}$, the map $\varkappa \mapsto D(\boldsymbol{\upsilon}_\varkappa^*)$ is decreasing.*

As an interesting feature, problem (3.4) 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 6.1. A sequential strategy (based on the notion of regularisation path) for solving problems (3.3) and (3.4) is also discussed in Section 5

Notice that, in view of Lemma 3.1-(iii) and Lemma 3.2, considering $\varkappa = \rho \mathbf{f}^T \boldsymbol{\omega}$ with $\rho \in [0,1]$ appears as a natural parameterisation for problem (3.4).

4. Analogy with one-class SVM. Following for instance [18], problems (3.3) and (3.4) can be interpreted as the dual formulations of *one-class distorted SVMs* (or discrepancy-SVMs) related to the squared kernel and the initial discrete measure μ. Soft-margin-type extensions of the one-class SVMs introduced in Sections 4.1 and 4.2 are also discussed in Appendix B.

We denote by $\mathcal{G}$ the RKHS associated with the squared kernel $K^2(\cdot,\cdot)$. We introduce the function $g_\mu \in \mathcal{G}$, defined by $g_\mu(x) = \int_{\mathscr{X}} K^2(t,x)\mathrm{d}\mu(t)$, with $\mu = \sum_{k=1}^N \omega_k \delta_{x_k}$; we shall refer to g_μ as the *primal distortion term*, see also (2.1).

4.1. One-class SVM related to the regularised problem. We first describe the SVM related to problem (3.3). For $g \in \mathcal{G}$, we consider the convex minimisation problem

$$\begin{array}{ll} \underset{g}{\text{minimise}} & \frac{1}{2}\|g\|_{\mathcal{G}}^2 + (g|g_\mu)_{\mathcal{G}} \\ \text{subject to} & g(x_k) \geqslant -\alpha f_k \text{ for all } k \in \{1,\cdots,N\}. \end{array} \tag{4.1}$$

The application $g \mapsto \|g\|_{\mathcal{G}}^2$ being strictly convex, a solution to problem (4.1) is necessarily unique.

Lemma 4.1. *If $\boldsymbol{\upsilon}_\alpha^*$ is a solution to* (3.3) *with $\alpha \geqslant 0$, then $g_\alpha^*(x) = \sum_{k=1}^N [\boldsymbol{\upsilon}_\alpha^* - \boldsymbol{\omega}]_k K^2(x,x_k)$ is the solution to* (4.1)*. For all $k \in \{1,\cdots,N\}$ such that $[\boldsymbol{\upsilon}_\alpha^*]_k > 0$, we have $g_\alpha^*(x_k) = -\alpha f_k$.*

Notice that for all k, we have $g_\alpha^*(x_k) = [\mathbf{S}(\boldsymbol{\upsilon}_\alpha^* - \boldsymbol{\omega})]_k$. By introducing the change of variable $\check{g} = g + g_\mu \in \mathcal{G}$, problem (4.1) leads to (up to an additive constant)

$$\begin{array}{ll} \underset{\check{g}}{\text{minimise}} & \frac{1}{2}\|\check{g}\|_{\mathcal{G}}^2 \\ \text{subject to} & \check{g}(x_k) \geqslant g_\mu(x_k) - \alpha f_k \text{ for all } k \in \{1,\cdots,N\}, \end{array} \tag{4.2}$$

which is an equivalent formulation for (4.1), with solution $\check{g}_\alpha^*(x) = \sum_{k=1}^N [\boldsymbol{\upsilon}_\alpha^*]_k K^2(x,x_k)$. In view of (2.1), if we denote by ν_α^* the discrete measure supported by $\mathcal{S}$ related to a solution $\boldsymbol{\upsilon}_\alpha^*$ to problem (3.3), then $\check{g}_\alpha^* = g_{\nu_\alpha^*}$. Problem (4.2) thus gives an interesting insight on the role of the penalisation vector $\mathbf{f}$, see Section 8 for a further discussion.

Remark 4.1. Problem (4.1) may also be considered in the framework of (3.1), 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 $\mathcal{G}_{\mathcal{S}} = \operatorname{span}\{K_{x_k}^2\}_{k=1}^N$ of $\mathcal{G}$, and an additive constant relative to the term $\|K\|_{L^2(\mu\otimes\mu)}^2$ is required. ◁

4.2. One-class SVM related to the constrained problem. We now describe the SVM related to problem (3.4). For $g \in \mathcal{G}$ and $\gamma \in \mathbb{R}$, we introduce the problem

$$\begin{array}{ll} \underset{g,\gamma}{\text{minimise}} & \frac{1}{2}\|g\|_{\mathcal{G}}^2 + (g|g_\mu)_{\mathcal{G}} - \gamma \\ \text{subject to} & g(x_k) \geqslant \gamma f_k/\varkappa \text{ for all } k \in \{1,\cdots,N\}. \end{array} \tag{4.3}$$

Again, a solution to problem (4.3) is necessarily unique.

Lemma 4.2. *If $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to* (3.4)*, then $g_\varkappa^*(x) = \sum_{k=1}^N [\boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega}]_k K^2(x,x_k)$ and $\gamma_\varkappa^* = (\boldsymbol{\upsilon}_\varkappa^*)^T \mathbf{S}(\boldsymbol{\upsilon}_\varkappa^* - \boldsymbol{\omega})$ is the solution to* (4.3)*. For all $k \in \{1,\cdots,N\}$ such that $[\boldsymbol{\upsilon}_\varkappa^*]_k > 0$, we have $g_\varkappa^*(x_k) = \gamma_\varkappa^* f_k/\varkappa$.*

From Lemma 4.2, we have $\gamma_\varkappa^* = \|g_\varkappa^*\|_{\mathcal{G}}^2 + (g_\varkappa^*|g_\mu)_{\mathcal{G}}$. In view of Lemma 3.2, for $0 < \varkappa \leqslant \mathbf{f}^T\boldsymbol{\omega}$, we know that $\boldsymbol{\upsilon}_\varkappa^*$ is a solution to (3.3) for $\alpha = -\gamma_\varkappa^*/\varkappa$, and since $\alpha \geqslant 0$, we therefore have $\gamma_\varkappa^* \leqslant 0$.

5. Regularisation path. Considering the framework of Section 3 and following the classical results relative to the regularisation paths for Lasso or SVM models (see e.g., [15, 10]), we now discuss the regularisation paths related to problems (3.3) and (3.4). In what follows, we mainly consider problem (3.3) (i.e., the regularised model); results related to problem (3.4) can then be obtained from Lemma 3.2.

5.1. Generalities. Let $\boldsymbol{v}_\alpha^*$ be a solution to (3.3) for $\alpha \geqslant 0$; we introduce the index sets

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

so that, by definition, $[\nabla D_\alpha(\boldsymbol{v}_\alpha^*)]_k > 0$ for all $k \in J_\alpha^c$. From Lemma 3.1, the index set J_α is unique, even when the solution to (3.3) is not (i.e., in case of non-uniqueness of the solution, J_α does not depends on the solution $\boldsymbol{v}_\alpha^*$ considered). We shall refer to J_α as the *sparsity pattern* of the solutions to problem (3.3) for $\alpha \geqslant 0$. We also recall that for all k such that $[\boldsymbol{v}_\alpha^*]_k > 0$, we necessarily have $k \in J_\alpha$, see for instance Lemma 3.1-(iv).

Knowing J_α, a solution $\boldsymbol{v}_\alpha^* \geqslant 0$ to (3.3) is characterised by the conditions

$$[\boldsymbol{v}_\alpha^*]_{J_\alpha^c} = 0, \text{ and } \mathbf{S}_{J_\alpha, J_\alpha} [\boldsymbol{v}_\alpha^*]_{J_\alpha} = [\mathbf{S}\boldsymbol{\omega}]_{J_\alpha} - \alpha \mathbf{f}_{J_\alpha}, \tag{5.1}$$

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 $n_\alpha = \operatorname{card}(J_\alpha)$, and where, for instance, $\mathbf{f}_{J_\alpha} \in \mathbb{R}^{n_\alpha}$ stands for the vector defined by the components of $\mathbf{f}$ with index in J_α.

Lemma 5.1. *Let $\boldsymbol{v}_{\alpha_1}^*$ and $\boldsymbol{v}_{\alpha_2}^*$ be solutions to problem* (3.3) *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* (3.3) *with $\alpha = \theta\alpha_1 + (1-\theta)\alpha_2$ and $J_\alpha = J$.*

If, for all $\alpha \geqslant 0$, the solution $\boldsymbol{v}_\alpha^*$ is unique (which is for instance the case when $\mathbf{S}$ is invertible), then in view of Lemma 5.1, the *regularisation map* $R : \alpha \mapsto \boldsymbol{v}_\alpha^*$ is a piecewise linear application from $\mathbb{R}_+$ into $\mathbb{R}^N$. In case of uniqueness of the solutions, the regularisation map is therefore piecewise continuous and has right and left limits for all α, these limits satisfying the optimality conditions for problem (3.3); by uniqueness of the solution, this entails that the map $R : \alpha \mapsto \boldsymbol{v}_\alpha^*$ is continuous. In case of non-uniqueness of the solution, Lemma 5.1 shows that the set of solutions related to a same sparsity pattern J is convex.

When α decreases or increases, we refer to a change in the sparsity pattern J_α as an *event*, and we call *kinks* the values of α where an event occurs. In particular, notice that since there cannot exists more than 2^N different subsets of $\{1, \cdots, N\}$, Lemma 5.1 implies that the number M_{ev} of events related to problem (3.3) necessarily satisfies $M_{ev} \leqslant 2^N - 1$ (i.e., M_{ev} is the number of kinks).

In the general case (i.e., even in case of non-uniqueness of the solutions), when α decreases, we can easily define the value α_0 and the underlying index set J_{α_0} at which the first event occurs. Indeed, from Lemma 3.1, we know that for $\alpha \geqslant \max_k [\mathbf{S}\boldsymbol{\omega}]_k / f_k$, we have $\boldsymbol{v}_\alpha^* = 0$. Therefore, the first event occurs at $\alpha_0 = \max_k [\mathbf{S}\boldsymbol{\omega}]_k / f_k$, and we have $J_{\alpha_0} = \{k | [\mathbf{S}\boldsymbol{\omega}]_k / f_k = \alpha_0\}$ (for larger values of α, the underlying sparsity pattern is the empty set). As detailed in the following Section 5.2, if the submatrix $\mathbf{S}_{J_{\alpha_0}, J_{\alpha_0}}$ is invertible, we can easily compute the value α_1 at which the next event occurs and obtain the underlying sparsity pattern J_{α_1}.

Remark 5.1. Lemma 5.1 can be generalised to regularised problems involving a general penalisation term $\mathbf{r} \in \mathbb{R}^N$, with $\mathbf{r} \geqslant 0$, i.e.,

$$\underset{\boldsymbol{v}}{\text{minimise}}\ D_{\mathbf{r}}(\boldsymbol{v}) = \tfrac{1}{2}(\boldsymbol{\omega} - \boldsymbol{v})^T \mathbf{S}(\boldsymbol{\omega} - \boldsymbol{v}) + \mathbf{r}^T \boldsymbol{v} \text{ subject to } \boldsymbol{v} \geqslant 0. \tag{5.2}$$

We denote by $\boldsymbol{v}_{\mathbf{r}}^*$ a solution to (5.2). Note that problem (3.3) thus consists in the particular case $\mathbf{r} = \alpha \mathbf{f}$. We can then define $J_{\mathbf{r}} = \{k | [\nabla D_{\mathbf{r}}(\boldsymbol{v}_{\mathbf{r}}^*)]_k = 0\}$ and $J_{\mathbf{r}}^c = \{1, \cdots, N\} \backslash J_{\mathbf{r}}$, and knowing $J_{\mathbf{r}}$, a solution $\boldsymbol{v}_{\mathbf{r}}^* \geqslant 0$ to (5.2) is characterised by the conditions

$$[\boldsymbol{v}_{\mathbf{r}}^*]_{J_{\mathbf{r}}^c} = 0, \text{ and } \mathbf{S}_{J_{\mathbf{r}}, J_{\mathbf{r}}} [\boldsymbol{v}_{\mathbf{r}}^*]_{J_{\mathbf{r}}} = [\mathbf{S}\boldsymbol{\omega}]_{J_{\mathbf{r}}} - \mathbf{r}_{J_{\mathbf{r}}}$$

From exactly the same argument as in Lemma 5.1, we obtain that if two penalisation terms $\mathbf{r}_1$ and $\mathbf{r}_2 \geqslant 0$ are such that $J_{\mathbf{r}_1} = J_{\mathbf{r}_2} = J$, then for all $\theta \in [0, 1]$, $\boldsymbol{v}_{\mathbf{r}}^* = \theta \boldsymbol{v}_{\mathbf{r}_1}^* + (1-\theta)\boldsymbol{v}_{\mathbf{r}_2}^*$ is a solution to problem (5.2) with $\mathbf{r} = \theta\mathbf{r}_1 + (1-\theta)\mathbf{r}_2$ and $J_{\mathbf{r}} = J$. ◁

5.2. Regularisation direction for non-singular submatrix. We now discuss, in case of uniqueness of the solutions, the computation of the regularisation path for decreasing values of α, i.e., we assume that the events occurs successively at $\alpha_0 = \max_k [\mathbf{S}\boldsymbol{\omega}]_k / f_k > \alpha_1 > \cdots > \alpha_{M_{ev}-1} \geqslant 0$, the last

event corresponding to the largest value of α such that $J_\alpha = \{1, \cdots, N\}$, since for $\alpha = 0$, we have $\upsilon^*_\alpha = \omega > 0$.

