

## OPTIMAL QUADRATURE-SPARSIFICATION FOR INTEGRAL OPERATOR APPROXIMATION\*

BERTRAND GAUTHIER<sup>†</sup> AND JOHAN A. K. SUYKENS<sup>‡</sup>

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

**Key words.** sparse quadrature, spectral approximation, RKHS, squared-kernel discrepancy,  $\ell^1$ -type penalization, convex quadratic programming, one-class SVM

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

**DOI.** 10.1137/17M1123614

**1. Introduction.** This work addresses the problem of designing sparse quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels. In parallel, we investigate the computation of accurate approximations of the main eigenpairs of a given initial operator (i.e., the pairs related to the largest eigenvalues) and the assessment of the accuracy of these approximations. From a numerical perspective, we pay special attention to *quadrature-sparsification* problems, which consist in designing a sparse quadrature from a fixed finite set of

---

\*Submitted to the journal’s Methods and Algorithms for Scientific Computing section March 31, 2017; accepted for publication (in revised form) July 5, 2018; published electronically October 25, 2018.

<http://www.siam.org/journals/sisc/40-5/M112361.html>

**Funding:** This work was supported by the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) ERC AdG A-DATADRIVE-B (290923); Research Council KUL: CoE PFV/10/002 (OPTec); the Flemish government: FWO projects G.0377.12 and G.088114N; and the Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO). This paper reflects only the authors’ views; the European Union is not liable for any use that may be made of the contained information.

<sup>†</sup>Corresponding author. School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, CF24 4AG, United Kingdom, and KU Leuven, ESAT-STADIUS Centre for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (gauthierb@cardiff.ac.uk).

<sup>‡</sup>KU Leuven, ESAT-STADIUS Centre for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (johan.suykens@esat.kuleuven.be).

candidate support points; this framework in particular encompasses the column-sampling problem (or landmark-selection problem) for the approximation of large-scale kernel matrices; see, for instance, [8, 11, 1].

**1.1. Motivations.** The spectral decomposition of an operator defined from a discrete measure supported by  $n$  points involves the diagonalization of an  $n \times n$  matrix; in the general case, the amount of computations required to perform this task scales as  $\mathcal{O}(n^3)$  and becomes numerically intractable for large values of  $n$  (not to mention storage issues). In practice, dealing with sparse quadratures, that is, discrete measures supported by a small number of points, is therefore especially important when one aims at computing the spectral decomposition of an approximate operator in order to approximate the eigendecomposition of an initial operator. Due to this sparsity constraint, the choice of the quadrature can strongly impact the quality of the induced approximation, naturally raising questions relative to the characterization and the construction of quadratures leading to accurate spectral approximations, and to the assessment of the accuracy of the induced approximations.

Following, for instance, [22, 23], under a trace-class condition, integral operators defined from a same positive-semidefinite kernel can be interpreted as Hilbert–Schmidt operators on the reproducing kernel Hilbert space (RKHS; see, for instance, [3]) associated with the kernel. In this framework, the squared Hilbert–Schmidt norm of the difference between the initial and approximate operators appears as a natural criterion to assess the approximation accuracy. Since the considered squared Hilbert–Schmidt norm can be expressed from integrals involving the square of the kernel, and by analogy with the notion of kernel discrepancy (see, for instance, [5, 21] and Appendix A), we refer to this criterion as the *squared-kernel discrepancy* between the initial and approximate measures (i.e., the measures defining, in combination with the kernel, the initial and approximate operators). The squared-kernel discrepancy can in addition be interpreted as a “weighted spectral sum-of-squared-errors-type criterion,” further highlighting the interest of low squared-kernel-discrepancy configurations for spectral approximation.

For a given initial measure and for a fixed quadrature size  $n$ , the search of an approximate measure minimizing the squared-kernel discrepancy among all measures supported by  $n$  points is generally a difficult nonconvex optimization problem. Nevertheless, for approximate measures with support included in a fixed finite set of points, the squared-kernel discrepancy can be expressed as a convex quadratic function, and sparsity of the approximate measure can be promoted through the introduction of an  $\ell^1$ -type penalization. In such a quadrature-sparsification framework, the induced penalized squared-kernel-discrepancy minimization problems consist in convex quadratic programs (QPs) that can be solved efficiently in the range of relatively sparse solutions, even for large-scale problems. From a matrix-approximation perspective, penalized squared-kernel-discrepancy minimization defines a deterministic, QP-based, weighted column-sampling scheme and appears as a complement to the existing column-sampling-based methodology for kernel-matrix approximation; see, e.g., [8, 26, 15, 25, 4, 11] for an overview.

**1.2. Contribution and organization of the paper.** This work aims at investigating the relevance of the penalized squared-kernel-discrepancy minimization framework for the computation of accurate approximations of the main eigenpairs of integral operators related to symmetric positive-semidefinite kernels. We are thus addressing two different, but nevertheless strongly intricate, problems: the design of sparse quadratures and the computation of accurate approximations of the main eigen-

pairs of a given initial operator. We present a careful analysis of the approach and describe numerical strategies to tackle large-scale penalized squared-kernel-discrepancy minimization problems.

To assess the accuracy of an approximate eigendirection (that is, an eigendirection of the approximate operator), we rely on the notion of *geometric approximate eigenvalues* (see Definition 3.2; we also use the *orthogonality test*, see Remark 3.1). For a given approximate eigendirection, the geometric approximate eigenvalues consist in four different approximations of the underlying eigenvalue. These approximations verify various optimality properties and are equal if and only if the related approximate eigendirection is an eigendirection of the initial operator; furthermore, the concordance between these four approximations is directly related to the accuracy of the approximate eigendirection, as detailed in Theorem 3.1.

As an important feature, the so obtained approximate eigenpairs are invariant under rescaling of the approximate measure, i.e., proportional approximate measures lead to the same spectral approximation of a given initial operator (see Lemma 3.1). Motivated by this invariance property, we introduce the notion of *conic squared-kernel discrepancy*, consisting in the minimum of the squared-kernel discrepancy on the rays of proportional approximate measures. The conic squared-kernel discrepancy is directly related to the overall accuracy of the spectral approximation, as detailed in Theorem 3.2.

For quadrature-sparsification problems, Theorem 5.1 gives an insight into the impact of the penalization on the trade-off between sparsity and accuracy of the spectral approximation. This result indeed provides a sufficient condition under which increasing the amount of penalization tends to increase the sparsity of the approximate measures (more precisely, this decreases an upper bound on the number of support points of the optimal approximate measures), at the expense of reducing the overall accuracy of the induced spectral approximations.

In the quadrature-sparsification framework, the  $\ell^1$ -type penalization can be introduced under the form of a regularization term or of a constraint and is based on the definition of a *penalization direction*. A penalization direction of special interest consists, for instance, in penalizing the trace of the approximate operators, leading to an interesting parallel with the approximation by spectral truncation; alternative choices for the penalization direction are nevertheless possible, and the definition of relevant problem-dependent penalization directions is discussed. The regularized and constrained formulations are equivalent, and the properties of the corresponding QPs are investigated. In particular, these QPs can be interpreted as the Lagrange duals of *distorted one-class support-vector machines* (SVMs; see, e.g., [24]) defined from the squared kernel, the initial measure, and the penalization term, so that the points selected through penalized squared-kernel-discrepancy minimization correspond to the support vectors of these SVMs.

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

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

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

**2.1. Integral operators.** We assume that  $\mathcal{X}$  is a measurable space and we denote by  $\mathcal{A}$  the underlying  $\sigma$ -algebra. We suppose that the kernel  $K(\cdot, \cdot)$  is measurable on  $\mathcal{X} \times \mathcal{X}$  for the product  $\sigma$ -algebra  $\mathcal{A} \otimes \mathcal{A}$  (see, for instance, [24, Chap. 4]), so that  $\mathcal{H}$  consists of measurable functions on  $\mathcal{X}$ . We also assume that the diagonal of  $K(\cdot, \cdot)$ , i.e., the function  $x \mapsto K(x, x)$ , is measurable on  $(\mathcal{X}, \mathcal{A})$ . We denote by  $\mathcal{M}$  the set of all measures on  $(\mathcal{X}, \mathcal{A})$  and we introduce

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

For  $\mu \in \mathcal{T}(K)$ , 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  $\mathcal{H}$ )

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

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

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

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

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

where  $(\cdot|\cdot)_{\mathcal{H}}$  and  $(\cdot|\cdot)_{L^2(\mu)}$  stand for the inner products of  $\mathcal{H}$  and  $L^2(\mu)$ , respectively; see, for instance, [9, 10] for more details.

We introduce the closed linear subspaces  $\mathcal{H}_{0\mu} = \{h \in \mathcal{H} \mid \|h\|_{L^2(\mu)} = 0\}$  and  $\mathcal{H}_\mu = \mathcal{H}_{0\mu}^\perp$  (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  $T_\mu$  (repeated according to their algebraic multiplicity) and let  $\{\tilde{\varphi}_k\}_{k \in \mathbb{I}_\mu^+}$  be a set of associated eigenfunctions, chosen to be orthonormal in  $L^2(\mu)$ , i.e.,  $\tilde{\varphi}_k \in L^2(\mu)$ ,  $T_\mu[\tilde{\varphi}_k] = \lambda_k \tilde{\varphi}_k$  in  $L^2(\mu)$ , and  $(\tilde{\varphi}_k | \tilde{\varphi}_{k'})_{L^2(\mu)} = \delta_{k,k'}$  (Kronecker delta). For  $k \in \mathbb{I}_\mu^+$ , let  $\varphi_k = \frac{1}{\lambda_k} T_\mu[\tilde{\varphi}_k] \in \mathcal{H}$  be the *canonical extension* of  $\tilde{\varphi}_k$  (the eigenfunctions  $\tilde{\varphi}_k$  are indeed only defined  $\mu$ -almost everywhere, while the extensions  $\varphi_k$  are defined for all  $x \in \mathcal{X}$ ). From (2.1), we obtain that  $\{\sqrt{\lambda_k} \varphi_k\}_{k \in \mathbb{I}_\mu^+}$  is an orthonormal basis (o.n.b.) of the subspace  $\mathcal{H}_\mu$  of  $\mathcal{H}$ , and the reproducing kernel  $K_\mu(\cdot, \cdot)$  of  $\mathcal{H}_\mu$  is thus given by, for all  $x$  and  $t \in \mathcal{X}$ ,

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

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

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

**2.2. Hilbert–Schmidt norm and squared-kernel discrepancy.** In view of section 2.1, for  $\mu \in \mathcal{T}(K)$ , the operator  $T_\mu$  can also be interpreted as an operator on  $\mathcal{H}$  (see, e.g., [22, 23]); with a slight abuse of notation, we keep the same notation for “ $T_\mu$  viewed as an operator on  $L^2(\mu)$ ,” and “ $T_\mu$  viewed as an operator on  $\mathcal{H}$ .” In both cases,  $T_\mu$  is a Hilbert–Schmidt operator.

We denote by  $\text{HS}(\mathcal{H})$  the Hilbert space of all Hilbert–Schmidt operators on  $\mathcal{H}$ . Let  $\mu$  and  $\nu \in \mathcal{T}(K)$ ; 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  $T_\mu$  and  $T_\nu$  on  $\mathcal{H}$  is given by

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

and we recall that the value of  $(T_\mu | T_\nu)_{\text{HS}(\mathcal{H})}$  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

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

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

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

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

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

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

In particular, notice that  $\|K\|_{L^2(\mu \otimes \nu)}^2 \leq \tau_\mu \tau_\nu$  and that  $\|T_\mu\|_{\text{HS}(\mathcal{H})}^2 = \sum_{k \in \mathbb{I}_\mu^+} \lambda_k^2$ , where  $\{\lambda_k\}_{k \in \mathbb{I}_\mu^+}$  is the set of all strictly positive eigenvalues of  $T_\mu$ . By definition, we always have  $D_{K^2}(\mu, \nu) \geq 0$ , and  $D_{K^2}(\mu, \mu) = 0$ . We can also remark that if  $\mu$  and  $\nu \in \mathcal{T}(K)$  are such that  $\mathcal{H}_\mu$  and  $\mathcal{H}_\nu$  are orthogonal subspaces of  $\mathcal{H}$ , then  $\|K\|_{L^2(\mu \otimes \nu)}^2 = 0$ .

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

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

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

The terminology “squared-kernel discrepancy” is motivated by the analogy with the notion of “kernel discrepancy” discussed, for instance, in [5, 21] (see Appendix A). Interestingly, the kernel discrepancy is related to approximate integration of functions in the RKHS  $\mathcal{H}$ , while the squared-kernel discrepancy is related to the approximation of integral operators defined from the reproducing kernel  $K(\cdot, \cdot)$  of  $\mathcal{H}$ ; by definition, the

squared-kernel discrepancy is thus also related to approximate integration of functions in the RKHS  $\mathcal{G}$  associated with the squared kernel  $K^2(\cdot, \cdot)$ .

LEMMA 2.2. *Let  $\mu$  and  $\nu \in \mathcal{T}(K)$  be such that  $\mathcal{H}_\nu \subset \mathcal{H}_\mu$  (i.e., for  $h \in \mathcal{H}$ , if  $\|h\|_{L^2(\mu)} = 0$ , then  $\|h\|_{L^2(\nu)} = 0$ ), and denote by  $\{\sqrt{\lambda_k}\varphi_k\}_{k \in \mathbb{I}_\mu^+}$  an o.n.b. of  $\mathcal{H}_\mu$  defined by the spectral decomposition of  $T_\mu$ . We have*

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

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

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

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

Since  $D_{K^2}(\mu, \mu) = 0$  (i.e., “the best approximation of  $T_\mu$  is  $T_\mu$  itself”), the unconstrained minimization of  $\nu \mapsto D_{K^2}(\mu, \nu)$  on  $\mathcal{T}(K)$  is of no interest. Furthermore, in the framework of sparse pointwise quadrature approximation, we aim at obtaining a discrete measure  $\nu$  supported by a relatively small number of points (in order to be able to compute the eigendecomposition of  $T_\nu$ ) 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  $\nu_n^*$  such that  $D_{K^2}(\mu, \nu_n^*)$  is minimal among all measures  $\nu_n$  supported by  $n$  points is in general a difficult (i.e., usually nonconvex) optimization problem on  $(\mathcal{X} \times \mathbb{R}_+)^n$ . To avoid this difficulty, we restrict the squared-kernel-discrepancy minimization to measures  $\nu$  with support included in a fixed finite set of points  $\mathcal{S} = \{x_k\}_{k=1}^N$  (with, in practice,  $N$  large); see section 4.2. In addition, instead of fixing a priori the number  $n$  of support points, we promote sparsity through the introduction of an  $\ell^1$ -type penalization, as considered in section 5.

**3. Approximate eigendecomposition.** We consider two measures  $\mu$  and  $\nu \in \mathcal{T}(K)$ , corresponding to an initial operator  $T_\mu$  and an approximate operator  $T_\nu$ .

**3.1. Geometric approximate eigenvalues.** Following section 2.1, we denote by  $\{\sqrt{\lambda_k}\varphi_k\}_{k \in \mathbb{I}_\mu^+}$  an o.n.b. of  $\mathcal{H}_\mu$  defined by the eigendecomposition of  $T_\mu$ . In the same way, let  $\{\sqrt{\vartheta_l}\psi_l\}_{l \in \mathbb{I}_\nu^+}$  be an o.n.b. of the subspace  $\mathcal{H}_\nu$  of  $\mathcal{H}$  related to  $T_\nu$ , i.e.,  $T_\nu[\psi_l] = \vartheta_l\psi_l \in \mathcal{H}$ , with  $\vartheta_l > 0$  and  $(\psi_l|\psi_{l'})_{L^2(\nu)} = \delta_{l,l'}$ ; in particular, the reproducing kernel  $K_\nu(\cdot, \cdot)$  of the subspace  $\mathcal{H}_\nu$  of  $\mathcal{H}$  thus verifies

$$(3.1) \quad K_\nu(x, t) = \sum_{l \in \mathbb{I}_\nu^+} \vartheta_l \psi_l(x) \psi_l(t) \text{ for all } x \text{ and } t \in \mathcal{X}.$$

We shall refer to the functions  $\psi_l$  as the *approximate eigendirections* of  $T_\mu$  induced by  $T_\nu$ . We recall that, from (2.1), we have

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

**DEFINITION 3.1.** For all  $l \in \mathbb{I}_\nu^+$  such that  $\|\psi_l\|_{L^2(\mu)} > 0$  (i.e.,  $\psi_l \in \mathcal{H}_\mu$ ), we introduce  $\widehat{\varphi}_l = \psi_l / \|\psi_l\|_{L^2(\mu)}$ , and we refer to  $\widehat{\varphi}_l$  as a *normalized approximate eigenfunction* of  $T_\mu$  induced by the spectral decomposition of  $T_\nu$ .