More precisely, considering a kink α_p with related sparsity pattern J_{α_p} (for $p \in \{0, \cdots, M_{ev} - 2\}$, i.e., $J_{\alpha_p} \neq \{1, \cdots, N\}$, i.e., $n_{\alpha_p} = \mathrm{card}(J_{\alpha_p}) < N$) and assuming that the submatrix $\mathbf{S}_{J_{\alpha_p}, J_{\alpha_p}}$ is invertible, we describe how to compute the value $\alpha_{p+1} < \alpha_p$ corresponding to the next event, and how to characterise the related sparsity pattern $J_{\alpha_{p+1}}$. For simplicity, we use the notation $J = J_{\alpha_p}$. We recall that, by definition, α_p is the largest value of α such that $J_\alpha = J$.

From (5.1), we introduce the vector υ_α such that $[\upsilon_\alpha]_{J^c} = 0$ and $[\upsilon_\alpha]_J = \mathbf{S}^{-1}_{J,J}([\mathbf{S}\omega]_J - \alpha \mathbf{f}_J)$; the vector υ_α is sometime referred to as the *regularisation direction*. By definition, α_{p+1} corresponds to the smallest α such that $0 \leqslant \alpha < \alpha_p$ and

$$[\upsilon_\alpha]_J \geqslant 0 \text{ and } [\mathbf{S}(\upsilon_\alpha - \omega) + \alpha \mathbf{f}]_{J^c} \geqslant 0. \tag{5.3}$$

The set $J_{\alpha_{p+1}}$ is then obtained by removing from $J = J_{\alpha_p}$ all the indices $k \in J$ such that $[\upsilon_{\alpha_{p+1}}]_k = 0$, and by adding all the indices $k \in J^c$ such that $[\nabla D_{\alpha_{p+1}}(\upsilon_{\alpha_{p+1}})]_k = 0$; see Lemma 5.2 for more details concerning the computation of α_{p+1}.

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

Lemma 5.2. *Consider a kink α_p with related sparsity pattern $J_{\alpha_p} = J \neq \{1, \cdots, N\}$, and assume that the submatrix $\mathbf{S}_{J,J}$ is invertible. We introduce the $(N - n_{\alpha_p}) \times n_{\alpha_p}$ matrix $\mathbf{M} = \mathbf{S}_{J^c,J}\mathbf{S}^{-1}_{J,J}$, and we define*

$$\alpha_+ = \max_l \left\{ \left[\mathbf{M}[\mathbf{S}\omega]_J - [\mathbf{S}\omega]_{J^c}\right]_l / \left[\mathbf{M}\mathbf{f}_J - \mathbf{f}_{J^c}\right]_l \middle| \left[\mathbf{M}\mathbf{f}_J - \mathbf{f}_{J^c}\right]_l < 0 \right\}, \text{ and}$$

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

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

In view of Lemma 5.2, once α_p and $J = J_{\alpha_p}$ are known, the computation of the next event (i.e., of α_{p+1} and $J_{\alpha_{p+1}}$) requires the calculation of $\mathbf{S}^{-1}_{J,J}\mathbf{f}_J$ and $\mathbf{S}^{-1}_{J,J}[\mathbf{S}\omega]_J$ (i.e., the resolution of a linear system). Starting "from scratch" (i.e., without taking into account the computations already performed to obtain the information relative to the kink α_p) and using a direct method (by for instance considering the Cholesky decomposition of the symmetric and positive-definite matrix $\mathbf{S}_{J,J}$), the amount of computations required for this operation scale as $\mathcal{O}(n^3_{\alpha_p})$. It is however possible to reduce the computational complexity by considering update formulae (by for instance iteratively updating the Cholesky decomposition of $\mathbf{S}_{J,J}$); in the favourable cases, the computational complexity may thus be reduced to $\mathcal{O}(n^2_{\alpha_p})$ (while still consisting in a direct approach). In order to further reduce the computational cost, an alternative may also consists in using an indirect iterative approach, like for example conjugate gradient methods; however, numerical errors may quickly lead to precision issues. Finally, the complexity of the two matrix vector products involving the matrix $\mathbf{S}_{J^c,J}$ scale as $\mathcal{O}(n_{\alpha_p}(N - n_{\alpha_p}))$. As a result, the computation of the regularisation direction thus becomes intractable once large values of n_{α_p} are reached. When N is large, the regularisation-path strategy may therefore only be used to explore the range of very sparse approximate measures ν.

The determination of the path is also extremely sensitive to numerical errors. For instance, very small value of the gap $\alpha_p - \alpha_{p+1}$ between two consecutive kinks or the simultaneous entry or exit of indices of the sparsity pattern may lead to numerical precision issues. See Sections 9 and 10 for illustrations.

6. Numerical solver for the constrained problem. In this section, we discuss a numerically tractable strategy to compute approximate solutions to problem (3.4) (i.e., the constrained problem) for any given 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.

6.1. Vertex-exchange QP solver. Consider problem (3.4); for $\varkappa > 0$, we can define the change of variable $\widetilde{\boldsymbol{\upsilon}} = \mathbf{D}\boldsymbol{\upsilon}$, with $\mathbf{D} = \mathrm{diag}(\mathbf{d})$, and $\mathbf{d} = (d_1, \cdots, d_N)^T = \mathbf{f}/\varkappa$, i.e., $\mathbf{D}$ is a diagonal matrix with i-th diagonal entry d_i (so that $\mathbf{d} = \mathbf{D}\mathbb{1}$). In this way, problem (3.4) is turned into (up to an additive constant), for $\widetilde{\boldsymbol{\upsilon}} \in \mathbb{R}^N$,

$$\underset{\widetilde{\boldsymbol{\upsilon}}}{\text{minimise}}\ C(\widetilde{\boldsymbol{\upsilon}}) = \tfrac{1}{2}\widetilde{\boldsymbol{\upsilon}}^T\mathbf{A}\widetilde{\boldsymbol{\upsilon}} - \mathbf{b}^T\widetilde{\boldsymbol{\upsilon}} \text{ subject to } \widetilde{\boldsymbol{\upsilon}} \geqslant 0 \text{ and } \mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1, \tag{6.1}$$

with $\mathbf{A} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$ and $\mathbf{b} = \mathbf{D}^{-1}\mathbf{S}\boldsymbol{\omega}$. We refer to (6.1) as the *canonical QP* related to the constrained squared-kernel-discrepancy minimisation (3.4). Since $\mathbf{A}_{i,j} = K^2(x_i, x_j)/(d_i d_j)$, any entry of $\mathbf{A}$ can be easily obtained from only the knowledge of the squared kernel $K^2(\cdot,\cdot)$, the set $\mathcal{S}$ and the vector $\mathbf{d}$. Importantly, we shall therefore not store the matrix $\mathbf{A}$, but rather compute on the fly any required entry of $\mathbf{A}$; this way, problems involving a large N may be considered. Notice that since $\boldsymbol{\omega}$ is non-sparse, for large N, the computation of the *dual distortion term* $\mathbf{S}\boldsymbol{\omega}$ is computationally demanding ($\mathcal{O}(N^2)$ complexity), but it may be parallelised; approximation of the underlying kernel-matrix vector product may also be obtained by a generalized fast multipole method (see [19] and references therein). Once $\mathbf{b}$ is known, the gradient $\nabla C(\widetilde{\boldsymbol{\upsilon}}) = \mathbf{A}\widetilde{\boldsymbol{\upsilon}} - \mathbf{b}$ can be easily obtained for any sparse feasible $\widetilde{\boldsymbol{\upsilon}}$.

The extreme points of the polytopes defined by the constraints in (6.1) are the vectors $\{\mathbf{e}_i\}_{i=1}^N$, where $\mathbf{e}_i \in \mathbb{R}^N$ is the i-th element of the canonical basis of $\mathbb{R}^N$ (that is $[\mathbf{e}_i]_i = 1$, all the other components being equal to zero). For a feasible $\widetilde{\boldsymbol{\upsilon}}$, let $I_{\widetilde{\boldsymbol{\upsilon}}} = \{k | \widetilde{\upsilon}_k > 0\}$ be the index set of all strictly positive components of $\widetilde{\boldsymbol{\upsilon}}$. An iteration of the vertex-exchange algorithm consists in searching ($\mathcal{O}(N)$ complexity)

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

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 inexpensive, 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 vertex-exchange strategy thus appears as a interesting candidate to scale up to relatively large N and may be used as a complement of the regularisation-path strategy described in Section 5.

Denoting by $\widetilde{\boldsymbol{\upsilon}}^*$ a solution to (6.1), the convergence of the vertex-exchange algorithm can be easily verified (see, e.g., [8]) by simply remarking that since $\widetilde{\boldsymbol{\upsilon}} \geqslant 0$ and $\mathbb{1}^T\widetilde{\boldsymbol{\upsilon}} = 1$, by definition of j^*, we have $\widetilde{\boldsymbol{\upsilon}}^T\nabla C(\widetilde{\boldsymbol{\upsilon}}) \leqslant \mathbf{e}_{j^*}^T\nabla C(\widetilde{\boldsymbol{\upsilon}})$, so that

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

and these inequalities can also be used to check distance from optimality. In Sections 9 and 10, the accuracy of an approximate solution $\widetilde{\boldsymbol{\upsilon}}$ is indicated by $\epsilon = (\widetilde{\boldsymbol{\upsilon}} - \mathbf{e}_{i^*})^T\nabla C(\widetilde{\boldsymbol{\upsilon}})$ (Frank-Wolfe error bound).

6.2. Enhancing sparsity through components merging. The canonical QP formulation introduced in Section 6.1 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 (6.1), or any vector related to an interesting low-discrepancy configuration $\boldsymbol{\upsilon}$ through the change of variable $\widetilde{\boldsymbol{\upsilon}} = \mathbf{D}\boldsymbol{\upsilon}$, with $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ and $\mathbf{d} = \mathbf{f}/(\mathbf{f}^T\boldsymbol{\upsilon})$, see Section 6.1. 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 9 and 10), 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 refer to this operation as the $\{i,j\}$-merging of $\widetilde{\boldsymbol{\upsilon}}$. We have

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

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

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 on the computational power at disposal, we may either consider

- *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 a "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 $\{\widetilde{\boldsymbol{\upsilon}}_{[0]}, \widetilde{\boldsymbol{\upsilon}}_{[1]}, \cdots, \widetilde{\boldsymbol{\upsilon}}_{[n_0-1]}\}$, 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 $\boldsymbol{\upsilon} = \mathbf{D}^{-1}\widetilde{\boldsymbol{\upsilon}}_0$, we may consider a sparser vector $\boldsymbol{\upsilon}_{[m]} = \mathbf{D}^{-1}\widetilde{\boldsymbol{\upsilon}}_{[m]}$ (notice that $\boldsymbol{\upsilon}_{[m]}$ and $\widetilde{\boldsymbol{\upsilon}}_{[m]}$ have the same number of strictly positive components); see Sections 9 and 10 for illustrations.

7. Approximate eigendecomposition. This section is devoted to a discussion relative to the accuracy of the approximation of the main eigenpairs of the integral operator T_μ induced by the eigendecomposition of the approximate operator T_ν.