We introduce  $\widetilde{\mathbb{I}}_\nu^+ = \{l \in \mathbb{I}_\nu^+ | \psi_l \in \mathcal{H}_\mu\}$ , so that the functions  $\widehat{\varphi}_l$  are well defined for all  $l \in \widetilde{\mathbb{I}}_\nu^+$ . Notice that if  $\mathcal{H}_\nu \subset \mathcal{H}_\mu$ , then we have  $\widetilde{\mathbb{I}}_\nu^+ = \mathbb{I}_\nu^+$ . In particular, if  $\psi_l \in \mathcal{H}_{0\mu}$ , then  $T_\mu[\psi_l] = 0$  and such a direction  $\psi_l$  is therefore of no use in approximating the eigendirections related to the strictly positive eigenvalues of  $T_\mu$ .

*Remark 3.1* (orthogonality test). The normalized approximate eigenfunctions  $\widehat{\varphi}_l$  are by definition orthogonal in  $L^2(\nu)$  and in  $\mathcal{H}$  and verify  $\|\widehat{\varphi}_l\|_{L^2(\mu)} = 1$ . Controlling the orthogonality, in  $L^2(\mu)$ , between the approximations  $\widehat{\varphi}_l$ , with  $l \in \widetilde{\mathbb{I}}_\nu^+$ , appears as a relatively affordable way to assess their accuracy. Indeed, from (2.1) and due to their orthogonality in  $\mathcal{H}$ , accurate normalized approximate eigenfunctions  $\widehat{\varphi}_l$  should be almost mutually orthogonal in  $L^2(\mu)$ . Notice that this condition is, however, only a necessary condition. See sections 10 and 11 for illustrations; a further insight into the relevance of the orthogonality test is given in Remark 3.2.

It is very instructive to try to estimate the eigenvalue, for the operator  $T_\mu$ , related to an approximate eigendirection  $\psi_l$  induced by  $T_\nu$ , as discussed hereafter.

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

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

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

We refer to  $\widehat{\lambda}_l^{[1]}$ ,  $\widehat{\lambda}_l^{[2]}$ ,  $\widehat{\lambda}_l^{[3]}$ , and  $\widehat{\lambda}_l^{[4]}$  as the *four geometric approximate eigenvalues* of  $T_\mu$  related to the approximate eigendirection  $\psi_l$  induced by  $T_\nu$ .

The intuition behind these four approximate eigenvalues  $\widehat{\lambda}_l^{[1]}$  is further discussed in the proof of Theorem 3.1 (Appendix B); see Remark 3.3 for comments relative to their computation. The various expressions characterizing  $\widehat{\lambda}_l^{[1]}$ ,  $\widehat{\lambda}_l^{[3]}$ , and  $\widehat{\lambda}_l^{[4]}$  follow from (2.1) and Definition 3.1; in particular, notice that if  $\|\psi_l\|_{L^2(\mu)} > 0$ , then  $\sqrt{\widehat{\lambda}_l^{[1]}} \widehat{\varphi}_l = \sqrt{\vartheta_l} \psi_l$ .

**THEOREM 3.1.** For all  $l \in \widetilde{\mathbb{I}}_\nu^+$ , we have  $\widehat{\lambda}_l^{[1]} \leq \widehat{\lambda}_l^{[2]} \leq \widehat{\lambda}_l^{[3]} \leq \widehat{\lambda}_l^{[4]}$ , with equality if and only if  $\psi_l$  is an eigendirection of the operator  $T_\mu$  (on  $L^2(\mu)$  or on  $\mathcal{H}$ ). In case of equality, the approximation  $\widehat{\lambda}_l^{[1]}$  corresponds exactly to the eigenvalue of  $T_\mu$  related to the eigendirection  $\psi_l$ ; in particular, equality between the four geometric approximate eigenvalues occurs as soon as two of them are equal.

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

$$(3.2) \quad \lambda \mapsto \|\lambda\sqrt{\vartheta_l}\psi_l - T_\mu[\sqrt{\vartheta_l}\psi_l]\|_{\mathcal{H}}^2 = \lambda^2 - 2\lambda\widehat{\lambda}_l^{[1]} + (\widehat{\lambda}_l^{[2]})^2$$

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

$$(3.3) \quad \lambda \mapsto \|\lambda\widehat{\varphi}_l - T_\mu[\widehat{\varphi}_l]\|_{L^2(\mu)}^2 = \lambda^2 - 2\lambda\widehat{\lambda}_l^{[3]} + (\widehat{\lambda}_l^{[4]})^2$$

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

In view of Theorem 3.1, for  $l \in \widetilde{\mathbb{I}}_\nu^+$  (so that  $\widehat{\lambda}_l^{[1]} > 0$ ), one may assess the accuracy of an approximate eigendirection  $\psi_l$  (as eigendirection of  $T_\mu$ ) by checking how close to each other are the approximations  $\widehat{\lambda}_l^{[1]}$ ,  $\widehat{\lambda}_l^{[2]}$ ,  $\widehat{\lambda}_l^{[3]}$ , and  $\widehat{\lambda}_l^{[4]}$ . From (3.2) and (3.3), we, for instance, have

$$(3.4) \quad \|\sqrt{\vartheta_l}\psi_l - T_\mu[\sqrt{\vartheta_l}\psi_l]/\widehat{\lambda}_l^{[1]}\|_{\mathcal{H}}^2 = (\widehat{\lambda}_l^{[2]}/\widehat{\lambda}_l^{[1]})^2 - 1 \text{ and}$$

$$(3.5) \quad \|\widehat{\varphi}_l - T_\mu[\widehat{\varphi}_l]/\widehat{\lambda}_l^{[3]}\|_{L^2(\mu)}^2 = (\widehat{\lambda}_l^{[4]}/\widehat{\lambda}_l^{[3]})^2 - 1,$$

so that the closer (3.4) and (3.5) are to zero, the more accurate is the approximate eigendirection  $\psi_l$ ; see sections 10 and 11 for illustrations. Notice that we have  $0 < \widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]} \leq 1$  and that this ratio corresponds to the inner product, in  $\mathcal{H}$ , between the normalized functions  $\sqrt{\vartheta_l}\psi_l$  and  $T_\mu[\sqrt{\vartheta_l}\psi_l]/\|T_\mu[\sqrt{\vartheta_l}\psi_l]\|_{\mathcal{H}}$ . In the same way, we have  $0 < \widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]} \leq 1$ , and this ratio corresponds to the inner product, in  $L^2(\mu)$ , between the normalized functions  $\widehat{\varphi}_l$  and  $T_\mu[\widehat{\varphi}_l]/\|T_\mu[\widehat{\varphi}_l]\|_{L^2(\mu)}$ .

*Remark 3.2.* Consider the spectral approximation of the initial operator  $T_\mu$  induced by the approximate operator  $T_\nu$ ; see Definitions 3.1 and 3.2. From (3.1), we obtain

$$(3.6) \quad \|K - K_\nu\|_{L^2(\mu \otimes \mu)}^2 = \|K_\mu - K_\nu\|_{L^2(\mu \otimes \mu)}^2 = \sum_{k \in \widetilde{\mathbb{I}}_\mu^+} \lambda_k^2 + \sum_{l \in \widetilde{\mathbb{I}}_\nu^+} (\widehat{\lambda}_l^{[1]})^2 - 2 \sum_{l \in \widetilde{\mathbb{I}}_\nu^+} \widehat{\lambda}_l^{[1]}\widehat{\lambda}_l^{[3]} + \sum_{l \neq l' \in \widetilde{\mathbb{I}}_\nu^+} \widehat{\lambda}_l^{[1]}\widehat{\lambda}_{l'}^{[1]} (\widehat{\varphi}_l | \widehat{\varphi}_{l'})_{L^2(\mu)}^2.$$

Equation (3.6) further illustrates the conclusions drawn from Remark 3.1 and Theorem 3.1. We can indeed, for instance, remark that if we have  $\widehat{\lambda}_l^{[1]}\widehat{\lambda}_l^{[3]} \approx (\widehat{\lambda}_l^{[1]})^2$ , for all  $l \in \widetilde{\mathbb{I}}_\nu^+$ , and if the normalized approximate eigenfunctions  $\widehat{\varphi}_l$  are almost mutually orthogonal in  $L^2(\mu)$ , then the kernel  $K_\nu(\cdot, \cdot)$  is an accurate low-rank approximation of the kernel  $K(\cdot, \cdot)$  in  $L^2(\mu \otimes \mu)$ , i.e., the kernel  $K_\nu(\cdot, \cdot)$  accurately approximates a low-rank approximation of  $K_\mu(\cdot, \cdot)$  obtained by truncation of the expansion (2.2). Notice that the reciprocal of this reasoning also holds and that this remark can be extended to the approximate kernels obtained by truncation of the expansion (3.1) of the kernel  $K_\nu(\cdot, \cdot)$ .

*Remark 3.3.* Once  $\vartheta_l$  and  $\psi_l$  are known, we obtain the normalized approximate eigenfunction  $\widehat{\varphi}_l$  and the approximate eigenvalue  $\widehat{\lambda}_l^{[1]}$  by simply evaluating  $\|\psi_l\|_{L^2(\mu)}^2$ . Computing the other approximate eigenvalues  $\widehat{\lambda}_l^{[2]}$ ,  $\widehat{\lambda}_l^{[3]}$ , and  $\widehat{\lambda}_l^{[4]}$  requires the knowledge of  $T_\mu[\psi_l]$ . We can then obtain  $\widehat{\lambda}_l^{[3]}$  and  $\widehat{\lambda}_l^{[4]}$  by evaluating an inner product in  $L^2(\mu)$ , and derive  $\widehat{\lambda}_l^{[2]}$  from the relation  $\widehat{\lambda}_l^{[2]} = \sqrt{\widehat{\lambda}_l^{[1]}\widehat{\lambda}_l^{[3]}}$ .



We have to compute  $T_\mu[\psi_l]$  (which may prove challenging) only when we are interested in assessing precisely the accuracy of an approximate eigendirection  $\psi_l$  of  $T_\mu$ . Otherwise, we might simply consider the approximate eigenpairs  $\{\widehat{\lambda}_l^{[1]}, \widehat{\varphi}_l\}_{l \in \widetilde{\mathbb{I}}_\nu^+}$  (see also Remark 3.4), while eventually checking the orthogonality, in  $L^2(\mu)$ , between the normalized approximate eigendirections (orthogonality test; see Remark 3.1).

The computation of the geometric approximate eigenvalues when  $\mu$  is a discrete measure with finite support is further discussed in section 4.3.

Following Remark 2.1, for any  $\nu \in \mathcal{T}(K)$  and for any  $c > 0$ , we have  $K_\nu(\cdot, \cdot) = K_{c\nu}(\cdot, \cdot)$  and  $\mathcal{H}_\nu = \mathcal{H}_{c\nu}$ ; also notice that, as operators on  $\mathcal{H}$ , we have  $T_{c\nu} = cT_\nu$ . Lemma 3.1 points out the invariance of the spectral approximations induced by proportional approximate measures; this invariance follows directly from Remark 2.1 and Definitions 3.1 and 3.2 (so that we don't further detail the proof).

LEMMA 3.1. *For any approximate measure  $\nu \in \mathcal{T}(K)$  and for a given initial operator  $T_\mu$ , 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 we replace  $\nu$  by  $c\nu$  for any  $c > 0$ .*

**3.2. Conic squared-kernel discrepancy.** In the framework of section 3.1 and in view of Lemma 3.1, proportional (nonnull) approximate measures lead to the same spectral approximation of  $T_\mu$ . For a given measure  $\nu \in \mathcal{T}(K)$ , we can thus search the value of  $c \geq 0$  for which  $D_{K^2}(\mu, c\nu)$  is minimal.

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

$$c_\nu = \frac{\|K\|_{L^2(\mu \otimes \nu)}^2}{\|K\|_{L^2(\nu \otimes \nu)}^2} \text{ and } \phi(c_\nu) = \|K\|_{L^2(\mu \otimes \mu)}^2 - \frac{\|K\|_{L^2(\mu \otimes \nu)}^4}{\|K\|_{L^2(\nu \otimes \nu)}^2}.$$

In particular,  $T_{c_\nu \nu}$  is the orthogonal projection, in  $\text{HS}(\mathcal{H})$ , of  $T_\mu$  onto the linear subspace spanned by  $T_\nu$ ; in addition,  $\|T_{c_\nu \nu} - \frac{1}{2}T_\mu\|_{\text{HS}(\mathcal{H})}^2 = \frac{1}{4}\|T_\mu\|_{\text{HS}(\mathcal{H})}^2$ , so that, in  $\text{HS}(\mathcal{H})$ , the approximate operator  $T_{c_\nu \nu}$  lies on a sphere centered at  $\frac{1}{2}T_\mu$  and with radius  $\frac{1}{2}\|T_\mu\|_{\text{HS}(\mathcal{H})}$ . We also have

$$(3.7) \quad \sum_{l \in \widetilde{\mathbb{I}}_\nu^+} \widehat{\lambda}_l^{[1]} \|T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l^{[1]} \widehat{\varphi}_l\|_{\mathcal{H}}^2 \leq \sum_{l \in \widetilde{\mathbb{I}}_\nu^+} \widehat{\lambda}_l^{[1]} \|T_\mu[\widehat{\varphi}_l] - c_\nu \vartheta_l \widehat{\varphi}_l\|_{\mathcal{H}}^2 \leq D_{K^2}(\mu, c_\nu \nu) \text{ and}$$

$$(3.8) \quad \sum_{l \in \widetilde{\mathbb{I}}_\nu^+} \widehat{\lambda}_l^{[1]} \|T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l^{[3]} \widehat{\varphi}_l\|_{L^2(\mu)}^2 \leq \sum_{l \in \widetilde{\mathbb{I}}_\nu^+} \widehat{\lambda}_l^{[1]} \|T_\mu[\widehat{\varphi}_l] - \widehat{\lambda}_l^{[1]} \widehat{\varphi}_l\|_{L^2(\mu)}^2 \leq \tau_\mu D_{K^2}(\mu, c_\nu \nu).$$

In Theorem 3.2, we are exploiting the positive cone structure of  $\mathcal{T}(K)$ ; we thus refer to  $\phi(c_\nu) = D_{K^2}(\mu, c_\nu \nu)$  as the *conic squared-kernel discrepancy* between  $\mu$  and  $\nu$  (notice that the measure  $\mu$  is fixed); to avoid confusion, we shall sometimes refer to  $D_{K^2}(\mu, \nu)$  as the *raw squared-kernel discrepancy* between  $\mu$  and  $\nu$ . The operator  $T_{c_\nu \nu}$  is the best approximation of  $T_\mu$  (in terms of squared-kernel discrepancy) among all operators defined from measures proportional to  $\nu$ , i.e., of the form  $c\nu$ , with  $c \geq 0$ . In view of (3.7) and (3.8), the conic squared-kernel discrepancy  $D_{K^2}(\mu, c_\nu \nu)$  is directly related to the overall accuracy of the spectral approximation of  $T_\mu$  induced by the operator  $T_\nu$ .

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

**4. The discrete case.** We now investigate in more detail the case of discrete measures with finite support. We pay particular attention to the situation where the initial measure  $\mu$  is discrete and the support of  $\nu$  is included in the support of  $\mu$ .

**4.1. Discrete measures and kernel matrices.** We first recall the connection between kernel matrices and integral operators related to discrete measures with finite support. Let  $\mu = \sum_{k=1}^N \omega_k \delta_{x_k}$  be a discrete measure supported by  $\mathcal{S} = \{x_k\}_{k=1}^N$ , with  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_N)^T \in \mathbb{R}^N$ ,  $\omega_k > 0$  for all  $k$  (in what follows, we use the notation  $\boldsymbol{\omega} > 0$ ), and where  $\delta_{x_k}$  is the Dirac measure (evaluation functional) at  $x_k \in \mathcal{X}$ ; we have  $\mu \in \mathcal{T}(K)$ , and for  $f \in L^2(\mu)$  and  $x \in \mathcal{X}$ , 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  $\mathbf{W} = \text{diag}(\boldsymbol{\omega})$ , and  $\mathbf{k}(x) = (K(x_1, x), \dots, K(x_N, x))^T$ , and  $\mathbf{f} = (f(x_1), \dots, f(x_N))^T \in \mathbb{R}^N$ . 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,  $f \in L^2(\mu)$  corresponds to  $\mathbf{f} \in \mathbb{R}^N$ , and the operator  $T_\mu$  then corresponds to the matrix  $\mathbf{K} \mathbf{W}$ , where  $\mathbf{K} \in \mathbb{R}^{N \times N}$  is the kernel matrix with  $i, j$  entry  $\mathbf{K}_{i,j} = K(x_i, x_j)$ ; in particular, we have  $\mathbf{K} \mathbf{W} \mathbf{f} = (T_\mu[f](x_1), \dots, T_\mu[f](x_N))^T$ .

We denote by  $\lambda_1 \geq \dots \geq \lambda_N \geq 0$  the eigenvalues of  $\mathbf{K} \mathbf{W}$  and by  $\mathbf{v}_1, \dots, \mathbf{v}_N$  a set of associated orthonormalized eigenvectors, i.e.,  $\mathbf{K} \mathbf{W} = \mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^{-1}$ , with  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_N)$  and  $\mathbf{P} = (\mathbf{v}_1 | \dots | \mathbf{v}_N)$ . The vectors  $\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$  form an o.n.b. of the Hilbert space  $\{\mathbb{R}^N, (\cdot | \cdot)_{\mathbf{W}}\}$ , i.e.,  $\mathbf{P}^T \mathbf{W} \mathbf{P} = \text{Id}_N$ , the  $N \times N$  identity matrix; since  $\boldsymbol{\omega} > 0$ , we also have

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

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

For a general  $\boldsymbol{\omega} > 0$ , the matrix  $\mathbf{K} \mathbf{W}$  is nonsymmetric; however, since  $\mathbf{K} \mathbf{W} \mathbf{v}_k = \lambda_k \mathbf{v}_k$ , we have

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

The symmetric matrix  $\mathbf{W}^{1/2} \mathbf{K} \mathbf{W}^{1/2}$  thus defines a symmetric and positive-semidefinite operator on the classical Euclidean space  $\{\mathbb{R}^N, (\cdot | \cdot)_{\text{Id}_N}\}$ , with eigenvalues  $\lambda_k$  and orthonormalized eigenvectors  $\mathbf{W}^{1/2} \mathbf{v}_k$ . We can thus easily deduce the eigendecomposition of the matrix  $\mathbf{K} \mathbf{W}$  viewed as an operator on  $\{\mathbb{R}^N, (\cdot | \cdot)_{\mathbf{W}}\}$  from the eigendecomposition of the symmetric matrix  $\mathbf{W}^{1/2} \mathbf{K} \mathbf{W}^{1/2}$ .

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

**4.2. Restricting the support of the approximate measure.** We consider a general measure  $\mu \in \mathcal{T}(K)$  and a fixed set  $\mathcal{S} = \{x_k\}_{k=1}^N$  of  $N$  points in  $\mathcal{X}$ . For a

measure  $\nu$  with support included in  $\mathcal{S}$ , i.e.,  $\nu = \sum_{k=1}^N v_k \delta_{x_k}$ , with  $\mathbf{v} = (v_1, \dots, v_N)^T \geq 0$  (that is,  $v_k \geq 0$  for all  $k$ ), we have

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

where  $\mathbf{S}$  is the matrix defined by the squared kernel  $K^2(\cdot, \cdot)$  and the set of points  $\mathcal{S}$ , i.e., with  $i, j$  entry  $\mathbf{S}_{i,j} = K^2(x_i, x_j) \geq 0$  (the kernel matrix  $\mathbf{S}$  is therefore nonnegative and symmetric positive-semidefinite), and where  $\mathbf{g}_\mu = (g_\mu(x_1), \dots, g_\mu(x_N))^T \in \mathbb{R}^N$ , with  $g_\mu(x_k) = \int_{\mathcal{X}} K^2(x_k, t) d\mu(t) \geq 0$ . Notice in particular that  $\mathbf{S} = \mathbf{K} * \mathbf{K}$  (Hadamard product), where we recall that  $\mathbf{K}$  is the kernel matrix defined by  $K(\cdot, \cdot)$  and  $\mathcal{S}$ , i.e.,  $\mathbf{K}_{i,j} = K(x_i, x_j)$ .

For such a discrete measure  $\nu$ , we obtain

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

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

Minimizing  $\mathbf{v} \mapsto \mathbf{v}^T \mathbf{S} \mathbf{v} - 2\mathbf{g}_\mu^T \mathbf{v}$  under the constraint  $\mathbf{v} \geq 0$  leads to the best approximation of  $\mu$ , in terms of squared-kernel discrepancy, among all discrete measures supported by  $\mathcal{S}$ . In practice, this minimization requires the knowledge of the vector  $\mathbf{g}_\mu \in \mathbb{R}^N$ , which might be problematic for general measures  $\mu$  (in this case, an approximation might be considered). In this work, we nevertheless more specifically aim at computing approximate measures supported by a number of points significantly smaller than  $N$ , so that we do not consider such a minimization; instead, we add an  $\ell^1$ -type penalization term to the squared-kernel discrepancy, as detailed in section 5.

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

For such measures  $\mu$  and  $\nu$  (related to vectors  $\boldsymbol{\omega} > 0$  and  $\mathbf{v} \geq 0$ , respectively), we have

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

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

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

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

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

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

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

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

Following section 3, we now illustrate how to compute the approximate eigendecomposition of the matrix  $\mathbf{KW}$  related to  $T_\mu$  induced by the matrix  $\mathbf{KV}$  related to  $T_\nu$ .

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

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

they satisfy  $\mathbf{KV} \mathbf{u}_l = \vartheta_l \mathbf{u}_l$  and  $\mathbf{u}_l^T \mathbf{V} \mathbf{u}_l = \delta_{l, l}$ . Notice that  $[\mathbf{u}_l]_I = (\mathbf{V}_{I, I})^{-1/2} \mathbf{a}_l$ , where  $[\mathbf{u}_l]_I \in \mathbb{R}^n$  consists in the components of  $\mathbf{u}_l$  with index in  $I$ .

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

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

Following Remark 3.3 and starting from a pair  $\{\vartheta_l, (\mathbf{V}_{I, I})^{-1/2} \mathbf{a}_l\}$ , the amount of computations required to obtain the extended components of the eigenvector  $\mathbf{u}_l$  scales as  $\mathcal{O}(n(N - n))$ . The measure  $\mu$  being supported by  $N$  points, computing an inner product in  $L^2(\mu)$  requires  $\mathcal{O}(N)$  operations. The computation of the normalized approximate eigenvector  $\widehat{\mathbf{v}}_l$  and of the approximate eigenvalue  $\widehat{\lambda}_l^{[1]}$  is therefore relatively inexpensive. To obtain  $\widehat{\lambda}_l^{[2]}$ ,  $\widehat{\lambda}_l^{[3]}$ , or  $\widehat{\lambda}_l^{[4]}$ , we need to compute

$$\mathbf{KW} \mathbf{u}_l = (T_\mu[\psi_l](x_1), \dots, T_\mu[\psi_l](x_N))^T,$$

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

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

Considering the reproducing kernel  $K_\nu(\cdot, \cdot)$  of the subspace  $\mathcal{H}_\nu$  of  $\mathcal{H}$  (see (3.1)), and following Remark 4.1, we introduce the  $N \times N$  kernel matrix  $\mathbf{K}_\nu$  defined by

$K_\nu(\cdot, \cdot)$  and  $\mathcal{S}$ , i.e., with  $i, j$  entry  $[\mathbf{K}_\nu]_{i,j} = K_\nu(x_i, x_j)$ . From the eigendecomposition  $[\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}]_{I,I} = \mathbf{A}\mathbf{\Theta}\mathbf{A}^T$  (with  $\mathbf{A}$  an  $n \times n$  orthogonal matrix), we deduce that

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

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

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

and in particular,  $\mathbf{V}^{1/2}\mathbf{K}_\nu\mathbf{V}^{1/2} = \mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}$ . Following, for instance, [8, 11], the matrix  $\mathbf{K}_\nu$  corresponds to the *Nyström approximation* of the kernel matrix  $\mathbf{K}$  induced by the approximate operator  $T_\nu$  (i.e., induced by the weighted column-sample defined by  $\mathbf{v}$ ). Low-rank approximations of  $\mathbf{K}_\nu$  can classically be obtained by spectral truncation, i.e., by considering a subset  $\mathbb{I}_{\nu, \text{trc}}^+$  of  $\mathbb{I}_\nu^+$  (the truncation subset usually corresponds to the largest eigenvalues of  $T_\nu$ ) and by defining  $\mathbf{K}_{\nu, \text{trc}} = \sum_{l \in \mathbb{I}_{\nu, \text{trc}}^+} \vartheta_l \mathbf{u}_l \mathbf{u}_l^T$ ; in practice, in view of section 3, one should in this case favor a truncation subset corresponding to accurately approximate eigendirections.

For  $\boldsymbol{\omega} = \mathbb{1}$ , the approximate eigenpairs  $\{\widehat{\lambda}_l^{[1]}, \widehat{\mathbf{v}}_l\}$  correspond to approximations of the eigenpairs of  $\mathbf{K}\mathbf{W} = \mathbf{K}$ . In this case, the matrix  $\mathbf{K}_{\nu, \text{trc}}$  approximates a low-row rank approximation of  $\mathbf{K}$  obtained by spectral truncation (i.e., obtained by truncating the spectrum of  $\mathbf{K}$ ; see, e.g., [8, 11]); following Remark 3.2, we can also notice that for  $\boldsymbol{\omega} = \mathbb{1}$ , we have  $\|K - K_\nu\|_{L^2(\mu \otimes \mu)}^2 = \|\mathbf{K} - \mathbf{K}_\nu\|_F^2$ .

**5. Optimal quadrature-sparsification as quadratic programming.** We consider the framework of section 4.3. From (4.3), for a fixed discrete measure  $\mu$  supported by  $\mathcal{S}$  (i.e.,  $\boldsymbol{\omega} > 0$  is fixed), we define, for  $\mathbf{v} \in \mathbb{R}^N$  (and in practice  $\mathbf{v} \geq 0$ ),

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

the scalar  $1/2$  being added for simplification purposes. To promote sparsity of the approximate measure and discard the trivial minimum at  $\mathbf{v} = \boldsymbol{\omega}$ , we now introduce squared-kernel-discrepancy minimization problems involving an  $\ell^1$ -type penalization.

Notice that we could as well consider the framework of section 4.2; in this case, the term  $\mathbf{S}\boldsymbol{\omega}$  has to be replaced by  $\mathbf{g}_\mu$  and  $\boldsymbol{\omega}^T \mathbf{S}\boldsymbol{\omega}$  by  $\|K\|_{L^2(\mu \otimes \mu)}^2$ . For simplicity, however, we do not discuss quadrature-sparsification problems involving a general initial measure  $\mu \in \mathcal{T}(K)$  in the remainder of this article.

**5.1. Regularized squared-kernel-discrepancy minimization.** For a given *penalization direction*  $\mathbf{d} = (d_1, \dots, d_N)^T \in \mathbb{R}^N$ , with  $\mathbf{d} > 0$  (see section 9 for a discussion on the choice of relevant penalization directions), and for  $\alpha \geq 0$ , we introduce the minimization problem, for  $\mathbf{v} \in \mathbb{R}^N$ ,

$$(5.1) \quad \underset{\mathbf{v}}{\text{minimize}} \quad D_\alpha(\mathbf{v}) = \frac{1}{2}(\boldsymbol{\omega} - \mathbf{v})^T \mathbf{S}(\boldsymbol{\omega} - \mathbf{v}) + \alpha \mathbf{d}^T \mathbf{v} \quad \text{subject to } \mathbf{v} \geq 0.$$

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

PROPOSITION 5.1. Denote by  $\mathbf{v}_\alpha^*$  a solution to (5.1) with  $\alpha \geq 0$ . We have

- (a) for  $\alpha = 0$ ,  $\mathbf{v}_\alpha^* = \boldsymbol{\omega}$  is a solution to (5.1),
- (b) if  $\alpha \geq \max_k \{[\mathbf{S}\boldsymbol{\omega}]_k/d_k\}$ , then  $\mathbf{v}_\alpha^* = 0$  (with  $[\mathbf{S}\boldsymbol{\omega}]_k$  the  $k$ th component of  $\mathbf{S}\boldsymbol{\omega}$ ),
- (c) for all  $\alpha \geq 0$ , we have  $0 \leq \alpha \mathbf{d}^T \mathbf{v}_\alpha^* \leq \alpha \mathbf{d}^T \boldsymbol{\omega} - (\boldsymbol{\omega} - \mathbf{v}_\alpha^*)^T \mathbf{S}(\boldsymbol{\omega} - \mathbf{v}_\alpha^*)$ ,
- (d)  $\nabla D_\alpha(\mathbf{v}_\alpha^*) \geq 0$  and  $(\mathbf{v}_\alpha^*)^T \nabla D_\alpha(\mathbf{v}_\alpha^*) = 0$ ,
- (e) if  $\tilde{\mathbf{v}}_\alpha^*$  is another solution to (5.1), then  $\mathbf{S}\tilde{\mathbf{v}}_\alpha^* = \mathbf{S}\mathbf{v}_\alpha^*$  and  $\mathbf{d}^T \tilde{\mathbf{v}}_\alpha^* = \mathbf{d}^T \mathbf{v}_\alpha^*$ ,
- (f) if  $[\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}]_k > 0$ , or if  $[\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}]_k = 0$  and  $\mathbf{S}_{k,k} > 0$  (see Remark 5.1), then  $[\mathbf{v}_\alpha^*]_k = 0$ ,
- (g) the maps  $\alpha \mapsto D(\mathbf{v}_\alpha^*)$  and  $\alpha \mapsto D_\alpha(\mathbf{v}_\alpha^*)$  are increasing,
- (h) the maps  $\alpha \mapsto \mathbf{d}^T \mathbf{v}_\alpha^*$ ,  $\alpha \mapsto (\mathbf{v}_\alpha^*)^T \mathbf{S}\mathbf{v}_\alpha^*$ , and  $\alpha \mapsto \boldsymbol{\omega}^T \mathbf{S}\mathbf{v}_\alpha^*$  are decreasing.

Remark 5.1. Assuming  $\mathbf{S}_{k,k} = K^2(x_k, x_k) > 0$  for all  $k \in \{1, \dots, N\}$  (what we shall denote by  $\text{diag}(\mathbf{S}) > 0$ ) is equivalent to assuming  $K(x_k, x_k) > 0$  for all  $k$ ; we recall that for all  $x \in \mathcal{X}$ , we have  $K(x, x) = \|K(x, \cdot)\|_{\mathcal{H}}^2 \geq 0$ . This assumption is thus nonrestrictive: indeed, if  $K(x_k, x_k) = 0$ , then  $h(x_k) = 0$  for all  $h \in \mathcal{H}$ ; if  $\mu$  and  $\nu$  are supported by  $\mathcal{S}$  (section 4.3), then such a point  $x_k$  can be removed from  $\mathcal{S}$  without inducing any modification of the operators  $T_\mu$  and  $T_\nu$ .