7.1. Eigendecomposition of the approximate operator. Let $\nu \sum_{k=1}^N \upsilon_k \delta_{x_k}$ be a discrete measure with support included in S, with related vector $\boldsymbol{\upsilon} \geqslant 0$; we assume that $\boldsymbol{\upsilon} \neq 0$.

We introduce the index set $I = \{i|\upsilon_i > 0\}$ and let $n = \operatorname{card}(I)$ be the number of strictly positive components of $\boldsymbol{\upsilon}$; we have in particular $\nu = \sum_{i\in I}\upsilon_i\delta_{x_i}$ (i.e., we have discarded all the points x_k such that $\upsilon_k = 0$, since ν is supported by only n points). Following Section 2.2, 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 easily obtained from the eigendecomposition of the $n\times n$ (symmetric and positive-semidefinite) principal submatrix $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I}$. We thus in particular obtain the o.n.b. $\{\sqrt{\vartheta_l}\psi_l | l \in \mathbb{I}_\nu^+\}$ of the subspace $\mathcal{H}_\nu$ of $\mathcal{H}$ related to T_ν.

As already mentioned, by definition and in view of Lemma 2.2, when $D_{K^2}(\mu,\nu)$ is small, we can expect the main eigendirections of the operator T_ν (the approximate eigendirection) to be relevant approximation of the main eigendirections of the operator T_μ.

7.2. Accuracy of the approximate eigendirections. For simplicity, we assume that $\mathcal{H}_\nu \subset \mathcal{H}_\mu$. We recall that we denote by $\{\sqrt{\lambda_k}\varphi_k | k \in \mathbb{I}_\mu^+\}$ the o.n.b. of $\mathcal{H}_\mu$ related to T_μ; the eigenfunctions φ_k being orthonormal in $L^2(\mu)$. For $l \in \mathbb{I}_\nu^+$, we introduce the *normalised approximate eigenfunctions* of T_μ induced by T_ν, given by

$$\widehat{\varphi}_l = \psi_l / \|\psi_l\|_{L^2(\mu)}. \tag{7.1}$$

The normalised approximate eigenfunctions are such that $\|\widehat{\varphi}_l\|_{L^2(\mu)} = 1$, but contrary to the true eigenfunctions φ_k, they are not necessarily orthogonal in $L^2(\mu)$. Notice that since $\mathcal{H}_\nu \subset \mathcal{H}_\mu$, we necessarily have $\|\psi_l\|_{L^2(\mu)} > 0$; in addition, by definition of T_μ,

$$\|\psi_l\|^2_{L^2(\mu)} = \big(\psi_l \big| T_\mu[\psi_l]\big)_{\mathcal{H}}. \tag{7.2}$$

Controlling the orthogonality, in $L^2(\mu)$, between the approximate eigenfunctions $\widehat{\varphi}_l$ appears as a relatively inexpensive way to evaluate the accuracy of the approximate eigendirections. Indeed,

accurate approximate eigenfunctions $\widehat{\varphi}_l$ should be almost mutually orthogonal in $L^2(\mu)$; this condition is however only a necessary condition. It is also very instructive to try to estimate the eigenvalues, for the operator T_μ, related to the approximate eigendirections ψ_l induced by T_ν, as discussed hereafter.

For all $k \in \mathbb{I}_\mu^+$, we have $\|\sqrt{\lambda_k}\varphi_k\|_{\mathcal{H}}^2 = 1$. By analogy, for all $l \in \mathbb{I}_\nu^+$, we may define $\widehat{\lambda}_l^{[1]}$ such that $\|\sqrt{\widehat{\lambda}_l^{[1]}}\widehat{\varphi}_l\|_{\mathcal{H}}^2 = 1$ (so that $\sqrt{\widehat{\lambda}_l^{[1]}}\widehat{\varphi}_l = \sqrt{\vartheta_l}\psi_l$). Thus, using (7.1) and (7.2), we obtain

$$\widehat{\lambda}_l^{[1]} = 1/\|\widehat{\varphi}_l\|_{\mathcal{H}}^2 = \vartheta_l\|\psi_l\|_{L^2(\mu)}^2 = \big(\sqrt{\vartheta_l}\psi_l\big|T_\mu[\sqrt{\vartheta_l}\psi_l]\big)_{\mathcal{H}} = \big(T_\nu[\psi_l]\big|T_\mu[\psi_l]\big)_{\mathcal{H}}. \tag{7.3}$$

If ψ_l is a true eigendirection of T_μ, then $\widehat{\lambda}_l^{[1]}$ corresponds to the associated eigenvalue. From the Cauchy-Schwarz inequality, we have

$$\big(\sqrt{\vartheta_l}\psi_l\big|T_\mu[\sqrt{\vartheta_l}\psi_l]\big)_{\mathcal{H}} \leqslant \|\sqrt{\vartheta_l}\psi_l\|_{\mathcal{H}}\big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}} = \big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}},$$

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

$$\widehat{\lambda}_l^{[2]} = \big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}}. \tag{7.4}$$

The ratio $\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}$ corresponds to the inner product, in $\mathcal{H}$, between the normalised vectors $\sqrt{\vartheta_l}\psi_l$ and $T_\mu[\sqrt{\vartheta_l}\psi_l]/\big\|T_\mu[\sqrt{\vartheta_l}\psi_l]\big\|_{\mathcal{H}}$, and we have $0 < \widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]} \leqslant 1$. See Remark 7.1 for a discussion relative to the computation of the approximate eigenvalues $\widehat{\lambda}_l^{[\cdot]}$.

Following (7.3) while considering the Hilbert structure of $L^2(\mu)$ instead of the one of $\mathcal{H}$, we can also define

$$\widehat{\lambda}_l^{[3]} = \big(\widehat{\varphi}_l\big|T_\mu[\widehat{\varphi}_l]\big)_{L^2(\mu)} = \big\|T_\mu[\widehat{\varphi}_l]\big\|_{\mathcal{H}}^2 = \big(\widehat{\lambda}_l^{[2]}\big)^2/\widehat{\lambda}_l^{[1]}, \tag{7.5}$$

so that $\widehat{\lambda}_l^{[2]} \leqslant \widehat{\lambda}_l^{[3]}$, with equality when ψ_l is an eigendirection of T_μ. Finally, from the Cauchy-Schwarz inequality and by analogy with (7.4), we have

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

suggesting the introduction of

$$\widehat{\lambda}_l^{[4]} = \big\|T_\mu[\widehat{\varphi}_l]\big\|_{L^2(\mu)}. \tag{7.6}$$

We therefore have $\widehat{\lambda}_l^{[3]} \leqslant \widehat{\lambda}_l^{[4]}$, with, again, equality when ψ_l is an eigendirection of T_μ. 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 two normalised vectors $\widehat{\varphi}_l$ and $T_\mu[\widehat{\varphi}_l]/\|T_\mu[\widehat{\varphi}_l]\|_{L^2(\mu)}$.

Remark 7.1. For $\mu = \sum_{k=1}^N \omega_k\delta_{x_k}$ and once ϑ_l and ψ_l are known, the amount of computations required to compute $\widehat{\lambda}_l^{[1]} = \vartheta_l\|\psi_l\|_{L^2(\mu)}^2$ scales as $\mathcal{O}(N)$, so that computing the approximate eigenvalues $\widehat{\lambda}_l^{[1]}$ is relatively inexpensive on a numerical point of view.

Computing $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ or $\widehat{\lambda}_l^{[4]}$ requires in particular the knowledge of $T_\mu[\psi_l]$, which consists in performing a kernel-matrix vector product, with complexity scaling as $\mathcal{O}(N^2)$; this operation is therefore costly, but may however be easily parallelised (and an approximation may for instance also be obtained by a generalised fast multipole method). Once $T_\mu[\psi_l]$ is known, we can obtain $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ by computing an inner product in $L^2(\mu)$, with complexity scaling as $\mathcal{O}(N)$; and we finally obtain $\widehat{\lambda}_l^{[2]}$ thanks to the relation $\widehat{\lambda}_l^{[2]} = \sqrt{\widehat{\lambda}_l^{[1]}\widehat{\lambda}_l^{[3]}}$.

Note that computing $\widehat{\lambda}_l^{[2]}$ directly from the Hilbert structure of $\mathcal{H}$ requires to perform an inner product with numerical complexity scaling as $\mathcal{O}(N^2)$; on a numerical point of view, this operation is therefore not interesting. ◁

Lemma 7.1 summarises the properties of the approximate eigenvalues $\widehat{\lambda}_l^{[1]}, \cdots, \widehat{\lambda}_l^{[4]}$.

Lemma 7.1. *Considering equations* (7.3)–(7.6)*, 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.*

For $\lambda \in \mathbb{R}$, the function

$$\lambda \mapsto \left\| \lambda \sqrt{\vartheta_l} \psi_l - T_\mu[\sqrt{\vartheta_l}\psi_l] \right\|_{\mathcal{H}}^2 = \lambda^2 - 2\lambda \widehat{\lambda}_l^{[1]} + \left(\widehat{\lambda}_l^{[2]}\right)^2 \tag{7.7}$$

reaches its minimum at $\lambda = \widehat{\lambda}_l^{[1]}$; in the same way, the function

$$\lambda \mapsto \left\| \lambda \widehat{\varphi}_l - T_\mu[\widehat{\varphi}_l] \right\|_{L^2(\mu)}^2 = \lambda^2 - 2\lambda \widehat{\lambda}_l^{[3]} + \left(\widehat{\lambda}_l^{[4]}\right)^2 \tag{7.8}$$

reaches its minimum at $\lambda = \widehat{\lambda}_l^{[3]} = \left(\widehat{\lambda}_l^{[2]}\right)^2 / \widehat{\lambda}_l^{[1]}$.

Remark 7.2. Consider any measure $\nu \in \mathscr{T}(K)$; the approximations $\widehat{\varphi}_l$, $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ remain unchanged if one replaces ν by $\theta\nu$ for any $\theta > 0$. ◁

In view of Lemma 7.1, 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 $\widehat{\lambda}_l^{[4]}$ (the ratios $\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}$ and $\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]}$ being of particular interest due to their geometric interpretation, as already discussed). From (7.7) and (7.8), one may for instance consider the quantities

$$\left[\left(\widehat{\lambda}_l^{[2]}\right)^2 - \left(\widehat{\lambda}_l^{[1]}\right)^2\right] / \left(\widehat{\lambda}_l^{[1]}\right)^2 \text{ and } \left[\left(\widehat{\lambda}_l^{[4]}\right)^2 - \left(\widehat{\lambda}_l^{[3]}\right)^2\right] / \left(\widehat{\lambda}_l^{[3]}\right)^2, \tag{7.9}$$

so that the closer the (positive) terms in (7.9) are from zero, the more accurate is the approximate eigendirection ψ_l; see Sections 9 and 10 for illustrations.

8. Penalisation direction. In the framework of Section 3 (i.e., ν is a discrete measure with support included in a fixed set of points S), sparsity of the approximate measure ν (related to the vector $\boldsymbol{\upsilon}$) is induced by the introduction of an ℓ^1-type penalisation based on the term $\mathbf{f}^T\boldsymbol{\upsilon}$ for a given $\mathbf{f} \in \mathbb{R}^N$ with $\mathbf{f} > 0$. The term $\mathbf{f}^T\boldsymbol{\upsilon}$ can be interpreted as the integral with respect to ν of a (measurable) real-valued function f on $\mathscr{X}$ satisfying $f(x_k) = f_k$, so that $\mathbf{f}^T\boldsymbol{\upsilon} = \int_{\mathscr{X}} f(x)\mathrm{d}\nu(x)$. In practice, we aim at obtaining a vector $\boldsymbol{\upsilon}$ which is both as sparse as possible and such that $D(\boldsymbol{\upsilon})$ is as small as possible, naturally raising questions relative to the choice of the penalisation direction $\mathbf{f}$.

Lemma 8.1 (Penalisation direction inducing no sparsity)**.** *If $\mathbf{f} = \theta\mathbf{S}\boldsymbol{\omega}$ with $\theta > 0$, then for $\alpha \leqslant 1/\theta$, $\boldsymbol{\upsilon}_\alpha^* = (1 - \alpha\theta)\boldsymbol{\omega} \geqslant 0$ is a solution to* (3.3) *(and $\boldsymbol{\upsilon}_\alpha^* = 0$ for $\alpha > 1/\theta$).*

Thus, for $\mathbf{f} = \theta\mathbf{S}\boldsymbol{\omega}$, the solutions to (3.3) are non-sparse, and such a choice for the penalisation term $\mathbf{f}$ is of no practical interest. More generally, we can remark that if $\mathbf{f} = \mathbf{S}\boldsymbol{\eta} \geqslant 0$, 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 (3.3).

We recall that, from (4.2), if $\boldsymbol{\upsilon}_\alpha^*$ is a solution to the regularised problem (3.3) (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 $g \in \mathcal{G}$ (with $\mathcal{G}$ the RKHS related to the squared kernel),

$$\underset{g}{\text{minimise}}\ \frac{1}{2}\|g\|_{\mathcal{G}}^2 \text{ subject to } g_\mu(x_k) - g(x_k) \leqslant \alpha f_k \ \text{ for all } k \in \{1, \cdots, N\}, \tag{8.1}$$

with $g_\mu(x_k) = \int_{\mathscr{X}} K^2(t, x_k)\mathrm{d}\mu(t) = [\mathbf{S}\boldsymbol{\omega}]_k$; in addition, $g_\mu(x_k) - g_{\nu_\alpha^*}(x_k) = \alpha f_k$ for all k such that $[\boldsymbol{\upsilon}_\alpha^*]_k > 0$. Also, for any μ and $\nu \in \mathscr{T}(K)$, we have $D_{K^2}(\mu, \nu) = \|g_\mu - g_\nu\|_{\mathcal{G}}^2$, and $(g_\mu | g_\nu)_{\mathcal{G}} = \int_{\mathscr{X}} g_\mu(t)\mathrm{d}\nu(t) = \int_{\mathscr{X}} g_\nu(t)\mathrm{d}\mu(t)$, see (2.1) and Appendix A.

As illustrated in Sections 9 and 10, considering $\mathbf{f} = \mathbb{1}$ generally leads to satisfactory results; it is however also possible to consider model-dependent penalisation directions leading to interesting interpretations. Following Remark 3.2, in the framework of the regularised problem (3.3), we can reasonably assume that $K(x_k, x_k) > 0$ for all k (i.e., $\mathrm{diag}(\mathbf{K}) > 0$), so that, in particular, $\mathbf{S}\boldsymbol{\omega} > 0$ (since $\boldsymbol{\omega} > 0$). In what follows, we discuss penalisation schemes related to the vectors $\mathbf{S}\boldsymbol{\omega}$ and $\mathrm{diag}(\mathbf{K})$. In practice, notice that we must always ensure that the considered penalisation direction do not coincide with the pathological case described in Lemma 8.1 (for instance, if $\mathbf{K}$ is a circulant matrix and if $\boldsymbol{\omega} \propto \mathbb{1}$, then $\mathbf{S}\boldsymbol{\omega} \propto \mathrm{diag}(\mathbf{K}) \propto \mathbb{1}$).

Distortion-term-based penalisation. In view of (8.1), considering $\mathbf{f} = 1/(\mathbf{S}\boldsymbol{\omega})^p$ with $p > 0$ results in a SVM model 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 view of (3.3), the positive components of $\boldsymbol{\upsilon}^*_\alpha$ are then more likely to correspond to large values of $g_\mu(x_k)$. Since, for any measure ν supported by S (with related weights υ_k), we have $(g_\mu|g_\nu)_\mathcal{G} = \sum_{k=1}^N \upsilon_k g_\mu(x_k)$, considering a penalisation of the form $\mathbf{f} = 1/(\mathbf{S}\boldsymbol{\omega})^p$ with $p > 0$ tends to promote the collinearity, in $\mathcal{G}$, between g_μ and $g_{\nu^*_\alpha}$ (by promoting a large value of their inner product).

Kernel-diagonal-based penalisation. We can first remark that for $\mathbf{f} = \mathrm{diag}(\mathbf{K})$, we have $\mathbf{f}^T\boldsymbol{\upsilon} = \mathrm{trace}(T_\nu)$, so that from Lemma 3.1-(iii) and by analogy with spectral truncation, a solution to the regularised problem (3.3) then satisfies $\mathbf{f}^T\boldsymbol{\upsilon}^*_\alpha \leqslant \mathbf{f}^T\boldsymbol{\omega} = \mathrm{trace}(T_\mu)$.

From the reproducing property in $\mathcal{G}$ and the Cauchy-Schwarz inequality, we have, for all $x \in \mathscr{X}$,

$$\forall \mu \text{ and } \nu \in \mathscr{T}(K),\ \left|g_\mu(x) - g_\nu(x)\right| \leqslant \sqrt{D_{K^2}(\mu,\nu)}K(x,x), \tag{8.2}$$

so that for $D_{K^2}(\mu,\nu)$ fixed, $K(x_k,x_k)$ appears as a tight bound on $g_\mu(x_k) - g_\nu(x_k)$. In view of (8.1) and (8.2), by considering $\mathbf{f} = 1/(\mathrm{diag}(\mathbf{K}))^p$ 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 relatively small for every $k \in \{1,\cdots,N\}$.

The impact of the penalisation direction on the tradeoff between sparsity and squared-kernel discrepancy is investigated and further discussed in the experiments of Sections 9.5 and 9.6.

9. 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 [14]), as illustrated in Figure 9.2. We set $\omega_k = 1/N$ for all k, so that the measure $\mu = \sum_k \omega_k \delta_{x_k}$ in particular 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 9.6). An overview of the spectrum of the operator T_μ is given in Figure 9.1. We first consider the penalisation direction $\mathbf{f} = \mathbb{1}$.

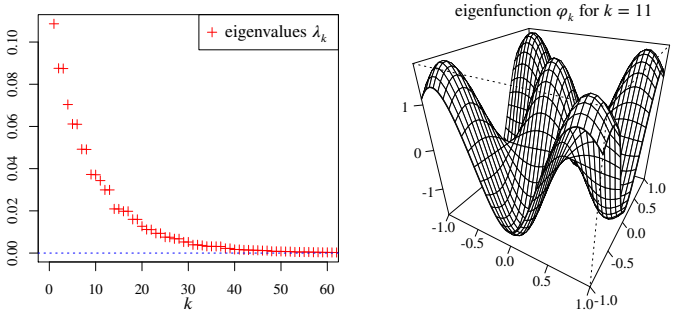

FIG. 9.1. *For the two-dimensional example (Gaussian kernel and $\boldsymbol{\omega} = \mathbb{1}/N$), eigenvalues λ_k of the integral operator T_μ (sorted in decreasing order, only the 62 largest eigenvalues are presented), and graph, on $[-1,1]^2$ of the canonically extended eigenfunction φ_k for $k = 11$.*

9.1. First experiment. Figure 9.2 shows the (approximate) solution $\boldsymbol{\upsilon}^*$ to problem (3.4) with $\varkappa = 0.81$, or equivalently, to problem (3.3) with $\alpha \approx 8.354214 \times 10^{-3}$ (with $\boldsymbol{\omega} = 1/N$ and $\mathbf{f} = \mathbb{1}$). The vector $\boldsymbol{\upsilon}^*$ has 160 strictly positive components, and the support of the related measure ν^* inherits an interesting "four-concentric-square" structure. We have $D(\boldsymbol{\upsilon}^*) = 7.631887 \times 10^{-4}$ (for comparison, notice that $D(0) = \frac{1}{2}\boldsymbol{\omega}^T\mathbf{S}\boldsymbol{\omega} = 2.661452 \times 10^{-2}$ and $D(\mathbf{e}_1) = 4.760566 \times 10^{-1}$, with $\mathbf{e}_1$ the first element of the canonical basis of $\mathbb{R}^N$).

The presented solution has been obtained from the regularisation-path strategy described in Section 5 (see Section 9.2 for more details). Considering the regularisation path for problem (3.3) with decreasing values of α, the underlying value of $\alpha \approx 8.354215 \times 10^{-3}$ satisfies

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

Correspondingly, considering the regularisation path for problem (3.4) with increasing values of $\varkappa$, the underlying value $\varkappa = 0.81$ satisfies

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

In the framework of Section 6.1, the presented solution is related to a Frank-Wolfe error bound $\epsilon = 3.989864 \times 10^{-17}$.

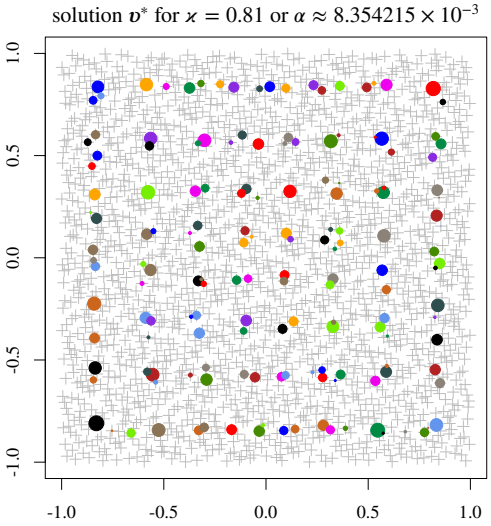

FIG. 9.2. *Graphical representation (two-dimensional example, Gaussian kernel, $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{f} = \mathbb{1}$) of the solution υ^* to problem* (3.4) *with $\varkappa = 0.81$, or equivalently, to problem* (3.3) *with $\alpha \approx 8.354215 \times 10^{-3}$. The grey crosses represent the points in S and the filled dots are the strictly positive components of υ^* (surface being proportional to υ_k^*).*

The accuracy of the approximate eigendecomposition of T_μ induced by the solution υ^* presented in Figure 9.2 (i.e., $\varkappa = 0.81$) is illustrated in Figure 9.3. In view of the similarity between the approximate eigenvalues $\widehat{\lambda}_l^{[\cdot]}$, and more particularly of the ratios $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ and $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ (see Section 7), we observe that the 21 main eigendirections of the operator T_{υ^*} (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 $k \geqslant 44$.

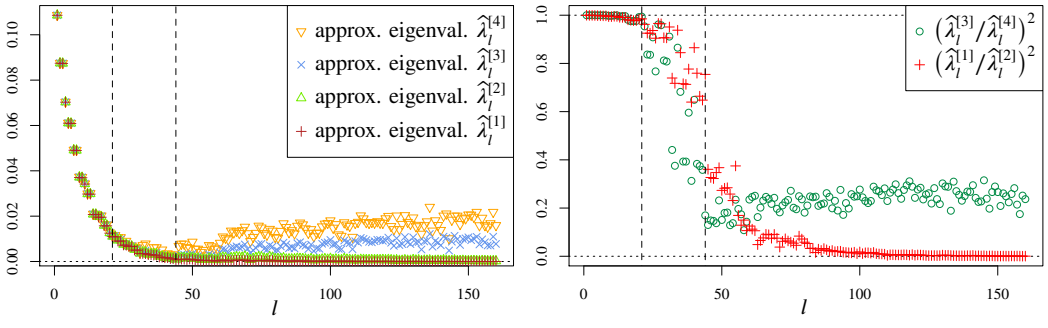

FIG. 9.3. *Approximate eigenvalues $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ induced by the solution υ^* presented in Figure 9.2; ratios $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ and $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ highlighting the accuracy of the approximate eigendirections ψ_l of T_μ induced by υ^* (right).*

A comparison between the true eigenvalues of T_μ and their approximations induced by the solution υ^* presented in Figure 9.2 (i.e., $\varkappa = 0.81$) is illustrated in Figure 9.3. We for instance observe that for $1 \leqslant l \leqslant 8$, the approximate eigenvalues $\widehat{\lambda}_l^{[4]}$ are the more accurate.