Since  $\mathbf{v} \geq 0$ , the term  $\mathbf{d}^T \mathbf{v}$  can be interpreted as a weighted  $\ell^1$ -type regularization, and  $\alpha$  as a regularization parameter. For appropriate  $\mathbf{d}$  and  $\alpha$ , we can therefore expect a solution  $\mathbf{v}_\alpha^*$  to (5.1) to be sparse, and sparsity of the solutions should tend to increase with  $\alpha$  (see, e.g., [14]). This intuition is confirmed by Proposition 5.1(f), which shows that the number of strictly positive components of  $\mathbf{v}_\alpha^*$  is bounded from above by the number of negative components of  $\alpha \mathbf{d} - \mathbf{S}\boldsymbol{\omega}$  (this bound is, however, generally not tight).

**5.2. Constrained squared-kernel-discrepancy minimization.** Instead of considering (5.1), for  $\varkappa \geq 0$  (and, in practice,  $\varkappa \leq \mathbf{d}^T \boldsymbol{\omega}$ ; see Proposition 5.2), we can equivalently introduce, for  $\mathbf{v} \in \mathbb{R}^N$ ,

$$(5.2) \quad \underset{\mathbf{v}}{\text{minimize}} \quad D(\mathbf{v}) = \frac{1}{2}(\boldsymbol{\omega} - \mathbf{v})^T \mathbf{S}(\boldsymbol{\omega} - \mathbf{v}) \quad \text{subject to } \mathbf{v} \geq 0 \text{ and } \mathbf{d}^T \mathbf{v} = \varkappa.$$

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

PROPOSITION 5.2. Let  $\mathbf{v}_\alpha^*$  be a solution to problem (5.1) with  $\alpha \geq 0$ ; then  $\mathbf{v}_\alpha^*$  is a solution to problem (5.2) with  $\varkappa = \mathbf{d}^T \mathbf{v}_\alpha^*$ . Reciprocally, assume that  $\mathbf{v}_\varkappa^*$  is a solution to problem (5.2) with  $0 < \varkappa \leq \mathbf{d}^T \boldsymbol{\omega}$ ; then  $\mathbf{v}_\varkappa^*$  is a solution to problem (5.1) with  $\alpha = (\mathbf{v}_\varkappa^*)^T \mathbf{S}(\boldsymbol{\omega} - \mathbf{v}_\varkappa^*)/\varkappa$ . For  $\varkappa = 0$ , we have  $\mathbf{v}_\varkappa^* = 0$ , which is the solution to problem (5.1) for  $\alpha \geq \max_k \{[\mathbf{S}\boldsymbol{\omega}]_k/d_k\}$ . For  $0 \leq \varkappa \leq \mathbf{d}^T \boldsymbol{\omega}$ , the maps  $\varkappa \mapsto D(\mathbf{v}_\varkappa^*)$  and  $\varkappa \mapsto (\mathbf{v}_\varkappa^*)^T \mathbf{S}(\boldsymbol{\omega} - \mathbf{v}_\varkappa^*)/\varkappa$  are decreasing.

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

$$(5.3) \quad \underset{\mathbf{v}}{\text{minimize}} \quad D(\mathbf{v}) = \frac{1}{2}(\boldsymbol{\omega} - \mathbf{v})^T \mathbf{S}(\boldsymbol{\omega} - \mathbf{v}) \quad \text{subject to } \mathbf{v} \geq 0 \text{ and } \mathbf{d}^T \mathbf{v} \leq \varkappa.$$

Problem (5.2) can be efficiently solved thanks to a sparse-descent-direction QP solver (and without storing the matrix  $\mathbf{S}$ ), like, for instance, the vertex-exchange strategy; see [18, Chap. 9] and section 8.1. A sequential strategy (based on the notion of regularization path) for solving problems (5.1) and (5.2) is discussed in section 7

**5.3. Penalization and conic squared-kernel discrepancy.** We now investigate the properties of the solutions to penalized squared-kernel-discrepancy minimization problems in light of Theorem 3.2 (i.e., in terms of conic squared-kernel discrepancy).

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

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

Theorem 5.1 thus gives a sufficient condition for the conic squared-kernel discrepancy of the solutions to the regularized problem (5.1) to increase with the regularization parameter  $\alpha$ ; in combination with Proposition 5.1(f), this result therefore shows that increasing the amount of penalization tends to increase the sparsity of the approximate measures (more precisely, this decreases the upper bound on the number of support points of the optimal approximate measures), at the expense of reducing the overall accuracy of the induced spectral approximations; see sections 10 and 11 for illustrations. This sufficient condition is further discussed in section 7.1; notice that it is, for instance, always verified when the matrix  $\mathbf{S}$  is nonsingular.

**6. Analogy with one-class SVM.** Following, for instance, [19], problems (5.1) and (5.2) can be interpreted as the dual formulations of *one-class distorted SVMs* defined from the squared kernel, the initial discrete measure  $\mu$ , and the penalization direction  $\mathbf{d}$ .

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

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

$$(6.1) \quad \begin{aligned} & \underset{g}{\text{minimize}} && \frac{1}{2} \|g\|_{\mathcal{G}}^2 + (g|g_\mu)_{\mathcal{G}} \\ & \text{subject to} && g(x_k) \geq -\alpha d_k \text{ for all } k \in \{1, \dots, N\}. \end{aligned}$$

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

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

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

$$(6.2) \quad \begin{aligned} & \underset{\check{g}}{\text{minimize}} && \frac{1}{2} \|\check{g}\|_{\mathcal{G}}^2 \\ & \text{subject to} && \check{g}(x_k) \geq g_\mu(x_k) - \alpha d_k \text{ for all } k \in \{1, \dots, N\}, \end{aligned}$$

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

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

$$(6.3) \quad \begin{aligned} & \underset{g, \gamma}{\text{minimize}} && \frac{1}{2} \|g\|_{\mathcal{G}}^2 + (g|g_\mu)g - \gamma \\ & \text{subject to} && g(x_k) \geq \gamma d_k / \varkappa \text{ for all } k \in \{1, \dots, N\}. \end{aligned}$$

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

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

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

*Remark 6.1.* Following the analogy with SVM models, we could also define soft-margin-type extensions of problems (6.1) and (6.3), i.e., we may consider models where the inequalities appearing in the constraints can potentially be violated, the level of violation being penalized. To be more precise, in (6.1), for instance, instead of considering the constraints  $g(x_k) \geq -\alpha d_k$ , we may consider the relaxed constraints  $g(x_k) \geq -\alpha d_k - \xi_k$ , with  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N)^T \in \mathbb{R}^N$ , while penalizing the values taken by the slack variables  $\xi_k$ ; the considered penalization corresponds to the choice of a *loss function*; see, for instance, [24]. Soft-margin extensions appear as a possible way to further constrain or penalize the optimal approximate measures.

**7. Regularization path.** In this section, we further discuss the properties of the solutions to problem (5.1); following, for instance, [17, 13], we also describe the regularization-path method (or homotopy method) for solving the regularized problem. Results related to the constrained problem (5.2) can be obtained from Proposition 5.2.

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

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

so that, from Proposition 5.1,  $[\nabla D_\alpha(\mathbf{v}_\alpha^*)]_k > 0$  for all  $k \in J_\alpha^c$ ; in addition, the index set  $J_\alpha$  is unique (i.e., for a given  $\alpha$ , in case of nonuniqueness of the solution to (5.1),  $J_\alpha$  does not depend on the solution considered). We shall refer to  $J_\alpha$  as the *sparsity pattern* of the solutions to problem (5.1) for  $\alpha \geq 0$ . From Proposition 5.1(d), if  $[\mathbf{v}_\alpha^*]_k > 0$ , then  $k$  necessarily belongs to  $J_\alpha$ ; in addition, the solutions to (5.1) are characterized by the conditions

$$(7.1) \quad \mathbf{v}_\alpha^* \geq 0, \text{ and } [\mathbf{v}_\alpha^*]_{J_\alpha^c} = 0, \text{ and } \mathbf{S}_{J_\alpha, J_\alpha} [\mathbf{v}_\alpha^*]_{J_\alpha} = [\mathbf{S}\boldsymbol{\omega} - \alpha \mathbf{d}]_{J_\alpha},$$

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

**PROPOSITION 7.1.** *Let  $\mathbf{v}_{\alpha_1}^*$  and  $\mathbf{v}_{\alpha_2}^*$  be solutions to problem (5.1) with  $\alpha_1$  and  $\alpha_2 \geq 0$ , respectively. Assume that  $J_{\alpha_1} = J_{\alpha_2} = J$ ; then for all  $\theta \in [0, 1]$ ,  $\mathbf{v}_\alpha^* = \theta \mathbf{v}_{\alpha_1}^* + (1 - \theta) \mathbf{v}_{\alpha_2}^*$  is a solution to problem (5.1) with  $\alpha = \theta \alpha_1 + (1 - \theta) \alpha_2$ , and  $J_\alpha = J$ .*



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

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

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

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

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

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

**7.2. Computing the path.** The regularization-path method consists in iteratively computing the kinks  $\alpha_0, \alpha_1$ , etc., while keeping track of the evolution of the sparsity pattern of the solutions to (5.1).

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

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

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

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

$$\alpha_+ = \max_l \{ [\mathbf{M}[\mathbf{S}\boldsymbol{\omega}]_J - [\mathbf{S}\boldsymbol{\omega}]_{J^c}]_l / [\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}]_l \mid [\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}]_l < 0 \} \text{ and}$$

$$\alpha_- = \max_m \{ [(\mathbf{S}_{J,J})^{-1}[\mathbf{S}\boldsymbol{\omega}]_J]_m / [(\mathbf{S}_{J,J})^{-1}\mathbf{d}_J]_m \mid [(\mathbf{S}_{J,J})^{-1}\mathbf{d}_J]_m < 0 \}.$$

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

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

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

In view of Lemma 7.1, once a kink  $\alpha_p$  and its left sparsity pattern  $J = L_{\alpha_p}$  are known, defining the next event (i.e., computing  $\alpha_{p+1}$  and  $L_{\alpha_{p+1}}$ ) involves the calculation of  $(\mathbf{S}_{J,J})^{-1}\mathbf{d}_J$  and  $(\mathbf{S}_{J,J})^{-1}[\mathbf{S}\boldsymbol{\omega}]_J$  (i.e., solving a linear system); without taking into account the computations already performed to obtain the information relative to the kink  $\alpha_p$ , and using a direct method (by, for instance, considering the Cholesky decomposition of the symmetric and positive-definite matrix  $\mathbf{S}_{J,J}$ ), the computational complexity of this task scales as  $\mathcal{O}(n_{\alpha_p}^3)$ , with  $n_{\alpha_p} = \text{card}(L_{\alpha_p})$ . Update formulae can nevertheless be used to reduce this complexity by, for instance, iteratively updating the Cholesky decomposition of  $\mathbf{S}_{J,J}$ . In the favorable cases, the computational complexity may thus reduce to  $\mathcal{O}(n_{\alpha_p}^2)$ ; an alternative might also consist in using an indirect iterative approach, like, for instance, a conjugate gradient method (but numerical errors could then quickly lead to precision issues). Finally, the complexity of the two matrix-vector products involving the matrix  $\mathbf{S}_{J^c,J}$  scales as  $\mathcal{O}(n_{\alpha_p}(N - n_{\alpha_p}))$ . As a result, the computation of the regularization path becomes intractable once large values of  $n_{\alpha_p}$  are reached. When  $N$  is large, the regularization-path method may therefore only be used to explore the range of very sparse approximate measures. See sections 10 and 11 for illustrations.

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

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

$$(8.1) \quad \underset{\tilde{\mathbf{v}}}{\text{minimize}} \quad C(\tilde{\mathbf{v}}) = \frac{1}{2}\tilde{\mathbf{v}}^T \mathbf{A}\tilde{\mathbf{v}} - \mathbf{b}^T \tilde{\mathbf{v}} \quad \text{subject to} \quad \tilde{\mathbf{v}} \geq 0 \quad \text{and} \quad \mathbf{1}^T \tilde{\mathbf{v}} = 1,$$

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

The extreme points of the polytopes defined by the constraints in (8.1) are the vectors  $\{\mathbf{e}_i\}_{i=1}^N$ , where  $\mathbf{e}_i \in \mathbb{R}^N$  is the  $i$ th element of the canonical basis of  $\mathbb{R}^N$  (that is,  $[\mathbf{e}_i]_i = 1$ , all the other components being 0). For a feasible  $\tilde{\mathbf{v}}$ , let  $I_{\tilde{\mathbf{v}}} = \{k | \tilde{v}_k > 0\}$  be the index set defined by the strictly positive components of  $\tilde{\mathbf{v}}$ . An iteration of the vertex-exchange algorithm consists in searching

$$i^* = \underset{i}{\text{argmin}} [\nabla C(\tilde{\mathbf{v}})]_i \quad \text{and} \quad j^* = \underset{j \in I_{\tilde{\mathbf{v}}}}{\text{argmax}} [\nabla C(\tilde{\mathbf{v}})]_j,$$

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  $\tilde{\mathbf{v}}$ ); in case of nonuniqueness of the extrema, an index is simply selected at random from among the ones satisfying the condition. The step size is then classically obtained by line search, the optimal step size  $\varrho$  being given by  $\varrho = \min \{\tilde{v}_{j^*}, -(\boldsymbol{\delta}^T \nabla C(\tilde{\mathbf{v}}))/(\boldsymbol{\delta}^T \mathbf{A}\boldsymbol{\delta})\}$ . Since the descent direction  $\boldsymbol{\delta}$  is sparse, the computation of the optimal step size is numerically affordable, and the same holds for the gradient update. Indeed, we have  $\nabla C(\tilde{\mathbf{v}} + \varrho \boldsymbol{\delta}) = \nabla C(\tilde{\mathbf{v}}) + \varrho \mathbf{A}\boldsymbol{\delta}$ , so that the gradient update only involves two columns of  $\mathbf{A}$ . The complexity of an iteration thus scales as  $\mathcal{O}(N)$ .

Denoting by  $\tilde{\mathbf{v}}^*$  a solution to (8.1), the convergence of the vertex-exchange algorithm can be easily verified (see, e.g., [12]) by simply noting that since  $\tilde{\mathbf{v}} \geq 0$  and  $\mathbf{1}^T \tilde{\mathbf{v}} = 1$ , by definition of  $j^*$ , we have  $\tilde{\mathbf{v}}^T \nabla C(\tilde{\mathbf{v}}) \leq \mathbf{e}_{j^*}^T \nabla C(\tilde{\mathbf{v}})$ , and thus (distance from optimality)

$$C(\tilde{\mathbf{v}}) - C(\tilde{\mathbf{v}}^*) \leq -(\mathbf{e}_{i^*} - \tilde{\mathbf{v}})^T \nabla C(\tilde{\mathbf{v}}) \leq -(\mathbf{e}_{i^*} - \mathbf{e}_{j^*})^T \nabla C(\tilde{\mathbf{v}}).$$

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

**8.2. Enhancing sparsity through components merging.** The formulation introduced in section 8.1 offers a convenient framework to enhance the sparsity of an approximate measure  $\nu$  while trying to keep its squared-kernel discrepancy as low as possible. Let  $\tilde{\mathbf{v}} \geq 0$  (with  $\tilde{\mathbf{v}} \in \mathbb{R}^N$ ) be such that  $\mathbf{1}^T \tilde{\mathbf{v}} = 1$ . In practice,  $\tilde{\mathbf{v}}$  will be an exact or approximate solution to problem (8.1), or any vector related to an interesting low-squared-kernel-discrepancy configuration  $\mathbf{v}$  through the change of variable  $\tilde{\mathbf{v}} = \mathbf{D}\mathbf{v}/\varkappa$ , with  $\mathbf{D} = \text{diag}(\mathbf{d})$  and  $\varkappa = \mathbf{d}^T \mathbf{v}$ ; see section 8.1. We assume that  $\tilde{\mathbf{v}}$  has  $n = n_0$  strictly positive components and we introduce  $I = \{i | \tilde{v}_i > 0\}$ . As illustrated in sections 10 and 11, it is generally possible, to a certain extent, to merge together some components of  $\tilde{\mathbf{v}}$  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  $\tilde{\mathbf{v}}$ .

We assume that  $n > 1$ . For an ordered pair  $\{i, j\}$ , with  $i$  and  $j \in I$  and  $i \neq j$ , we define  $\tilde{\mathbf{v}}_{\{i,j\}} = \tilde{\mathbf{v}} + \tilde{v}_j(\mathbf{e}_i - \mathbf{e}_j)$ . The vector  $\tilde{\mathbf{v}}_{\{i,j\}}$  thus has  $n - 1$  strictly positive components, the  $i$ th component of  $\tilde{\mathbf{v}}$  having absorbed the  $j$ th. We have

$$C(\tilde{\mathbf{v}}_{\{i,j\}}) = C(\tilde{\mathbf{v}}) + \frac{1}{2}\tilde{v}_j^2(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{A}(\mathbf{e}_i - \mathbf{e}_j) + \tilde{v}_j(\mathbf{e}_i - \mathbf{e}_j)^T \nabla C(\tilde{\mathbf{v}}).$$

Thus, knowing  $\nabla C(\tilde{\mathbf{v}})$ , the computation  $C(\tilde{\mathbf{v}}_{\{i,j\}})$  only involves four entries of the matrix  $\mathbf{A}$  and two entries of  $\nabla C(\tilde{\mathbf{v}})$ .

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

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

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

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

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

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

Thus, for  $\mathbf{d} \propto \mathbf{S}\boldsymbol{\omega}$ , the solutions to (5.1) are nonsparse, and such a choice for  $\mathbf{d}$  is of no practical interest; in order to promote sparsity through penalized squared-kernel-discrepancy minimization, one therefore has to check that the considered penalization direction does not correspond to this pathological case. More generally (and as a proof for Lemma 9.1), we can remark that if  $\mathbf{d} = \mathbf{S}\boldsymbol{\eta} \geq 0$ , with  $\boldsymbol{\eta} \in \mathbb{R}^N$ , then for all  $\alpha$  such that  $\boldsymbol{\omega} - \alpha\boldsymbol{\eta} \geq 0$ , we have  $\nabla D_\alpha(\boldsymbol{\omega} - \alpha\boldsymbol{\eta}) = 0$ , and  $\mathbf{v}_\alpha^* = \boldsymbol{\omega} - \alpha\boldsymbol{\eta}$  is in this case a solution to (5.1); notice that in the framework of section 7.1, this situation corresponds to solutions with full sparsity pattern, i.e.,  $J_\alpha = \{1, \dots, N\}$ .

In the examples presented in sections 10 and 11, considering  $\mathbf{d} = \mathbb{1}$  leads to satisfactory results (notice that  $\nu(\mathcal{X}) = \mathbb{1}^T \mathbf{v}$ ); it is nevertheless possible to define problem-dependent penalization directions, leading to models inheriting interesting interpretations. Following Remark 5.1, we recall that we can reasonably assume that

$\text{diag}(\mathbf{K}) > 0$ , so that, in particular,  $\mathbf{S}\boldsymbol{\omega} > 0$  (since  $\boldsymbol{\omega} > 0$ ). In Remarks 9.1, 9.2, and 9.3, we discuss specific penalization directions defined from the vectors  $\mathbf{S}\boldsymbol{\omega}$  and  $\text{diag}(\mathbf{K})$ .

*Remark 9.1* (penalizing the trace). For  $\mathbf{d} = \text{diag}(\mathbf{K})$ , we have  $\mathbf{d}^T \mathbf{v} = \text{trace}(T_\nu)$ ; by analogy with spectral truncation, from Proposition 5.1(c) and Theorem 5.1, a solution  $\mathbf{v}_\alpha^*$  to the regularized problem (5.1) then satisfies, for  $0 \leq \alpha < \alpha_0$ ,

$$\text{trace}(T_{\nu_\alpha^*}) \leq \text{trace}(T_{c_\alpha \nu_\alpha^*}) \leq \text{trace}(T_\mu).$$

Also notice that the parameter  $\varkappa$  of the constrained problem (5.2) corresponds in this case to the trace of the approximate operator.

Following section 6, if  $\mathbf{v}_\alpha^*$  is a solution to the regularized problem (5.1) (with related measure  $\nu_\alpha^*$ ), then  $g_{\nu_\alpha^*} = \sum_{k=1}^N [\mathbf{v}_\alpha^*]_k K_{x_k}^2$  is the solution to, for  $g \in \mathcal{G}$ ,

$$(9.1) \quad \text{minimize } \frac{1}{2} \|g\|_{\mathcal{G}}^2 \text{ subject to } g_\mu(x_k) - g(x_k) \leq \alpha d_k \text{ for all } k \in \{1, \dots, N\}$$

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

*Remark 9.2* (inverse-distortion-based penalization). In view of (9.1), considering a penalization direction  $\mathbf{d}$  such that  $d_k = 1/[\mathbf{S}\boldsymbol{\omega}]_k^p$ , with  $p > 0$ , results in an SVM where the upper bound on  $g_\mu(x_k) - g(x_k)$  is inversely proportional to a positive power of  $g_\mu(x_k)$ , so that the larger  $g_\mu(x_k)$  is, the smaller the bound on  $g_\mu(x_k) - g(x_k)$ . Since the most constrained inequalities in (9.1) are more likely to be active, and since we have  $(g_\mu | g_\nu)_\mathcal{G} = \sum_{k=1}^N v_k g_\mu(x_k)$  for any measure  $\nu$  supported by  $\mathcal{S}$  and with related weights  $v_k$  (see Lemma 2.1), such a penalization tends to promote large values of the inner product between  $g_\mu$  and  $g_{\nu_\alpha^*}$  in  $\mathcal{G}$ .

*Remark 9.3* (inverse-kernel-diagonal-based penalization). For all  $x \in \mathcal{X}$ , from the reproducing property in  $\mathcal{G}$  and the Cauchy–Schwarz inequality, we have

$$(9.2) \quad \text{for all } \mu \text{ and } \nu \in \mathcal{T}(K), \quad |g_\mu(x) - g_\nu(x)| \leq \sqrt{D_{K^2}(\mu, \nu)} K(x, x).$$

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

**10. Two-dimensional example.** We assume that  $\mathcal{S} = \{x_k\}_{k=1}^N$  consists of the  $N = 2016$  first points of a uniform Halton sequence on  $[-1, 1]^2$  (see [16]), as illustrated in Figure 10.1. We set  $\boldsymbol{\omega} = \mathbb{1}/N$ , so that the measure  $\mu = \sum_k \omega_k \delta_{x_k}$  appears as a quadrature approximation of the uniform probability measure on  $[-1, 1]^2$ . We consider the Gaussian kernel  $K(x, y) = \exp(-\ell \|x - y\|^2)$ , where  $\|x - y\|$  is the Euclidean norm on  $\mathbb{R}^2$ , and we set  $\ell = 1/0.16$  (a different kernel is considered in section 10.6). An overview of the spectrum of the operator  $T_\mu$  (obtained from the eigendecomposition of the matrix  $\mathbf{K}/N$ ) is given in Figure 10.2. We first consider the penalization direction  $\mathbf{d} = \mathbb{1}$ .

**10.1. First experiment.** Figure 10.1 shows the (approximate) solution  $\mathbf{v}^*$  to problem (5.2) with  $\varkappa = 0.81$  or, equivalently, to problem (5.1) with  $\alpha \approx 8.354214 \times 10^{-3}$  (for  $\boldsymbol{\omega} = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ ). The vector  $\mathbf{v}^*$  has 160 strictly positive components, and the support of the related measure  $\nu^*$  inherits an interesting “four-concentric-squares” structure. We have  $D(\mathbf{v}^*) = 7.631887 \times 10^{-4}$  (for comparison, notice that

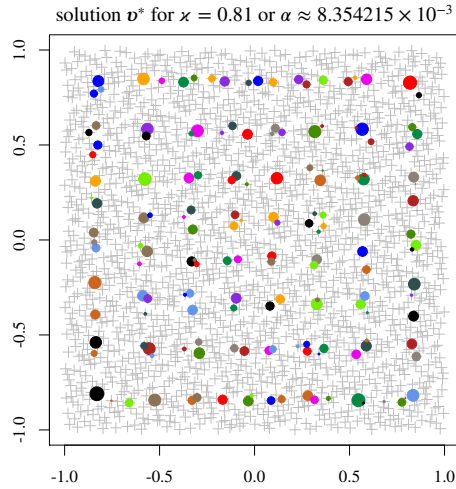


FIG. 10.1. Graphical representation (two-dimensional example, Gaussian kernel,  $\omega = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ ) of the solution  $\mathbf{v}^*$  to problem (5.2) with  $\varkappa = 0.81$  or, equivalently, to problem (5.1) with  $\alpha \approx 8.354215 \times 10^{-3}$ . The gray crosses represent the points in  $\mathcal{S}$  and the filled dots are the strictly positive components of  $\mathbf{v}^*$  (surface being proportional to  $v_k^*$ ).

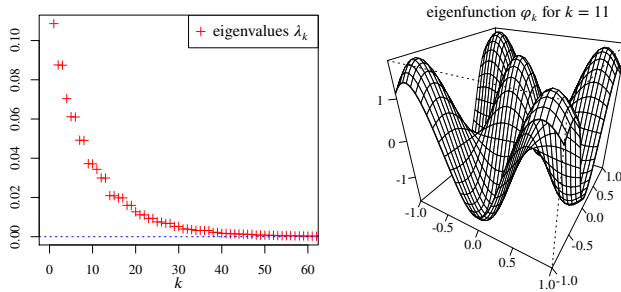


FIG. 10.2. For the two-dimensional example (Gaussian kernel and  $\omega = \mathbb{1}/N$ ), eigenvalues  $\lambda_k$  of the integral operator  $T_\mu$  (sorted in decreasing order; only the 62 largest eigenvalues are presented), and graph, on  $[-1, 1]^2$  of the canonically extended eigenfunction  $\varphi_k$  for  $k = 11$ .

$D(\varkappa \mathbf{e}_1) = 3.041066 \times 10^{-1}$ , with  $\mathbf{e}_1$  the first element of the canonical basis of  $\mathbb{R}^N$ . In the framework of section 8.1, the presented solution is related to a Frank–Wolfe error bound  $\epsilon = 3.989864 \times 10^{-17}$ .

The solution has been obtained using the regularization-path strategy (see section 10.2 for more details). Considering the regularization path for problem (5.1) with decreasing values of  $\alpha$ , the underlying value of  $\alpha \approx 8.354215 \times 10^{-3}$  satisfies

$$\alpha_{p+1} = 8.352970 \times 10^{-3} \leq \alpha \leq \alpha_p = 8.355244 \times 10^{-3} \text{ with } p = 4047;$$

correspondingly, for problem (5.2) (with increasing values of  $\varkappa$ ), the value  $\varkappa = 0.81$  satisfies

$$\varkappa_p = 0.8099788 \leq \varkappa \leq \varkappa_{p+1} = 0.8100256.$$

The accuracy of the approximate eigendecomposition of  $T_\mu$  induced by the solution  $\mathbf{v}^*$  presented in Figure 10.1 (i.e.,  $\varkappa = 0.81$ ) is illustrated in Figure 10.3. In view of the similarity between the geometric approximate eigenvalues  $\widehat{\lambda}_l^{[1]}$ , and

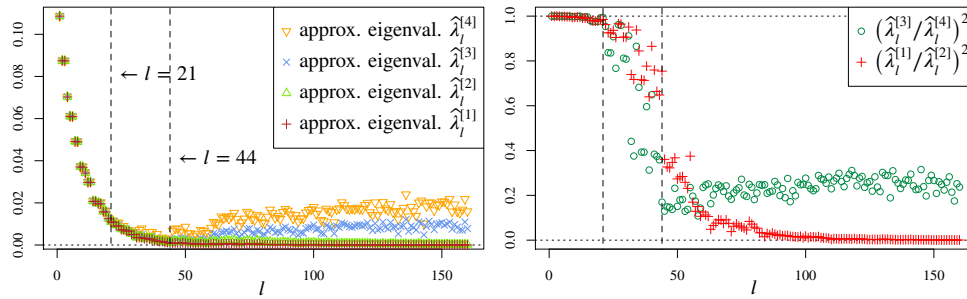


FIG. 10.3. Approximate eigenvalues  $\widehat{\lambda}_l^{[1]}$ ,  $\widehat{\lambda}_l^{[2]}$ ,  $\widehat{\lambda}_l^{[3]}$ , and  $\widehat{\lambda}_l^{[4]}$  induced by the solution  $\mathbf{v}^*$  presented in Figure 10.1 (left); ratios  $(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]})^2$  and  $(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]})^2$  highlighting the accuracy of the approximate eigendirections  $\psi_l$  of  $T_\mu$  (right).

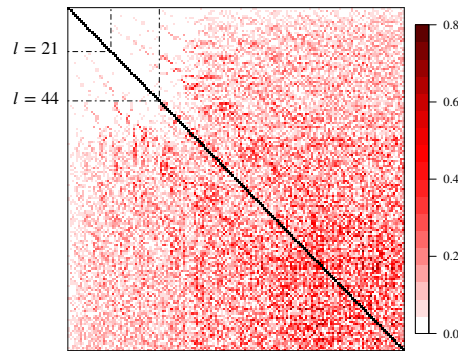


FIG. 10.4. Graphical representation of the matrix with  $l, l'$  entry  $|(\widehat{\varphi}_l | \widehat{\varphi}_{l'})_{L^2(\mu)}|$  for the 160 normalized approximate eigendirections induced by the solution  $\mathbf{v}^*$  presented in Figure 10.1 (i.e.,  $\varkappa = 0.81$ ).

more particularly of the ratios  $(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]})^2$  and  $(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]})^2$  (see section 3), we observe that the 21 main eigendirections of the operator  $T_{\nu^*}$  (i.e., for  $l \in \{1, \dots, 21\}$ ) lead to remarkably accurate approximations of the eigenpairs of  $T_\mu$  related to the 21 largest eigenvalues  $\lambda_k$ . The accuracy of the approximate eigenpairs decreases for  $l \in \{22, \dots, 44\}$  and becomes very poor for  $k > 44$ . The orthogonality, in  $L^2(\mu)$ , between the normalized approximate eigenfunctions  $\widehat{\varphi}_l$  is in perfect agreement with this observation, as illustrated in Figure 10.4 (see Remark 3.1).

A comparison between the true eigenvalues of  $T_\mu$  and their approximations induced by the solution  $\mathbf{v}^*$  of Figure 10.1 is presented in Figure 10.5; we, for instance, observe that for  $1 \leq l \leq 8$ , the approximate eigenvalues  $\widehat{\lambda}_l^{[4]}$  are the most accurate. Table 10.1 gives the errors  $\|\widehat{\varphi}_l - \varphi_l\|_{L^2(\mu)}^2$  for  $1 \leq l \leq 20$ ; in accordance with our previous conclusions, these approximations are remarkably accurate. Since orthonormalized sets of eigenfunctions are not unique, to perform this comparison, notice that we have when required replaced  $\widehat{\varphi}_l$  by  $-\widehat{\varphi}_l$  and applied a two-dimensional rotation to pairs of eigendirections related to the approximation of an eigensubspace of dimension two (corresponding to the case where the operator is defined with respect to a uniform measure on  $[-1, 1]^2$ ).

Following Theorem 5.1, we denote by  $c_\varkappa$  and  $c_\alpha$  the argument of the minimum of the functions  $c \mapsto D(c\mathbf{v}_\varkappa^*)$  and  $c \mapsto D(c\mathbf{v}_\alpha^*)$ . For the solution presented in (5.2)

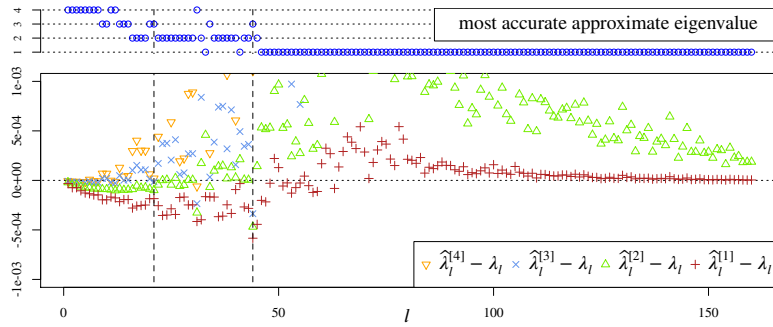


FIG. 10.5. Errors  $\hat{\lambda}_l^{[i]} - \lambda_l$  for the geometric approximate eigenvalues induced by the solution  $\mathbf{v}^*$  presented in Figure 10.1 (bottom), and indication of the most accurate (smallest absolute error) approximation among  $\hat{\lambda}_l^{[1]}$ ,  $\hat{\lambda}_l^{[2]}$ ,  $\hat{\lambda}_l^{[3]}$ , and  $\hat{\lambda}_l^{[4]}$  (top).