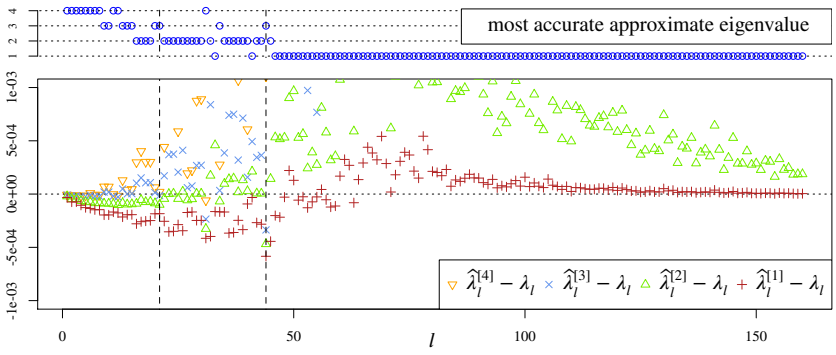

FIG. 9.4. *Errors $\widehat{\lambda}_l^{[\cdot]} - \lambda_l$ for the approximate eigenvalues induced by the solution $\boldsymbol{\upsilon}^*$ presented in Figure 9.2 (bottom), and indication of the most accurate (i.e., with the smallest error in absolute value) approximate eigenvalues among $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ (top).*

9.2. Regularisation path. We now investigate the impact of the parameters α and $\varkappa$ related to problems (3.3) and (3.4) respectively. We compute the 12 786 first events of the regularisation-path related to problem (3.3) with decreasing values of α (see Section 5), i.e., until we reach a precision issue; in particular, we have $\alpha_0 = 6.310163 \times 10^{-2}$ and $\alpha_{12785} = 1.514626 \times 10^{-5}$. Correspondingly, for the regularisation path related to problem (3.4) with increasing $\varkappa$, we have $\varkappa_0 = 0$ and $\varkappa_{12785} = 0.9995426$ (we recall that $\mathbf{f}^T\boldsymbol{\omega} = 1$).

Figure 9.5 shows that the number of strictly positive components of the solution $\boldsymbol{\upsilon}_\varkappa^*$ to problem (3.4) tends to increase when $\varkappa$ increases. As expected from Lemma 3.1-(vii), the functions $\varkappa \mapsto D(\boldsymbol{\upsilon}_\varkappa^*)$ is decreasing; in the same way, when $\varkappa$ increases, the corresponding value of the regularisation parameter α decreases (see Lemmas 3.1 and 3.2).

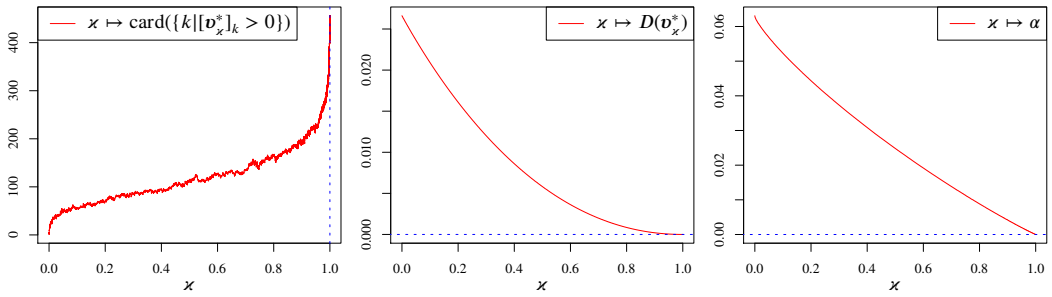

FIG. 9.5. *For the two-dimensional example (Gaussian kernel, $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{f} = \mathbb{1}$), graphical representation of the* 12 786 *first events of the regularisation path related to problem* (3.4) *for increasing $\varkappa$; number of strictly positive components of $\boldsymbol{\upsilon}_\varkappa^*$ as function of $\varkappa$ (left); graph of $\varkappa \mapsto D(\boldsymbol{\upsilon}_\varkappa^*)$ (middle), and relation between $\varkappa$ and the parameter α of problem* (3.3).

For 51 values of $\varkappa$ evenly spread between 0 and $\varkappa_{12785}$, Figure 9.6 shows the evolution of the ratio $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ for the approximate eigendecompositions induced by the 51 different solutions $\boldsymbol{\upsilon}_\varkappa^*$. As expected, the number of accurately approximate eigendirections increases with $\varkappa$. Remarkably, for each of the considered values of $\varkappa$, the number of eigendirections approximated with a high accuracy appears to be in close relation with the decay of the spectrum of T_μ; we recall that we have $\mathrm{trace}(T_{\nu_\varkappa^*}) = \varkappa$, since $\mathrm{diag}(\mathbf{K}) = \mathbb{1}$ for the Gaussian kernel.

9.3. Components merging. We now perform the strong-pairwise-merging (see Section 6.2) of the solution $\boldsymbol{\upsilon}^*$ presented in Figure 9.2 (i.e., problem (3.4) with $\varkappa = 0.81$). As illustrated in Figure 9.7, for the first merging iterations, $D(\boldsymbol{\upsilon}_{[k]})$ stays very close to $D(\boldsymbol{\upsilon}^*) = 7.631890 \times 10^{-4}$. After 90 iterations, we have $D(\boldsymbol{\upsilon}_{[90]}) - D(\boldsymbol{\upsilon}^*) = 3.494809 \times 10^{-5}$ (i.e., increasing 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

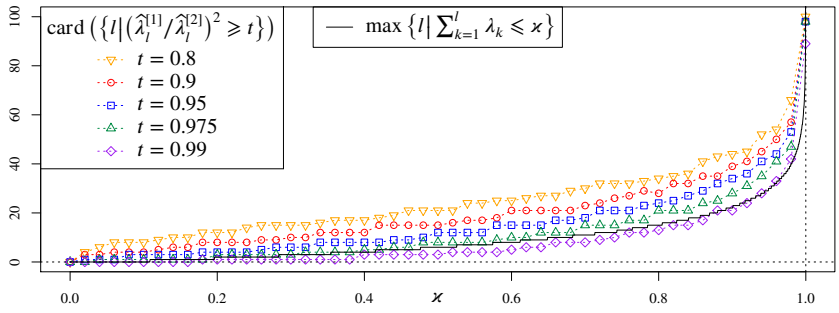

FIG. 9.6. *Evolution of the accuracy of the approximate eigendecomposition of T_μ induced by $\boldsymbol{\upsilon}_\varkappa^*$ for 51 values of $\varkappa$ between $\varkappa_0 = 0$ and $\varkappa_{12785} = 0.9995426$; the accuracy of the approximate eigendirections is measured trough the ratios $\left(\hat{\lambda}_l^{[1]}/\hat{\lambda}_l^{[2]}\right)^2$, 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 $\mathbf{f} = \mathbb{1}$).*

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 9.7. 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.

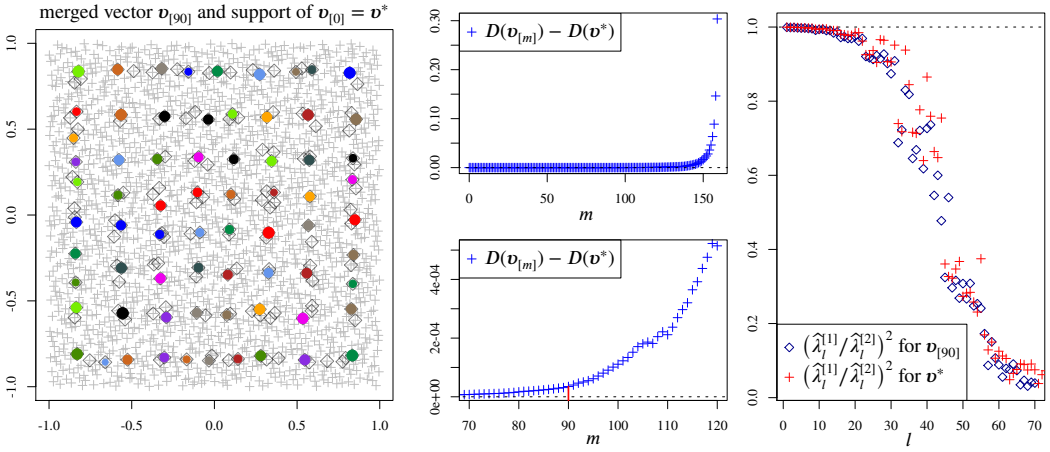

FIG. 9.7. *Graphical representation of the merged solution $\boldsymbol{\upsilon}_{[90]}$ (two-dimensional example with $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{f} = \mathbb{1}$) obtained after 90 iterations of the strong-pairwise-merging strategy applied to the solution $\boldsymbol{\upsilon}^*$ presented in Figure 9.2; 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 $\left(\hat{\lambda}_l^{[1]}/\hat{\lambda}_l^{[2]}\right)^2$ obtained from the merged vector $\boldsymbol{\upsilon}_{[90]}$ and comparison with the same ratios for the solution $\boldsymbol{\upsilon}^*$ (right).*

9.4. Comparison with random sampling. For comparison purpose, 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 $\mathcal{S}$, and we consider the uniform probability measure supported by the points selected); for each sample size, we perform 100 repetitions. Figure 9.8 illustrates the accuracy of the obtained approximate eigendirections, measured through the ratios $\left(\hat{\lambda}_l^{[1]}/\hat{\lambda}_l^{[2]}\right)^2$. As we could expect, the accuracy of the approximations increases with the size of the sample; however, in terms of trade-off between sparsity and accuracy, the results are far behind the one obtained using the penalised squared-kernel-discrepancy minimisation.

9.5. Impact of the penalisation direction. We now study the impact of the penalisation direction $\mathbf{f}$. For the two-dimensional example (Gaussian kernel and $\boldsymbol{\omega} = \mathbb{1}/N$), we compute the

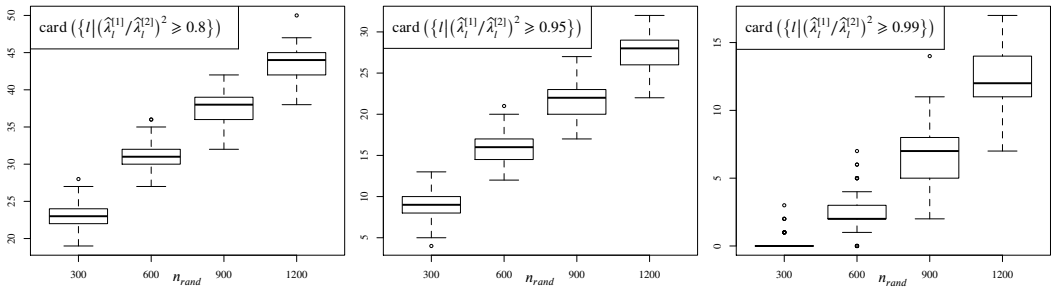

FIG. 9.8. *For the two-dimensional example, accuracy of the approximate eigendecompositions induced by random samples of size n_{rand} (without replacement); for each values of n_{rand}, Tukey's boxplot, over* 100 *repetitions, of the number of approximate eigendirections such that $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2 \geqslant 0.8$ (left),* 0.95 *(middle) and* 0.99 *(right).*

regularisation path (until precision issues) of the regularised problem (3.3) for seven different vectors $\mathbf{f} > 0$. We consider $\mathbf{f} = \mathbf{v}_{\max}(\mathbf{S})$ (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., $f_k = [\mathbf{S}\boldsymbol{\omega}]_k^2$), $\sqrt{\mathbf{S}\boldsymbol{\omega}}$ (i.e., $f_k = \sqrt{[\mathbf{S}\boldsymbol{\omega}]_k}$), $\mathbb{1}$, $1/\sqrt{\mathbf{S}\boldsymbol{\omega}}$, $1/(\mathbf{S}\boldsymbol{\omega})$ and $1/(\mathbf{S}\boldsymbol{\omega})^2$. In Figure 9.9, we represent, for each case, the number of strictly positive components of the solution to (3.3) as function of the squared-kernel discrepancy of the solution. We recall that for the Gaussian kernel, we have $\mathbb{1} = \operatorname{diag}(\mathbf{K}) = 1/\operatorname{diag}(\mathbf{K})$.

In accordance with Section 8, the results obtained for $\mathbf{f} = \mathbb{1}$ and $\mathbf{f} = 1/(\mathbf{S}\boldsymbol{\omega})^p$ (with in this case $p = 1/2$, 1 and 2) appears as the more interesting in terms of the overall tradeoff between sparsity and squared-kernel discrepancy (i.e, for a same value of the squared-kernel discrepancy, the underlying solutions are generally sparser); the best results are obtained for $\mathbf{f} = 1/(\mathbf{S}\boldsymbol{\omega})^2$. We can however remark that $\mathbf{f} = \mathbf{v}_{\max}(\mathbf{S})$ and $\mathbf{f} = (\mathbf{S}\boldsymbol{\omega})^2$ lead to sparser solutions in the range of large values of the squared-kernel discrepancy (i.e., on the right-hand side of Figure 9.9).