TABLE 10.1

Approximation error  $\|\hat{\varphi}_l - \varphi_l\|_{L^2(\mu)}^2$ , with  $1 \leq l \leq 20$ , for the normalized approximate eigendirections induced by the solution  $\mathbf{v}^*$  presented in Figure 10.1 (i.e.,  $\varkappa = 0.81$ ); the values of  $l$  grouped together correspond to pairs of eigendirections related to the approximation of an eigensubspace of dimension two.

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

(i.e.,  $\varkappa = 0.81$ ), we obtain  $c_\varkappa = 1.177289$ , and  $D(c_\varkappa \mathbf{v}_\varkappa^*) = 1.633391 \times 10^{-4}$ , and  $c_\varkappa \mathbf{d}^T \mathbf{v}_\varkappa^* = 0.9536041$ .

**10.2. Regularization path.** Following section 7, we compute the 12,818 first events of the regularization path related to problem (5.1) with decreasing values of  $\alpha$ ; we have in particular  $\alpha_0 = 6.310163 \times 10^{-2}$  and  $\alpha_{12817} = 1.495359 \times 10^{-5}$ . Correspondingly, for problem (5.2) and increasing  $\varkappa$ , we have  $\varkappa_0 = 0$  and  $\varkappa_{12817} = 0.9995482$  (we recall that  $\mathbf{d}^T \boldsymbol{\omega} = 1$ ).

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

For 51 values of  $\varkappa$  evenly spread between  $\varkappa_0$  and  $\varkappa_{12817}$ , Figure 10.7 shows the evolution of the ratio  $(\hat{\lambda}_l^{[1]} / \hat{\lambda}_l^{[2]})^2$  for the approximate eigendecompositions induced by the various solutions  $\mathbf{v}_\varkappa^*$ . As expected, the number of accurately approximate eigendirections increases with  $\varkappa$ . Remarkably, the number of eigendirections approximated with high accuracy appears to be in close relation with the decay of the spectrum of  $T_\mu$ ; we recall that we have  $\text{trace}(T_{\nu_\varkappa}) = \varkappa$ , since  $\text{diag}(\mathbf{K}) = \mathbb{1}$  for the Gaussian kernel.

**10.3. Components merging.** We now perform the strong-pairwise merging (see section 8.2) of the solution  $\mathbf{v}^*$  presented in Figure 10.1 (i.e., problem (5.2) with



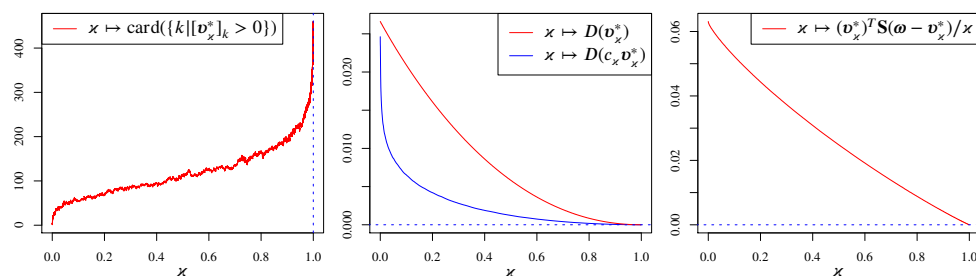


FIG. 10.6. For the two-dimensional example (Gaussian kernel,  $\omega = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ ), graphical representation of the 12,818 first events of the regularization path related to problem (5.2) for increasing  $\varkappa$ ; number of strictly positive components of  $\mathbf{v}_\varkappa^*$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

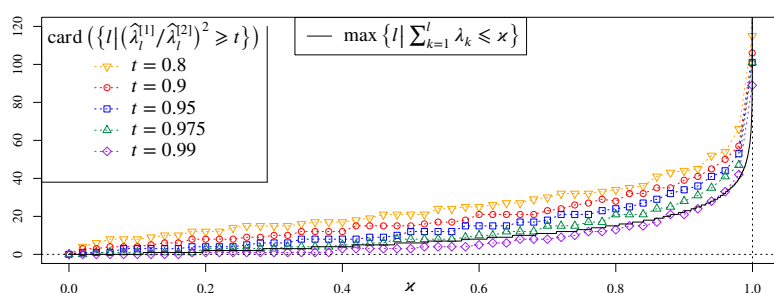


FIG. 10.7. Evolution of the accuracy of the approximate eigendecomposition of  $T_\mu$  induced by  $\mathbf{v}_\varkappa^*$  for 51 values of  $\varkappa$  between  $\varkappa_0 = 0$  and  $\varkappa_{12817} = 0.9995482$ ; the accuracy of the approximate eigendirections is measured through the ratios  $(\widehat{\lambda}_i^{[1]}/\widehat{\lambda}_i^{[2]})^2$ ; for illustration purposes, the map  $\varkappa \mapsto \max\{t \mid \sum_{k=1}^t \lambda_k \leq \varkappa\}$  is also presented (two-dimensional example, Gaussian kernel,  $\omega = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ ).

$\varkappa = 0.81$ ). As illustrated in Figure 10.8, for the first merging iterations,  $D(\mathbf{v}_{[k]})$  stays very close to  $D(\mathbf{v}^*) = 7.631887 \times 10^{-4}$ . After 90 iterations, we have  $D(\mathbf{v}_{[90]}) - D(\mathbf{v}^*) = 3.494809 \times 10^{-5}$  (i.e., an increase of 4.58%), and  $\mathbf{v}_{[90]}$  is supported by 70 points (instead of 160 for  $\mathbf{v}^*$ ); a graphical representation of  $\mathbf{v}_{[90]}$  is given in the left-hand part of the figure. The accuracy of the approximate eigendecomposition induced by  $\mathbf{v}_{[90]}$  is presented in the right-hand part of Figure 10.8. We observe that although being slightly less accurate than the approximate eigendecomposition induced by  $\mathbf{v}^*$ , the approximation induced by  $\mathbf{v}_{[90]}$  remains very satisfactory while being related to a vector more than two times sparser. Notice that the conic squared-kernel discrepancy of the merged solution is  $D(c_\diamond \mathbf{v}_{[90]}) = 2.091099 \times 10^{-4}$ , where  $c_\diamond$  stands for the optimal rescaling parameter  $c$  related to  $\mathbf{v}_{[90]}$ ; see Theorem 3.2.

**10.4. Comparison with random sampling.** We compute the approximate eigendecompositions induced by random uniform samples, without replacement, of size  $n_{rand} = 300, 600, 900,$  and  $1200$  (i.e., we randomly select  $n_{rand}$  distinct points among the  $N = 2016$  points in  $\mathcal{S}$ , and we consider the uniform probability measure supported by the points selected); for each sample size, we perform 100 repetitions. Figure 10.9 illustrates the accuracy of the obtained approximate eigendirections, measured through the ratios  $(\widehat{\lambda}_i^{[1]}/\widehat{\lambda}_i^{[2]})^2$ . As we could expect, the accuracy of the approximation increases with the size of the sample. In terms of trade-off between sparsity

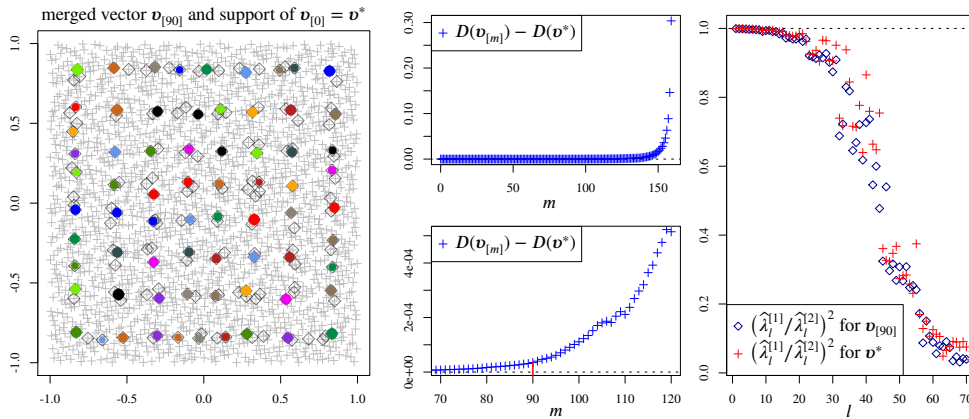


FIG. 10.8. Graphical representation of the merged solution  $\mathbf{v}_{[90]}$  (two-dimensional example with  $\omega = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ ) obtained after 90 iterations of the strong-pairwise-merging strategy applied to the solution  $\mathbf{v}^*$  presented in Figure 10.1; the gray diamonds indicate the support of  $\mathbf{v}^*$  (left). Increase of the cost  $D(\cdot)$  induced by each merging iteration, for the whole 159 iterations (top middle), and zoom around the 90th iteration (bottom middle). Representation of the ratios  $(\hat{\lambda}_i^{[1]}/\hat{\lambda}_i^{[2]})^2$  obtained from the merged vector  $\mathbf{v}_{[90]}$  and comparison with the same ratios for the solution  $\mathbf{v}^*$  (right).

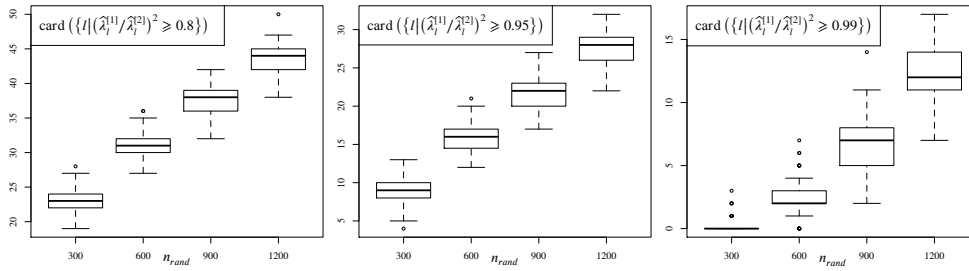


FIG. 10.9. For the two-dimensional example, accuracy of the approximate eigendecompositions induced by random samples of size  $n_{rand}$  (without replacement); for each values of  $n_{rand}$ , Tukey's boxplot, over 100 repetitions, of the number of approximate eigendirections such that  $(\hat{\lambda}_i^{[1]}/\hat{\lambda}_i^{[2]})^2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

and number of eigendirections accurately approximated, the results are, however, far behind the ones obtained using penalized squared-kernel-discrepancy minimization (see Figures 10.6 and 10.7). For instance, and in comparison to Figure 10.9, penalized squared-kernel-discrepancy minimization leads to the following trade-offs:

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

**10.5. Impact of the penalization direction.** For the two-dimensional example (Gaussian kernel and  $\omega = \mathbb{1}/N$ ), we compute the regularization path of problem (5.1) for seven different vectors  $\mathbf{d} > 0$ . We consider  $\mathbf{d} = \mathbf{v}_{\max}(\mathbf{S})$  (i.e., the eigenvector

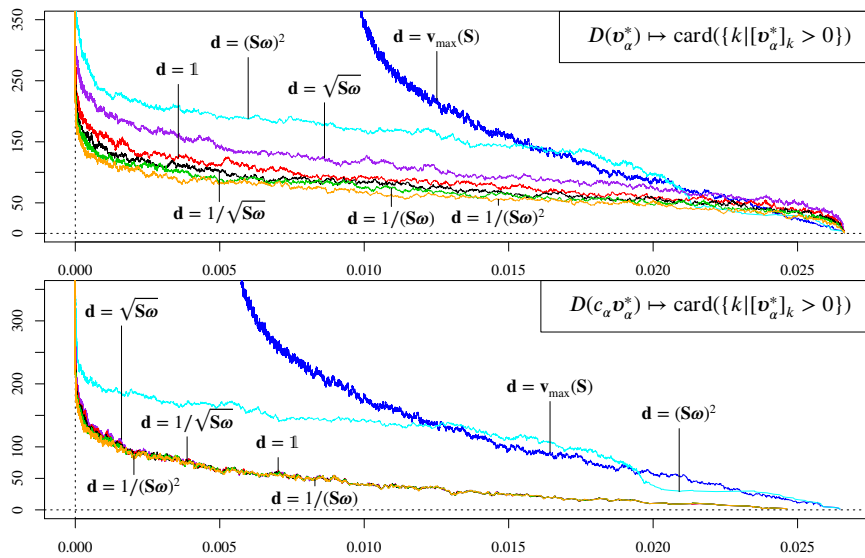


FIG. 10.10. For the two-dimensional example (Gaussian kernel and  $\boldsymbol{\omega} = 1/N$ ), number of strictly positive components of the solution  $\mathbf{v}_\alpha^*$  to problem (5.1) as a function of the squared-kernel discrepancy  $D(\mathbf{v}_\alpha^*)$  (top) and of the conic squared-kernel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

related to the largest eigenvalue of the matrix  $\mathbf{S}$ ; see the Perron–Frobenius theorem),  $(\mathbf{S}\boldsymbol{\omega})^2$  (i.e.,  $d_k = [\mathbf{S}\boldsymbol{\omega}]_k^2$ ),  $\sqrt{\mathbf{S}\boldsymbol{\omega}}$  (i.e.,  $d_k = \sqrt{[\mathbf{S}\boldsymbol{\omega}]_k}$ ),  $\mathbb{1}$ ,  $1/\sqrt{\mathbf{S}\boldsymbol{\omega}}$ ,  $1/(\mathbf{S}\boldsymbol{\omega})$ , and  $1/(\mathbf{S}\boldsymbol{\omega})^2$ . In Figure 10.10, we compare the trade-offs between sparsity and (raw and conic) squared-kernel discrepancy yield by these penalization directions. We recall that for the Gaussian kernel, we have  $\text{diag}(\mathbf{K}) = \mathbb{1}$ .

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

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

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

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

**11. Application to medium/large-scale problems.** This section aims at illustrating the ability of the proposed framework to tackle relatively large-scale

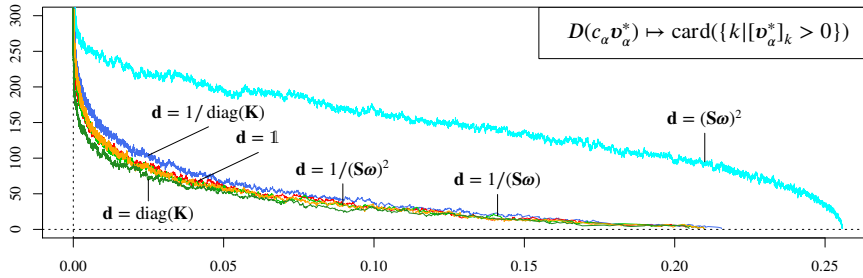


FIG. 10.11. For the two-dimensional example (modified Gaussian kernel (10.1) and  $\omega = \mathbb{1}/N$ ), number of strictly positive components of the solution  $\mathbf{v}_\alpha^*$  to problem (5.1) as a function of the conic squared-kernel discrepancy  $D(c_\alpha \mathbf{v}_\alpha^*)$ , for various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

problems. The datasets have been obtained from the UCI Machine Learning Repository; see [7]. All the computations have been performed on a 2015 desktop with an Intel Core i7-4790 processor with 16 GB of RAM; the various methods have been entirely implemented in C.

**11.1. MiniBooNE dataset.** We consider the standardized 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 (see section 10) with  $\ell = 0.02$ , and we set  $\omega = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$  (notice that  $\ell = 0.02$  belongs to the range of “good parameter values” for the SVM binary classification of this dataset).

We compute the 3000 first events of the regularization path related to problems (5.1) and (5.2). We have  $\alpha_0 = 0.2188961$  and  $\alpha_{2999} = 3.546703 \times 10^{-3}$ , and correspondingly  $\varkappa_0 = 0$  and  $\varkappa_{2999} = 0.655808$  (notice that  $\mathbf{d}^T \omega = 1$ ); a graphical representation of the properties of these solutions is proposed in Figure 11.1. We can observe that for  $\varkappa \geq 0.5$ , the number of strictly positive components of  $\mathbf{v}_\varkappa^*$  increases quickly with  $\varkappa$ ; the computation of the regularization path then becomes intractable (notice that the calculation of the 3000 first events of the regularization path took around 3 hours on our aforementioned 2015 desktop).

From the regulation path, we build the solutions to problem (5.2) for  $\varkappa = 0.3$  and  $\varkappa = 0.655$  (i.e., for problem (5.1),  $\alpha \approx 4.400276 \times 10^{-2}$  and  $\alpha \approx 3.571413 \times 10^{-3}$ ); these solutions have 76 and 1902 strictly positive components, respectively. The efficiency of the induced approximate eigendecompositions is illustrated in Figure 11.2. For  $\varkappa = 0.3$ , we obtain a relatively accurate approximation of the three main eigenpairs of  $T_\mu$  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  $T_\mu$  are approximate with high accuracy (i.e.,  $1 \leq l \leq 8$ ), and the approximations remains relatively accurate until  $l = 29$ . Interestingly, we observe that contrary to the ratios  $(\widehat{\lambda}_l^{[3]}/\widehat{\lambda}_l^{[4]})^2$ , the ratios  $(\widehat{\lambda}_l^{[1]}/\widehat{\lambda}_l^{[2]})^2$  remain relatively high for all the values of  $l$  presented in the graph (this behavior could be a consequence of the decay of the spectrum).

To explore the type of solutions obtained for larger values of  $\varkappa$ , we consider the vertex-exchange strategy described in section 8.1. We compute an approximate solution for  $\varkappa = 0.8$ ; the vertex-exchange algorithm is initialized at  $\tilde{\mathbf{v}} = \mathbf{e}_1$  and after 300,000 iterations, we obtain a Frank–Wolfe error bound of  $\epsilon = 1.692408 \times 10^{-8}$ ; the obtained approximate solution  $\widehat{\mathbf{v}}^*$  to problem (5.2) verifies  $D(\widehat{\mathbf{v}}^*) = 4.934072 \times 10^{-5}$

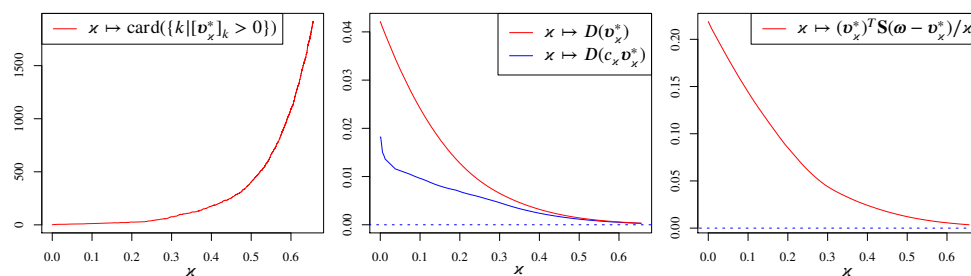


FIG. 11.1. For the MiniBooNE dataset (Gaussian kernel,  $\omega = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ ), graphical representation of the 3,000 first events of the regularization path related to problem (5.2) for increasing  $\varkappa$ : number of strictly positive components of  $\mathbf{v}_\varkappa^*$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

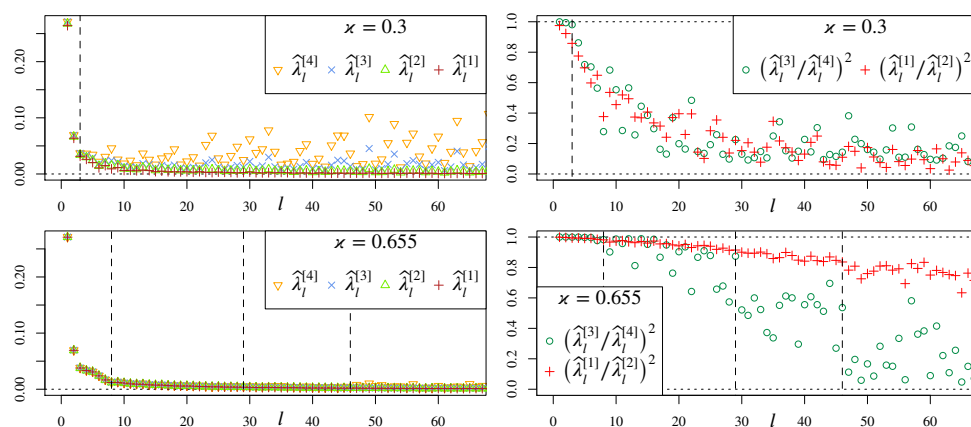


FIG. 11.2. For the MiniBooNE dataset (Gaussian kernel,  $\omega = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ ), approximate eigenvalues  $\hat{\lambda}_l^{[1]}$ ,  $\hat{\lambda}_l^{[2]}$ ,  $\hat{\lambda}_l^{[3]}$ , and  $\hat{\lambda}_l^{[4]}$  induced by the solution to problem (5.1) with  $\varkappa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

and has 9544 strictly positive components (in terms of conic squared-kernel discrepancy, we obtain  $D(c_\circ \hat{\mathbf{v}}^*) = 4.672895 \times 10^{-5}$ ).

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

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

**11.2. Test subsample of the SUSY dataset.** We consider the standardized entries of the test subsample of the SUSY dataset (without labels), so that  $\mathcal{S}$  consists

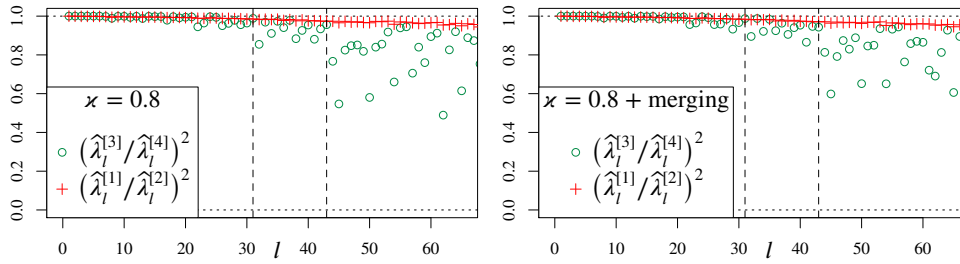


FIG. 11.3. For the MiniBooNE dataset, accuracy of the approximate eigendecompositions induced by the solution  $\hat{\mathbf{v}}^*$  to problem (5.1) with  $\varkappa = 0.8$  obtained from the vertex-exchange algorithm (left) and from the merged solution  $\mathbf{v}_{[5044]}$  (right).

TABLE 11.1

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

	Batch 1	Batch 2	Batch 3	Batch 4
Time (in sec.)	1148.7	1158.3	1158.5	1159.1
Total iterations	50,000	100,000	150,000	200,000
$\epsilon$	$3.1413 \times 10^{-7}$	$6.5477 \times 10^{-8}$	$2.7049 \times 10^{-8}$	$7.0928 \times 10^{-9}$
$n$	19,721	20,619	20,693	20,674

of  $N = 500,000$  points in  $\mathbb{R}^{18}$ . We still use a Gaussian kernel (see section 10) with  $\ell = 0.4$ , and we set  $\boldsymbol{\omega} = \mathbb{1}/N$  and  $\mathbf{d} = \mathbb{1}$ . The computation of the distortion term  $\mathbf{S}\boldsymbol{\omega}$  took 5,665.6 seconds.

We compute an approximate solution (vertex-exchange strategy) for the constrained problem (5.2) with  $\varkappa = 0.3$ ; we perform four consecutive batches of 50,000 iterations each, the solver being initialized at  $\tilde{\mathbf{v}} = \mathbf{e}_1$ . After 200,000 iterations (i.e., at the end of the fourth batch), the obtained approximate solution  $\hat{\mathbf{v}}^*$  verifies  $D(\hat{\mathbf{v}}^*) = 3.931629 \times 10^{-5}$  and has  $n = 20,664$  strictly positive components. Execution times and evolution of the Frank–Wolfe error bound  $\epsilon$  and of the sparsity of the approximate solution are reported in Table 11.1. We observe that a batch of 50,000 iterations of the vertex-exchange algorithm took around 19 minutes; the approximate solution obtained at the end of the first batch is already relatively accurate.

To enhance sparsity, we perform a weak-pairwise merging of the approximate solution  $\hat{\mathbf{v}}^*$ ; the computation of 20,673 merging iterations took 78.86 seconds. The merged solution  $\mathbf{v}_{[13674]}$  is supported by 7,000 points and  $D(\mathbf{v}_{[13674]}) = D(\hat{\mathbf{v}}^*) + 5.271960 \times 10^{-7}$  (i.e., an increase of only 1.34%). We then study the approximate eigendecomposition induced by  $\mathbf{v}_{[13674]}$ . Computing the 300 first normalized approximate eigenvectors  $\hat{\mathbf{v}}_l$  of  $\mathbf{KW}$  induced by  $\mathbf{v}_{[13674]}$  (i.e.,  $\hat{\mathbf{v}}_l \in \mathbb{R}^N$  is the vector corresponding to  $\hat{\varphi}_l$ ; see section 4.3) took 3,278.2 seconds (time for canonical extension and rescaling), and we thus also obtain the approximate eigenvalues  $\hat{\lambda}_l^{[1]}$ . For  $l$  and  $l' \in \{1, \dots, 300\}$ , we have  $\max_{l \neq l'} |(\hat{\varphi}_l | \hat{\varphi}_{l'})_{L^2(\mu)}| \approx 0.003734$ , so that we can expect the approximations  $\hat{\varphi}_l$  to be relatively accurate. To access precisely their accuracy, we compute  $T_\mu[\hat{\varphi}_l]$  (i.e.,  $\mathbf{KW}\hat{\mathbf{v}}_l$ ) for these 300 first approximate eigendirections; this operation took 191,622.3 seconds (i.e., around 53 hours). The results are presented in Figure 11.4. As already observed, the accuracy of the approximate eigendirections



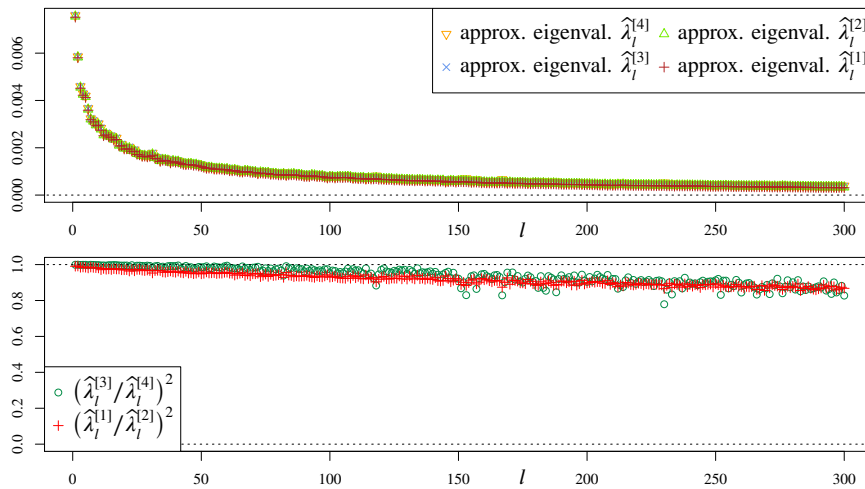


FIG. 11.4. For the test subsample of the SUSY dataset, graphical representation of the 300 first approximate eigenvalues  $\hat{\lambda}_l^{[1]}$  induced by the merged solution  $\mathbf{v}_{[13674]}$  obtained from the approximate solution  $\hat{\mathbf{v}}^*$  to problem (5.2) with  $\varkappa =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

decreases when  $l$  increases (we recall that the eigenvalues of the approximate operator are stored in descending order); all the obtained approximate eigenpairs are remarkably accurate (while considering only 7,000 points among 500,000).

**12. Conclusion.** We have studied a QP-based strategy to design sparse quadratures for the approximation of integral operators related to symmetric positive-semidefinite kernels in a quadrature-sparzification framework, i.e., when only quadratures with support included in a fixed finite set of points are considered. The points selected through penalized squared-kernel-discrepancy minimization can in particular be interpreted as the support vectors of one-class distorted SVMs defined from the squared kernel, the initial measure, and the  $\ell^1$ -type penalization term.

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

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

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

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

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

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

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

From the reproducing property of  $K(\cdot, \cdot)$  and the Cauchy–Schwarz inequality, we have for all  $h \in \mathcal{H}$  and for all  $\mu \in \mathcal{I}(K)$ ,

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

The linear functional  $I_\mu$  on  $\mathcal{H}$ , defined as  $I_\mu[h] = \int_{\mathcal{X}} h(x) d\mu(x)$ , 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  $x \in \mathcal{X}$ ,  $h_\mu(x) = \int_{\mathcal{X}} K(x,t) d\mu(t)$ .

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

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

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

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

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

Thus, when the integrands belong to the RKHS  $\mathcal{H}$ , the error induced by approximating integrals with respect to  $\mu$  by integrals with respect to  $\nu$  has a tight bound in terms



of kernel discrepancy; to approximate integrals with respect to  $\mu$ , it is therefore of interest to deal with a measure  $\nu$  such that  $D_K(\nu, \mu)$  is small; see, for instance, [5, 6] for a further discussion.

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

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

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

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

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

Equalities (B.1) and (B.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  $T_\mu$ . From the linearity and continuity of  $T_\mu$ , we then obtain

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

and we conclude by using the Tonelli theorem.  $\square$

*Proof of Lemma 2.1.* From the properties of  $K(\cdot, \cdot)$ , the squared kernel  $K^2(\cdot, \cdot)$  is symmetric and positive-semidefinite (see in particular the Schur product theorem); in addition the squared kernel is nonnegative, i.e.,  $K^2(x, t) \geq 0$  for all  $x$  and  $t \in \mathcal{X}$ . Considering the framework of Appendix A, we can remark that  $\mathcal{T}(K) = \mathcal{S}(K^2)$ , so that the result directly follows from Proposition 2.1 and the definition of  $g_\mu$  and  $g_\nu$ .  $\square$

*Proof of Lemma 2.2.* The proof directly follows from the properties discussed in sections 2.1 and 2.2. In particular, (2.3) is obtained by considering the o.n.b.  $\{\sqrt{\lambda_k} \varphi_k\}_{k \in \mathbb{I}_\mu^+}$  of  $\mathcal{H}_\mu$  defined by  $T_\mu$  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  $\tau_\mu$  is consequence of the relation  $\|h\|_{L^2(\mu)}^2 \leq \tau_\mu \|h\|_{\mathcal{H}}^2$  for all  $h \in \mathcal{H}$ .  $\square$

*Proof of Theorem 3.1.* We can first remark that if  $\|\psi_l\|_{L^2(\mu)} = 0$ , then  $T_\mu[\psi_l] = 0 = \widehat{\lambda}_l^{[1]} \psi_l$ . For all  $k \in \mathbb{I}_\mu^+$ , we have  $\|\sqrt{\lambda_k} \varphi_k\|_{\mathcal{H}} = 1$ . By analogy, for  $l \in \mathbb{I}_\nu^+$  with  $\|\psi_l\|_{L^2(\mu)} > 0$  (i.e.,  $l \in \widetilde{\mathbb{I}}_\nu^+$ ), we define  $\widehat{\lambda}_l^{[1]}$  so that  $\|\sqrt{\widehat{\lambda}_l^{[1]}} \widehat{\varphi}_l\|_{\mathcal{H}} = 1$ . From the Cauchy-Schwarz inequality, we have

$$\widehat{\lambda}_l^{[1]} = \left( \sqrt{\vartheta_l} \psi_l \mid T_\mu[\sqrt{\vartheta_l} \psi_l] \right)_{\mathcal{H}} \leq \|\sqrt{\vartheta_l} \psi_l\|_{\mathcal{H}} \|T_\mu[\sqrt{\vartheta_l} \psi_l]\|_{\mathcal{H}} = \|T_\mu[\sqrt{\vartheta_l} \psi_l]\|_{\mathcal{H}} = \widehat{\lambda}_l^{[2]}$$

with equality if and only if  $\psi_l$  and  $T_\mu[\psi_l]$  are collinear, i.e.,  $\psi_l$  is an eigendirection of  $T_\mu$ . In particular, since  $\|\sqrt{\vartheta_l} \psi_l\|_{\mathcal{H}} = 1$ , if  $\psi_l$  is an eigendirection of  $T_\mu$ , then  $\widehat{\lambda}_l^{[2]}$  corresponds by definition to the associated eigenvalue, i.e.,  $T_\mu[\psi_l] = \widehat{\lambda}_l^{[2]} \psi_l$  (a similar argument also holds for  $\widehat{\lambda}_l^{[4]}$ ).