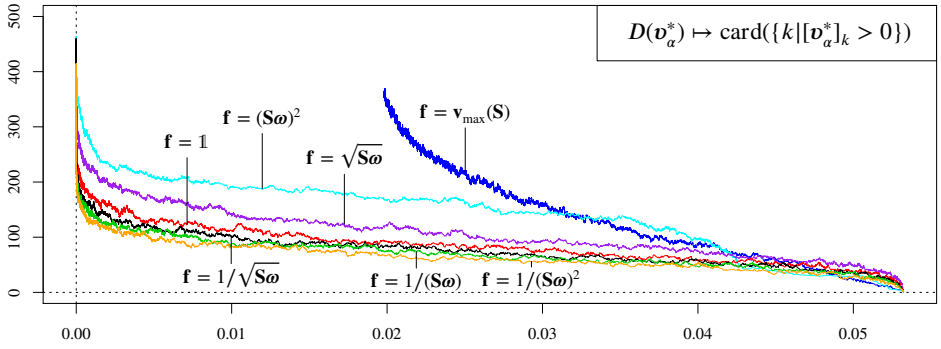

FIG. 9.9. *For the two-dimensional example (Gaussian kernel and $\boldsymbol{\omega} = \mathbb{1}/N$), number of strictly positive components of the solution $\boldsymbol{\upsilon}_\alpha^*$ to problem* (3.3) *as function of the squared-kernel discrepancy $D(\boldsymbol{\upsilon}_\alpha^*)$, for various penalisation vectors* $\mathbf{f}$*; all the curves have been obtained thanks to the regularisation path strategy (until precision issues).*

9.6. Modified kernel. We further illustrate the impact of the penalisation direction by now considering an alternative kernel (with same set $\mathcal{S}$ and same initial measure μ, i.e., $\boldsymbol{\omega} = \mathbb{1}/N$, as in the previous experiments). 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)

$$K(x,y) = s(x)s(y)\exp(-\ell\|x-y\|^2); \tag{9.1}$$

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 (9.5), while considering $\mathbf{f} = \mathbb{1}$, $\operatorname{diag}(\mathbf{K})$, $1/\operatorname{diag}(\mathbf{K})$, $1/(\mathbf{S}\boldsymbol{\omega})$, $1/(\mathbf{S}\boldsymbol{\omega})^2$ and $(\mathbf{S}\boldsymbol{\omega})^2$. The results are presented in Figure 9.10. The overall tradeoff between sparsity and squared-kernel discrepancy obtained for $\mathbf{f} = (\mathbf{S}\boldsymbol{\omega})^2$ is very poor in comparison with the tradeoffs obtained

for the five other penalisation directions, in accordance with the discusion made Section 8. The best overall tradeoffs are obtained for $\mathbf{f} = 1/(\mathbf{S}\boldsymbol{\omega})$ and $1/(\mathbf{S}\boldsymbol{\omega})^2$. Among the five "efficient penalisation directions", the trace $\mathbf{f} = \mathrm{diag}(\mathbf{K})$ appears to be the less interesting in the range of large values of the squared-kernel discrepancy (right-hand side of the graph), but becomes the more efficient in the range of small values of the squared-kernel discrepancy (left-hand side of the graph).

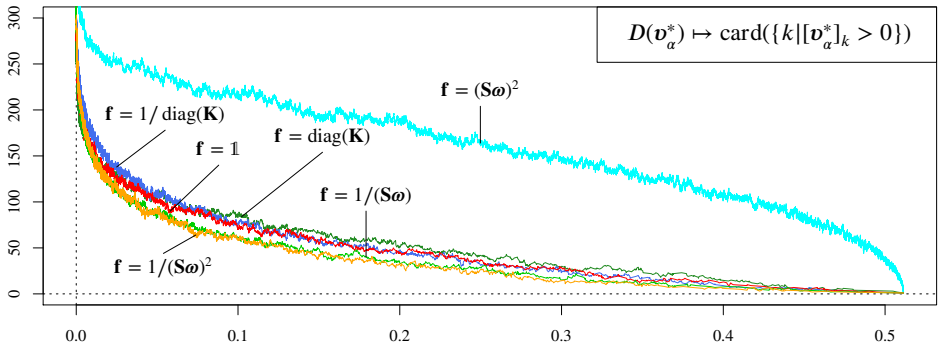

FIG. 9.10. *For the two-dimensional example (modified Gaussian kernel* (9.1) *and* $\boldsymbol{\omega} = \mathbb{1}/N$*), number of strictly positive components of the solution* $\boldsymbol{\upsilon}^*_\alpha$ *to problem* (3.3) *as function of the squared-kernel discrepancy* $D(\boldsymbol{\upsilon}^*_\alpha)$*, for various penalisation vectors* $\mathbf{f}$*; all the curves have been obtained thanks to the regularisation path strategy (until precision issues).*

10. 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 [13]. 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.

10.1. MiniBooNE dataset. We consider the standardised entries of the MiniBooNE dataset, without labels; $\mathcal{S}$ thus consists of $N = 129\,596$ points in $\mathbb{R}^{50}$. We use a Gaussian kernel (same expression as in Section 9) with $\ell = 0.02$, and we set $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{f} = \mathbb{1}$. Notice that $\ell = 0.02$ belongs to the range of "good parameters" for the SVM binary classification of this dataset.

We compute the 3 000 first events of the regularisation path related to problems (3.3) and (3.4). We have $\alpha_0 = 0.2188961$ and $\alpha_{2999} = 3.546703 \times 10^{-3}$, and correspondingly $\varkappa_0 = 0$ and $\varkappa_{2999} = 0.655808$ (note that $\mathbf{f}^T\boldsymbol{\omega} = 1$). A graphical representation of the obtained results is proposed in Figure 10.1. We can observe that for $\varkappa \geqslant 0.5$, the number of strictly positive components of $\boldsymbol{\upsilon}^*_\varkappa$ increases with a signifiant rate; as a consequence, the computation of the regularisation path quickly becomes numerically intractable (notice that the calculation of the 3 000 first events of the regularisation path took around 3 hours on our aforementioned 2015 desktop).

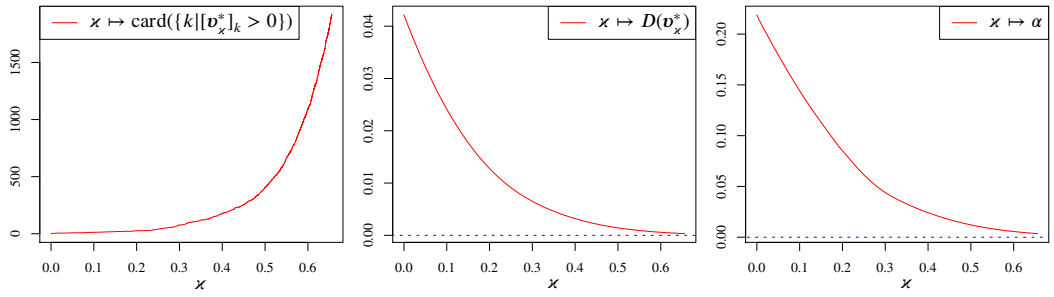

FIG. 10.1. *For the MiniBooNE dataset (Gaussian kernel,* $\boldsymbol{\omega} = \mathbb{1}/N$ *and* $\mathbf{f} = \mathbb{1}$*), graphical representation of the* 3 000 *first events of the regularisation path related to problem* (3.4) *for increasing* $\varkappa$*: number of strictly positive components of* $\boldsymbol{\upsilon}^*_\varkappa$ *as function of* $\varkappa$ *(left); graph of* $\varkappa \mapsto D(\boldsymbol{\upsilon}^*_\varkappa)$ *(middle), and relation between* $\varkappa$ *and the parameter* α *of problem* (3.3)*.*

We use the regulation path to compute the solutions to problem (3.4) for $\varkappa = 0.3$ and $\varkappa = 0.655$

(i.e., for problem (3.3), $\alpha \approx 4.400276\times10^{-2}$ and $\alpha \approx 3.571413\times10^{-3}$). The considered solutions have 76 and 1 902 strictly positive components, respectively. The efficiency of the induced approximate eigendecompositions is illustrated in Figure 10.2. 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 $\big(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]}\big)^2$, the ratios $\big(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\big)^2$ remain relatively high for all the values of l presented in the graph.

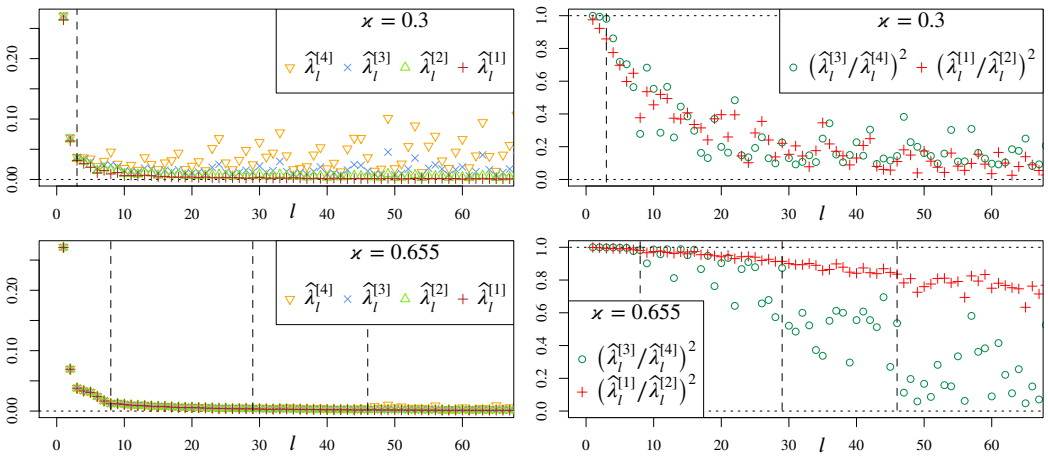

FIG. 10.2. *For the MiniBooNE dataset (Gaussian kernel, $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{f} = \mathbb{1}$), approximate eigenvalues $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ induced by the solution to problem* (3.3) *with $\varkappa = 0.3$ (top-left), and ratios $\big(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]}\big)^2$ 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).*

To explore the type of solutions obtained for larger values of $\varkappa$, we consider the vertex-exchange strategy described in Section 6.1. We compute an approximate solution for $\varkappa = 0.8$; the vertex-exchange algorithm is initialised at $\widetilde{\boldsymbol{\upsilon}} = \mathbf{e}_1$ and after 300 000 iterations, we obtain a Frank-Wolfe error bound of $\epsilon = 1.692408 \times 10^{-8}$. The obtained approximate solution $\widehat{\boldsymbol{\upsilon}}^*$ to problem (3.3) verifies $D(\widehat{\boldsymbol{\upsilon}}^*) = 4.934072 \times 10^{-5}$ and has 9544 strictly positive components.

To enhance sparsity, we perform a weak-pairwise merging of the solution $\widehat{\boldsymbol{\upsilon}}^*$ for $\varkappa = 0.8$ (see Section 6.2; notice that performing a strong-pairwise merging is in this case numerically prohibitive). 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., increasing of 2.15%).

We next compute the approximate eigendecompositions induced by $\widehat{\boldsymbol{\upsilon}}^*$ and $\boldsymbol{\upsilon}_{[5044]}$; the result are presented in Figure 10.3. In particular, in both case, the 31 main eigendirections of T_μ are approximated with high accuracy. We observe also 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.

10.2. 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 9) with $\ell = 0.4$, and we set $\boldsymbol{\omega} = \mathbb{1}/N$ and $\mathbf{f} = \mathbb{1}$. 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 (3.4) with $\varkappa = 0.3$; we perform four consecutive batches of 50 000 iterations each, the solver being initialised at $\widetilde{\boldsymbol{\upsilon}} = \mathbf{e}_1$. After 200 000 iterations (i.e., at the end of the 4-th batch), the obtained approximate solution $\widehat{\boldsymbol{\upsilon}}^*$ verifies $D(\widehat{\boldsymbol{\upsilon}}^*) = 3.931629 \times 10^{-5}$ 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 10.1. We observe that a batch of 50 000 iterations of the vertex-exchange algorithm

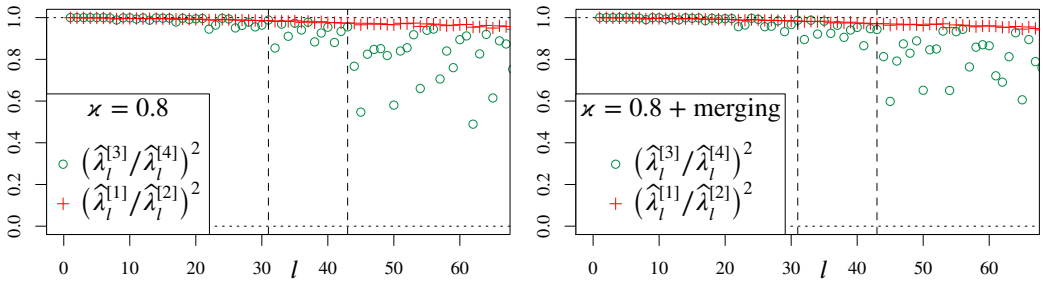

FIG. 10.3. *For the MiniBooNE dataset, accuracy of the approximate eigendecompositions induced by the solution* $\widehat{\upsilon}^*$ *to problem* (3.3) *with* $\varkappa = 0.8$ *obtained from the vertex-exchange algorithm (left), and from the merged solution* $\upsilon_{[5044]}$ *(left).*

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

TABLE 10.1

For the test subsample of the SUSY dataset, information relative to the approximate solutions to problem (3.4) *with* $\varkappa = 0.3$ *returned by the vertex-exchange algorithm for four consecutive batches of* 50 000 *iterations, the solver being initialised at* $\widetilde{\upsilon} = \mathbf{e}_1$*; for each batch, execution time, total number of iterations, Frank-Wolfe error bound* ϵ *and number* n *of strictly positive component of the approximate solution.*

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

To enhance the sparsity of the sampling, we perform a weak-pairwise merging of the approximate solution $\widehat{\upsilon}^*$; the computation of 20 673 merging iterations took 78.86 seconds. The merged solution $\upsilon_{[13674]}$ is supported by 7 000 points and $D(\upsilon_{[13674]}) = D(\widehat{\upsilon}^*) + 5.271960 \times 10^{-7}$ (i.e., increasing of only 1.34%). We next study the approximate eigendecomposition induced by $\upsilon_{[13674]}$. Computing the 300 first normalised approximate eigenvectors $\widehat{\mathbf{v}}_l$ of $\mathbf{KW}$ induced by $\upsilon_{[13674]}$ (i.e., $\widehat{\mathbf{v}}_l \in \mathbb{R}^N$ is the vector corresponding to the approximate eigendirection $\widehat{\varphi}_l$, see Sections 2.2 and 7) took 3 313.6 seconds (time for canonical extension and rescaling), and we thus also obtain the approximate eigenvalues $\widehat{\lambda}_l^{[1]}$. To access the accuracy of the approximate directions $\widehat{\varphi}_l$, we compute $T_\mu[\widehat{\varphi}_l]$ (i.e., $\mathbf{KW}\widehat{\mathbf{v}}_l$), for the 300 first approximate eigendirections, i.e., $l \in \{1, \cdots, 300\}$; this operation took 191 622.3 seconds (i.e., around 53 hours). The results are presented in Figure 10.4. 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). The obtained approximate eigenpairs are remarkably accurate.

11. Conclusion. We studied a QP-based strategy to design sparse (pointwise) quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels in a quadrature-sparsification framework (i.e., quadratures with support included in a fixed finite set of N points). For a given kernel, the considered criterion consists in the squared Hilbert-Schmidt norm for operators defined on the underlying RKHS, and sparsity of the approximate quadratures is obtained through the introduction of an ℓ^1-type penalisation, under the form of a regularisation term or of a constraint. We investigated the relations between the approximation of integral operators and the notions of squared-kernel discrepancy and one-class distorted SVMs related to the squared kernel. From a methodological point of view, the considered approximation scheme is deterministic, numerically scalable (i.e., large-scale problem can be tackled) and enjoys an optimality property; it in particular

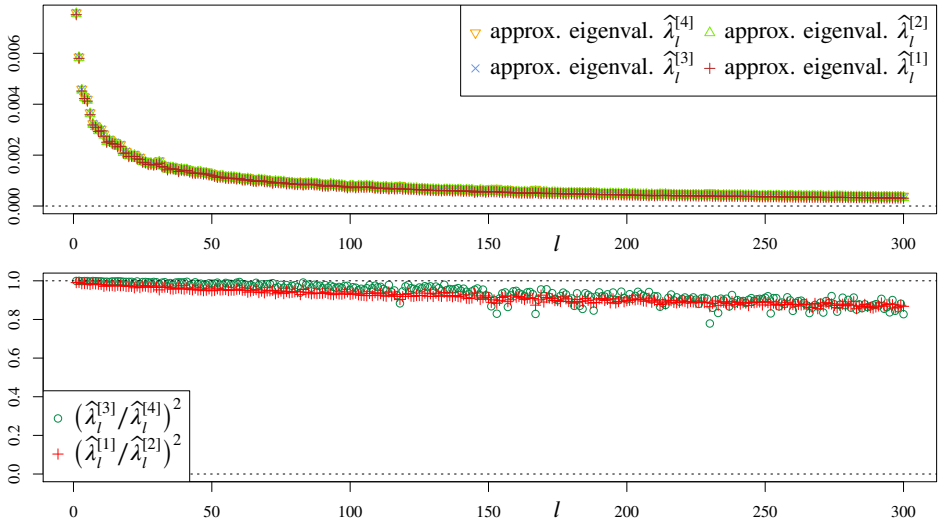

FIG. 10.4. *For the test subsample of the SUSY dataset, graphical representation of the* 300 *first approximate eigenvalues* $\widehat{\lambda}_l^{[1]}$, $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ *and* $\widehat{\lambda}_l^{[4]}$ *induced by the merged solution* $\mathfrak{v}_{[13674]}$ *obtained from the approximate solution* $\widehat{\mathfrak{v}}^*$ *to problem* (3.4) *(top); ratios* $\left(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]}\right)^2$ *and* $\left(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]}\right)^2$ *highlighting the accuracy of the underlying approximate eigendirections (bottom).*

offers an interesting alternative for the sampling problem related to the Nyström approximation of large-scale kernel-matrices.

We described two numerical strategies for solving regularised or constrained squared-kernel-discrepancy minimisation problems. The regularisation-path approach can be used to explore the range of very sparse solutions, with the interest of leading to exact solutions (up to precision errors); the vertex-exchange 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 are also proposed, aiming at enhancing sparsity while trying to keep squared-kernel discrepancy as low as possible. To assess the accuracy of a given approximate eigendirection, we in particular considered four different approximations (based on geometrical considerations) of the eigenvalue related to a given approximate eigendirection. We also discussed questions relative to selection of relevant penalisation directions, and we more specifically highlighted the benefits of considering penalisation terms promoting collinearity in the RKHS related to the squared kernel.

As illustrated in Section 10 (where the support of the measures μ and ν is included in a fixed set of N points), the proposed approach can be used to tackle relatively large-scale problems (i.e., large N). As far as the kernel $K(\cdot,\cdot)$ can be easily evaluated, the considered methodology leads to accurate sparse representations of the main eigenpairs of the initial operator T_μ, in a reasonable amount of time and without necessarily resorting to powerful computing hardwares. Indeed, in the range of sparse solutions, the most computationally demanding task required to obtain an approximate measure ν is the calculation of the distortion term $\mathbf{S}\boldsymbol{\omega}$, with numerical complexity scaling as $\mathcal{O}(N^2)$. Then, following Section 7, we can easily compute the eigendecomposition of the approximate operator T_ν, and obtain the approximate eigenpairs $\{(\widehat{\lambda}_l^{[1]}, \widehat{\varphi}_l)\}_{l\in\mathbb{I}_\nu^+}$. Assessing the accuracy of an approximate eigendecomposition through the evaluation of the approximate eigenvalues $\widehat{\lambda}_l^{[2]}$, $\widehat{\lambda}_l^{[3]}$ and $\widehat{\lambda}_l^{[4]}$ is more challenging (and optional) since it requires the computation of $T_\mu[\widehat{\varphi}_l]$, with complexity scaling as $\mathcal{O}(N^2)$. However, as already mentioned, any of the operations with $\mathcal{O}(N^2)$ complexity considered in this work consists of kernel-matrix vector products that can be easily parallelised.

Acknowledgements. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information – Research Council KUL:

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

Appendix A. Kernel discrepancy and integration in RKHS.

Consider the framework of Section 2 and introduce the subspace $\mathscr{I}(K)$ 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 the kernel $K(\cdot,\cdot)$ and the Cauchy-Schwarz inequality, we have, for all $h \in \mathcal{H}$ and $\mu \in \mathscr{I}(K)$,

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

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 given by

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

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

Let us consider that we are searching for a measure ν in order to approximate integrals relative to a fixed measure μ (such a situation occurs for instance in case of quadrature approximation).If both measures μ and ν belong to $\mathscr{I}(K)$, 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, and it is therefore of interest to deal with a measure ν such that $D_K(\nu,\mu)$ is small; see [3] for a further discussion.

Appendix B. Soft-margin-type extensions for the one-class SVM related to the regularised problem. Pursuing the analogy with one-class SVMs, we can define soft-margin-type extensions of problems (4.1) and (4.3), i.e, we can consider models where the inequalities appearing in the constraints can, potentially, be violated, the level of violation being penalised. In this section, we only discuss extensions related to problem (4.1), but a similar discussion holds for problem (4.3).

We introduce $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_N)^T \in \mathbb{R}^N$; the components of $\boldsymbol{\xi}$ are referred to as *slack variables*. Instead of considering the contraints $g(x_k) \geqslant -\alpha f_k$, 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 [22]. In what follows, we discuss the model obtained for two popular types of loss functions: the (weighted) hinge loss and the (generalised) square loss.

In view of our initial problem (3.3) (regularised squared-kernel-discrepancy minimisation), soft-margin extensions of problem (4.1) appear as tool to further constrain or penalise the measure ν (i.e., the vector $\boldsymbol{\upsilon}$) used to approximate the initial measure μ (i.e., the vector $\boldsymbol{\omega}$).

B.1. Weighted hinge loss. Let $\mathbf{c} \in \mathbb{R}^N$, with $\mathbf{c} \geqslant 0$; the soft-margin extension of problem (4.1) corresponding to a weighted hinge loss consists in the problem, for $g \in \mathcal{G}$ and $\boldsymbol{\xi} \in \mathbb{R}^N$,

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

The Lagrange dual of problem (B.1) is given by

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

Problem (B.2) only differs from problem (3.3) by the presence of the additional constraints $\boldsymbol{\upsilon} \leqslant \mathbf{c}$, which acts as an upper bound on the values taken by the components of $\boldsymbol{\upsilon}$ (i.e., an upper bound on the weights of the points in the quadrature related to ν).

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 (4.1) corresponding to a generalised square loss is

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

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

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

In comparison to problem (3.3), 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 underlying solution $\boldsymbol{\upsilon}^*$. In particular, $D_{\alpha,\boldsymbol{\Sigma}}$ is then strongly convex.

Appendix C. Proofs.

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

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

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

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

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 linear continuity of T_μ, we then obtain

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

and we conclude by using the Tonelli theorem. □

Proof of Lemma 2.2. The proof directly follows from the properties discussed in Sections 2.1 and 2.3. In particular, (2.2) 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 Lemma 3.1. Assertion (i) follows directly from $D_{K^2}(\mu,\mu) = 0$ (since $D_{K^2}(\mu,\nu) \geqslant 0$). From the first order optimality condition, for $\alpha \geqslant 0$, a feasible $\boldsymbol{\upsilon}_\alpha^*$ is solution to (3.3) if and only if, for any feasible $\boldsymbol{\upsilon}$, we have $(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_\alpha^*)^T\nabla D_\alpha(\boldsymbol{\upsilon}_\alpha^*) \geqslant 0$. Then, considering $\boldsymbol{\upsilon}_\alpha^* = 0$ leads to assertion (ii), and since $\boldsymbol{\omega}$ is by assumption feasible for (3.3), assertion (iii) is obtained by taking $\boldsymbol{\upsilon} = \boldsymbol{\omega}$. For assertion (iv), we first remark that the first order optimality condition for $\boldsymbol{\upsilon} = 0$ gives $(\boldsymbol{\upsilon}_\alpha^*)^T\nabla D_\alpha(\boldsymbol{\upsilon}_\alpha^*) \leqslant 0$. If we assume that there exists k such that $[\nabla D_\alpha(\boldsymbol{\upsilon}_\alpha^*)]_k < 0$, we obtain $(\mathbf{e}_k - \boldsymbol{\upsilon}_\alpha^*)^T\nabla D_\alpha(\boldsymbol{\upsilon}_\alpha^*) < 0$, and the first order optimality condition is then violated ($\mathbf{e}_k$ stands for the k-th element of the canonical

basis of $\mathbb{R}^N$, so that $\mathbf{e}_k^T \nabla D_\alpha(\boldsymbol{v}_\alpha^*) = [\nabla D_\alpha(\boldsymbol{v}_\alpha^*)]_k$). Therefore, we necessarily have $\nabla D_\alpha(\boldsymbol{v}_\alpha^*) \geqslant 0$ and $(\boldsymbol{v}_\alpha^*)^T \nabla D_\alpha(\boldsymbol{v}_\alpha^*) = 0$ (since $\boldsymbol{v}_\alpha^* \geqslant 0$). To prove (v), we first remark that for all $\theta \in [0,1]$, we have

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

Since $D_\alpha(\widetilde{\boldsymbol{v}}_\alpha^*) = D_\alpha(\theta \boldsymbol{v}_\alpha^* + (1-\theta)\widetilde{\boldsymbol{v}}_\alpha^*)$ and $(\boldsymbol{v}_\alpha^* - \widetilde{\boldsymbol{v}}_\alpha^*)^T \nabla D_\alpha(\widetilde{\boldsymbol{v}}_\alpha^*) = 0$, we necessarily have $(\boldsymbol{v}_\alpha^* - \widetilde{\boldsymbol{v}}_\alpha^*)^T \mathbf{S}(\boldsymbol{v}_\alpha^* - \widetilde{\boldsymbol{v}}_\alpha^*) = 0$ and therefore $\mathbf{S}(\boldsymbol{v}_\alpha^* - \widetilde{\boldsymbol{v}}_\alpha^*) = 0$ (since the matrix $\mathbf{S}$ is symmetric and positive-semidefinite); this completes the proof. Assertion (vi) follows from the expansion, for all $\boldsymbol{v} \geqslant 0$,

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

Since all the entries of $\mathbf{S}$ are non-negative and $\mathbf{S}_{k,k} > 0$, if $[\alpha \mathbf{f} - \mathbf{S}\boldsymbol{\omega}]_k \geqslant 0$ and $v_k > 0$, we necessarily have $D(\boldsymbol{v} - v_k \mathbf{e}_k) < D_\alpha(\boldsymbol{v})$. To obtain assertion (vii), consider $\alpha_1 < \alpha_2$ and let $\boldsymbol{v}_{\alpha_1}^*$ and $\boldsymbol{v}_{\alpha_2}^*$ be solutions to problem (3.3) with $\alpha = \alpha_1$ and $\alpha = \alpha_2$ respectively. We have $D(\boldsymbol{v}_{\alpha_1}^*) - D(\boldsymbol{v}_{\alpha_2}^*) \leqslant \alpha_1 \mathbf{f}^T(\boldsymbol{v}_{\alpha_2}^* - \boldsymbol{v}_{\alpha_1}^*)$ and $D(\boldsymbol{v}_{\alpha_1}^*) - D(\boldsymbol{v}_{\alpha_2}^*) \geqslant \alpha_2 \mathbf{f}^T(\boldsymbol{v}_{\alpha_2}^* - \boldsymbol{v}_{\alpha_1}^*)$, so that, necessarily, $\mathbf{f}^T(\boldsymbol{v}_{\alpha_2}^* - \boldsymbol{v}_{\alpha_1}^*) \leqslant 0$, and therefore $D(\boldsymbol{v}_{\alpha_1}^*) - D(\boldsymbol{v}_{\alpha_2}^*) \leqslant 0$. □

Proof of Lemma 3.2. If $\boldsymbol{v}_\alpha^*$ is a solution to problem (3.3) with $\alpha \geqslant 0$, then by definition, $\boldsymbol{v}_\alpha^*$ in particular minimises $D(\cdot)$ over the set of all $\boldsymbol{v} \geqslant 0$ such that $\mathbf{f}^T \boldsymbol{v} = \mathbf{f}^T \boldsymbol{v}_\alpha^*$, and $\boldsymbol{v}_\alpha^*$ is therefore a solution to problem (3.4) with $\varkappa = \mathbf{f}^T \boldsymbol{v}_\alpha^*$.

The condition $\varkappa \leqslant \mathbf{f}^T \boldsymbol{\omega}$ follows directly from Lemma 3.1-(iii): a solution $\boldsymbol{v}_\alpha^*$ to problem (3.3) indeed necessarily satisfies $\mathbf{f}^T \boldsymbol{v}_\alpha^* \leqslant \mathbf{f}^T \boldsymbol{\omega}$. For $\varkappa = 0$, we have $\boldsymbol{v}_\varkappa^* = 0$ and the result follows from Lemma 3.1-(ii). For $0 < \varkappa \leqslant \mathbf{f}^T \boldsymbol{\omega}$, in order to be a solution to problem (3.3), a solution $\boldsymbol{v}_\varkappa^*$ to problem (3.4) must satisfy (first order optimality condition), $(\boldsymbol{v} - \boldsymbol{v}_\varkappa^*)^T \nabla D_\alpha(\boldsymbol{v}_\varkappa^*) \geqslant 0$ for all $\boldsymbol{v} \geqslant 0$. In particular, for all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{v}_\varkappa^*]_k > 0$, the constraint $v_k \geqslant 0$ is not active; therefore, for all $\theta \geqslant 0$, the parameter α must satisfy $(\theta \mathbf{e}_k - \boldsymbol{v}_\varkappa^*)^T \nabla D_\alpha(\boldsymbol{v}_\varkappa^*) = 0$ where $\mathbf{e}_k$ stands for the k-th element of the canonical basis of $\mathbb{R}^N$ (so that $[\boldsymbol{v}_\varkappa^*]_k = \mathbf{e}_k^T \boldsymbol{v}_\varkappa^*$). Considering $\theta = 0$ directly leads to the expected result (and one can also easily verify that the obtained value of α does not depend on $\theta \geqslant 0$). The last assertion follows directly from Lemma 3.1-(vii) and the correspondence between solutions to problems (3.3) and (3.4). □

Proof of Lemma 4.1. 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 (4.1) to $\mathcal{G}_S$ then yields, for $\boldsymbol{\beta} \in \mathbb{R}^N$,

$$\underset{\boldsymbol{\beta}}{\text{minimise}}\ \tfrac{1}{2}\boldsymbol{\beta}^T \mathbf{S}\boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{S}\boldsymbol{\omega} \text{ subject to } \mathbf{S}\boldsymbol{\beta} \geqslant -\alpha \mathbf{f}. \tag{C.3}$$

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

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

The primal optimality conditions give $\mathbf{S}\boldsymbol{\beta} = \mathbf{S}(\boldsymbol{v} - \boldsymbol{\omega})$, leading to the Lagrange dual problem (3.3) (written as a minimisation problem). If $\boldsymbol{v}_\alpha^*$ is a solution to (3.4), then a solution $\boldsymbol{\beta}^*$ to (C.3) needs to satisfy $\mathbf{S}\boldsymbol{\beta}^* = \mathbf{S}(\boldsymbol{v}_\alpha^* - \boldsymbol{\omega})$, so that we can in particular consider $\boldsymbol{\beta}^* = \boldsymbol{v}_\varkappa^* - \boldsymbol{\omega}$. 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 $g_\alpha^*(x_k) = -\alpha f_k$ for all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{v}_\alpha^*]_k > 0$ is consequence of the complementary slackness condition $(\boldsymbol{v}_\alpha^*)^T\left[\mathbf{S}(\boldsymbol{v}_\alpha^* - \boldsymbol{\omega}) + \alpha \mathbf{f}\right] = 0$. □

Proof of Lemma 4.2. We follow the same reasoning than in the proof of Lemma 4.1. By restricting problem (4.3) to $\mathcal{G}_S$, we obtain, for $\boldsymbol{\beta} \in \mathbb{R}^N$,

$$\underset{\boldsymbol{\beta},\gamma}{\text{minimise}}\ \tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\omega} - \gamma \text{ subject to } \mathbf{S}\boldsymbol{\beta} \geqslant \gamma\mathbf{f}/\varkappa. \tag{C.4}$$

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

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

The primal optimality conditions give $\mathbf{S}\boldsymbol{\beta} = \mathbf{S}(\boldsymbol{\upsilon} - \boldsymbol{\omega})$ and $\mathbf{f}^T\boldsymbol{\upsilon} = \varkappa$, leading to the Lagrange dual problem (3.4). If $\boldsymbol{\upsilon}^*_\varkappa$ is a solution to (3.4), then a solution $\boldsymbol{\beta}^*$ to (C.4) 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 $(\boldsymbol{\upsilon}^*_\varkappa)^T\big[\mathbf{S}(\boldsymbol{\upsilon}^*_\varkappa - \boldsymbol{\omega}) - \gamma^*_\varkappa\mathbf{f}/\varkappa\big] = 0$. The equality $g^*_\varkappa(x_k) = \gamma^*_\varkappa f_k/\varkappa$ for all $k \in \{1, \cdots, N\}$ such that $[\boldsymbol{\upsilon}^*_\varkappa]_k > 0$ is also consequence of the complementary slackness condition. □

Proof of Lemma 5.2. Let $\boldsymbol{\upsilon}_\alpha$ be such that $[\boldsymbol{\upsilon}_\alpha]_{J^c} = 0$ and $[\boldsymbol{\upsilon}_\alpha]_J = \mathbf{S}^{-1}_{J,J}([\mathbf{S}\boldsymbol{\omega}]_J - \alpha\mathbf{f}_J)$. Following (5.3), from the condition $[\mathbf{S}(\boldsymbol{\upsilon}_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{f}]_{J^c} \geqslant 0$, we define α_+ as the smallest α satisfying the constraint $\alpha\big[\mathbf{M}\mathbf{f}_J - \mathbf{f}_{J^c}\big]_l \leqslant \big[\mathbf{M}[\mathbf{S}\boldsymbol{\omega}]_J - [\mathbf{S}\boldsymbol{\omega}]_{J^c}\big]_l$, for all $l \in \{1, \cdots, \mathrm{card}(J^c)\}$. By definition, this constraint is satisfied by α_p, the components l such that $\big[\mathbf{M}\mathbf{d}_J - \mathbf{f}_{J^c}\big]_l \geqslant 0$ therefore carry no information, and the problem consists in searching for the smallest α such that

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

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

Proof of Lemma 5.1. 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

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

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

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

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 $\boldsymbol{\upsilon}^T_\alpha\big(\mathbf{S}(\boldsymbol{\upsilon}_\alpha - \boldsymbol{\omega}) + \alpha\mathbf{f}\big) = 0$, so that for all $\boldsymbol{\upsilon} \geqslant 0$, the optimality condition $(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_\alpha)^T\nabla D_\alpha(\boldsymbol{\upsilon}_\alpha) \geqslant 0$ holds, i.e., $\boldsymbol{\upsilon}_\alpha$ is a solution to (3.3), and $J_\alpha = J$. □

REFERENCES

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

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

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

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

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

[6] Bertrand Gauthier and Luc Pronzato. Convex relaxation for IMSE optimal design in random-field models. *Computational Statistics and Data Analysis*, 2016. `http://dx.doi.org/10.1016/j.csda.2016.10.018`.

[7] 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.

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

[9] Wolfgang Hackbusch. *Integral Equations: Theory and Numerical Treatment*, volume 120. Birkhäuser, 2012.

[10] 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.
[11] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press, 2015.
[12] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.
[13] Moshe Lichman. UCI Machine Learning Repository, 2013.
[14] Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, 1992.
[15] 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.
[16] Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.
[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] Christoph Schwab and Radu Alexandru Todor. Karhunen–Loève approximation of random fields by generalized fast multipole methods. *Journal of Computational Physics*, 217(1):100–122, 2006.
[20] Laurent Schwartz. *Analyse Hilbertienne*. Hermann, 1978.
[21] Steve Smale and Ding-Xuan Zhou. Geometry on probability spaces. *Constructive Approximation*, 30(3):311–323, 2009.
[22] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.
[23] 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.