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

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

with, again, equality if and only if  $\psi_l$  is an eigendirection of  $T_\mu$ .

The expansions (3.2) and (3.3) follow from the definition of the four geometric approximate eigenvalues related to an approximate eigendirection of  $T_\mu$  induced by  $T_\nu$ , and the optimality properties of  $\widehat{\lambda}_l^{[1]}$  and  $\widehat{\lambda}_l^{[3]}$  are obtained by minimizing the underlying second degree polynomials.  $\square$

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

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

The characterization of  $T_{c_\nu \nu}$  as an orthogonal projection and the fact that all such operators lie on a sphere in  $\text{HS}(\mathcal{H})$  is a direct consequence of the definition of  $c_\nu$ ; notice, for instance, that, in  $\text{HS}(\mathcal{H})$ ,

$$T_{c_\nu \nu} = c_\nu T_\nu = \left( (T_\mu | T_\nu)_{\text{HS}(\mathcal{H})} / \|T_\nu\|_{\text{HS}(\mathcal{H})}^2 \right) T_\nu.$$

By definition,  $\{\sqrt{\vartheta_l} \psi_l\}_{l \in \mathbb{I}_\nu^+}$  is an o.n.b. of  $\mathcal{H}_\nu = \mathcal{H}_{c_\nu \nu}$ , and we have  $\sqrt{\widehat{\lambda}_l^{[1]}} \widehat{\varphi}_l = \sqrt{\vartheta_l} \psi_l$  for all  $l \in \mathbb{I}_\nu^+$ . Introducing an o.n.b.  $\{h_m\}_{m \in \mathbb{J}}$  of the subspace  $\mathcal{H}_{0\nu}$  of  $\mathcal{H}$ , we obtain

$$(B.3a) \quad D_{K^2}(\mu, c_\nu \nu) = \sum_{l \in \mathbb{I}_\nu^+} \left\| T_\mu \left[ \sqrt{\widehat{\lambda}_l^{[1]}} \widehat{\varphi}_l \right] - c_\nu \vartheta_l \sqrt{\widehat{\lambda}_l^{[1]}} \widehat{\varphi}_l \right\|_{\mathcal{H}}^2$$

$$(B.3b) \quad + \sum_{l \in \mathbb{I}_\nu^+ \setminus \mathbb{I}_\nu^+} \|c_\nu \vartheta_l \sqrt{\vartheta_l} \psi_l\|_{\mathcal{H}}^2 + \sum_{m \in \mathbb{J}} \|T_\mu [h_m]\|_{\mathcal{H}}^2.$$

Since all the terms appearing in (B.3b) are positive, (B.3a) can be turned into the required inequality. We conclude by using the optimality properties of the approximate eigenvalues  $\widehat{\lambda}_l^{[1]}$  and  $\widehat{\lambda}_l^{[3]}$  described in Theorem 3.1; for (3.8), we also use the inequality  $\|h\|_{L^2(\mu)}^2 \leq \tau_\mu \|h\|_{\mathcal{H}}^2$  for all  $h \in \mathcal{H}$ .  $\square$

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

canonical basis of  $\mathbb{R}^N$ , so that  $\mathbf{e}_k^T \nabla D_\alpha(\mathbf{v}_\alpha^*) = [\nabla D_\alpha(\mathbf{v}_\alpha^*)]_k$ . We thus necessarily have  $\nabla D_\alpha(\mathbf{v}_\alpha^*) \geq 0$  and  $(\mathbf{v}_\alpha^*)^T \nabla D_\alpha(\mathbf{v}_\alpha^*) = 0$  (since  $\mathbf{v}_\alpha^* \geq 0$ ). To prove (e), we first remark that

$$D_\alpha(\tilde{\mathbf{v}}_\alpha^*) = D_\alpha(\mathbf{v}_\alpha^*) + (\tilde{\mathbf{v}}_\alpha^* - \mathbf{v}_\alpha^*)^T \nabla D_\alpha(\mathbf{v}_\alpha^*) + \frac{1}{2}(\tilde{\mathbf{v}}_\alpha^* - \mathbf{v}_\alpha^*)^T \mathbf{S}(\tilde{\mathbf{v}}_\alpha^* - \mathbf{v}_\alpha^*).$$

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

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

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

*Proof of Theorem 5.1.* For  $0 \leq \alpha < \max_k \{[\mathbf{S}\boldsymbol{\omega}]_k / d_k\} = \alpha_0$ , we have  $\mathbf{v}_\alpha^* \neq 0$ ; see Proposition 5.1(b). In addition, if  $\alpha$  is such that  $\mathbf{S}(\mathbf{v}_\alpha^* - \boldsymbol{\omega}) = 0$ , then  $c_\alpha = 1$ . We now assume that  $\mathbf{S}(\mathbf{v}_\alpha^* - \boldsymbol{\omega}) \neq 0$ ; from Proposition 5.1(d), we have  $(\mathbf{v}_\alpha^*)^T \mathbf{S}(\mathbf{v}_\alpha^* - \boldsymbol{\omega}) + \alpha \mathbf{d}^T \mathbf{v}_\alpha^* = 0$ , leading to  $c_\alpha = 1 + \frac{\alpha \mathbf{d}^T \mathbf{v}_\alpha^*}{(\mathbf{v}_\alpha^*)^T \mathbf{S} \mathbf{v}_\alpha^*} \geq 1$ . By definition of  $c_\alpha$ , we also have  $c_\alpha (\mathbf{v}_\alpha^*)^T \mathbf{S} \mathbf{v}_\alpha^* = \boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_\alpha^*$ , so that

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

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

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

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

Consider  $0 \leq \alpha_1 < \alpha_2 < \alpha_0$ ; from Proposition 5.1(d) and by definition of  $c_\alpha$ , we have

$$\begin{aligned} (c_{\alpha_1} \mathbf{v}_{\alpha_2}^* - \mathbf{v}_{\alpha_1}^*)^T [\mathbf{S}(\mathbf{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \alpha_1 \mathbf{d}] &= (c_{\alpha_1} \mathbf{v}_{\alpha_2}^*)^T \mathbf{S}(\mathbf{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \alpha_1 c_{\alpha_1} \mathbf{d}^T \mathbf{v}_{\alpha_2}^* \\ &= (\mathbf{v}_{\alpha_2}^*)^T \mathbf{S}(c_{\alpha_1} \mathbf{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \alpha_1 c_{\alpha_1} \mathbf{d}^T \mathbf{v}_{\alpha_2}^* - (c_{\alpha_1} - 1) \boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_{\alpha_2}^* \\ &= (\mathbf{v}_{\alpha_2}^*)^T \mathbf{S}(c_{\alpha_1} \mathbf{v}_{\alpha_1}^* - \boldsymbol{\omega}) + \frac{\alpha_1}{(\mathbf{v}_{\alpha_1}^*)^T \mathbf{S} \mathbf{v}_{\alpha_1}^*} [(\boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_{\alpha_1}^*) \mathbf{d}^T \mathbf{v}_{\alpha_2}^* - (\boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_{\alpha_2}^*) \mathbf{d}^T \mathbf{v}_{\alpha_1}^*] \geq 0. \end{aligned}$$

(B.4)

Since  $(\boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_{\alpha_2}^*) \mathbf{d}^T \mathbf{v}_{\alpha_1} \geq (\boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_{\alpha_1}^*) \mathbf{d}^T \mathbf{v}_{\alpha_2}$  (we indeed assume that  $\alpha \mapsto \boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_{\alpha}^* / \mathbf{d}^T \mathbf{v}_{\alpha}^*$  is increasing), inequality (B.4) entails that  $(\mathbf{v}_{\alpha_2}^*)^T \mathbf{S} (c_{\alpha_1} \mathbf{v}_{\alpha_1}^* - \boldsymbol{\omega}) \geq 0$ , so that  $(c_{\alpha_2} \mathbf{v}_{\alpha_2}^* - c_{\alpha_1} \mathbf{v}_{\alpha_1}^*)^T \mathbf{S} (c_{\alpha_1} \mathbf{v}_{\alpha_1}^* - \boldsymbol{\omega}) \geq 0$ , and thus, by convexity,  $D(c_{\alpha_2} \mathbf{v}_{\alpha_2}^*) \geq D(c_{\alpha_1} \mathbf{v}_{\alpha_1}^*)$ ; we recall that  $\nabla D(\mathbf{v}) = \mathbf{S}(\mathbf{v} - \boldsymbol{\omega})$ . For all  $\alpha \geq 0$ ,  $c \geq 0$ , and  $\mathbf{v} \geq 0$ , from Proposition 5.1(d), we also have

$$(B.5) \quad (\mathbf{v} - c\mathbf{v}_{\alpha}^*)^T [\mathbf{S}(\mathbf{v}_{\alpha}^* - \boldsymbol{\omega}) + \alpha \mathbf{d}] \geq 0.$$

We introduce  $\tau = \boldsymbol{\omega}^T \mathbf{S} (c_{\alpha_2} \mathbf{v}_{\alpha_2}^* - c_{\alpha_1} \mathbf{v}_{\alpha_1}^*) = 2(D(c_{\alpha_1} \mathbf{v}_{\alpha_1}^*) - D(c_{\alpha_2} \mathbf{v}_{\alpha_2}^*))$ , and  $\beta = \mathbf{d}^T (c_{\alpha_2} \mathbf{v}_{\alpha_2}^* - c_{\alpha_1} \mathbf{v}_{\alpha_1}^*)$ . From (B.5) and by definition of  $c_{\alpha}$ , we deduce that

$$(B.6) \quad \begin{aligned} (c_{\alpha_1} \mathbf{v}_{\alpha_1}^* - c_{\alpha_2} \mathbf{v}_{\alpha_2}^*)^T \mathbf{S} (\mathbf{v}_{\alpha_2}^* - \boldsymbol{\omega}) &= \tau + (\mathbf{v}_{\alpha_2}^*)^T \nabla D(c_{\alpha_1} \mathbf{v}_{\alpha_1}^*) \\ &= \tau + \frac{1}{c_{\alpha_2}} (c_{\alpha_2} \mathbf{v}_{\alpha_2}^* - c_{\alpha_1} \mathbf{v}_{\alpha_1}^*)^T \nabla D(c_{\alpha_1} \mathbf{v}_{\alpha_1}^*) \geq \alpha_2 \beta. \end{aligned}$$

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

$$(B.7) \quad -\frac{1}{2} \tau \geq (c_{\alpha_2} \mathbf{v}_{\alpha_2}^* - c_{\alpha_1} \mathbf{v}_{\alpha_1}^*)^T \nabla D(c_{\alpha_1} \mathbf{v}_{\alpha_1}^*).$$

Since  $0 < 1/c_{\alpha_2} \leq 1$  and  $\tau \leq 0$ , inequalities (B.6) and (B.7) imply  $\beta \leq 0$ , as expected.  $\square$

*Proof of Proposition 6.1.* Define the closed linear subspace  $\mathcal{G}_S = \text{span}\{K^2(x_k, \cdot)\}_{k=1}^N$  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

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

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

$$(B.8) \quad \underset{\boldsymbol{\beta}}{\text{minimize}} \quad \frac{1}{2} \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\omega} \quad \text{subject to} \quad \mathbf{S} \boldsymbol{\beta} \geq -\alpha \mathbf{d}.$$

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

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

The primal optimality condition gives  $\mathbf{S} \boldsymbol{\beta} = \mathbf{S}(\mathbf{v} - \boldsymbol{\omega})$ , leading to the Lagrange dual (5.1) (written as a minimization problem). If  $\mathbf{v}_{\alpha}^*$  is a solution to (5.2), then a solution  $\boldsymbol{\beta}^*$  to (B.8) needs to satisfy  $\mathbf{S} \boldsymbol{\beta}^* = \mathbf{S}(\mathbf{v}_{\alpha}^* - \boldsymbol{\omega})$ , so that we can in particular consider  $\boldsymbol{\beta}^* = \mathbf{v}_{\alpha}^* - \boldsymbol{\omega}$ . Notice that when  $\mathbf{S}$  is noninvertible, 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 solution  $g_\alpha^* \in \mathcal{G}_S$  does not depend on such a  $\boldsymbol{\varepsilon}$ . The equality  $g_\alpha^*(x_k) = -\alpha d_k$  for all  $k \in \{1, \dots, N\}$  such that  $[\mathbf{v}_{\alpha}^*]_k > 0$  is a consequence of the complementary slackness condition  $(\mathbf{v}_{\alpha}^*)^T [\mathbf{S}(\mathbf{v}_{\alpha}^* - \boldsymbol{\omega}) + \alpha \mathbf{d}] = 0$ .  $\square$

*Proof of Proposition 6.2.* We follow the same reasoning as in the proof of Proposition 6.1. By restricting problem (6.3) to  $\mathcal{G}_S$ , we obtain, for  $\boldsymbol{\beta} \in \mathbb{R}^N$ ,

$$(B.9) \quad \underset{\boldsymbol{\beta}, \gamma}{\text{minimize}} \quad \frac{1}{2} \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\omega} - \gamma \quad \text{subject to} \quad \mathbf{S} \boldsymbol{\beta} \geq \gamma \mathbf{d} / \varkappa.$$

The underlying Lagrangian function is then given by, for  $\mathbf{v} \in \mathbb{R}^N$  with  $\mathbf{v} \geq 0$ ,

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

The primal optimality condition gives  $\mathbf{S} \boldsymbol{\beta} = \mathbf{S}(\mathbf{v} - \boldsymbol{\omega})$  and  $\mathbf{d}^T \mathbf{v} = \varkappa$ , leading to the Lagrange dual (5.2). If  $\mathbf{v}_\varkappa^*$  is a solution to (5.2), then a solution  $\boldsymbol{\beta}^*$  to (B.9) needs to satisfy  $\mathbf{S} \boldsymbol{\beta}^* = \mathbf{S}(\mathbf{v}_\varkappa^* - \boldsymbol{\omega})$ , so that we can in particular consider  $\boldsymbol{\beta}^* = \mathbf{v}_\varkappa^* - \boldsymbol{\omega}$ . The expression of  $\gamma_\varkappa^*$  follows from the complementary slackness condition  $(\mathbf{v}_\varkappa^*)^T [\mathbf{S}(\mathbf{v}_\varkappa^* - \boldsymbol{\omega}) - \gamma_\varkappa^* \mathbf{d} / \varkappa] = 0$ , as well as the equality  $g_\varkappa^*(x_k) = \gamma_\varkappa^* d_k / \varkappa$  for all  $k \in \{1, \dots, N\}$  such that  $[\mathbf{v}_\varkappa^*]_k > 0$ .  $\square$

*Proof of Proposition 7.1.* Let  $\mathbf{v}_\alpha = \theta \mathbf{v}_{\alpha_1}^* + (1 - \theta) \mathbf{v}_{\alpha_2}^*$ , and consider  $J = J_{\alpha_1} = J_{\alpha_2}$ ; we have

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

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

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

By construction,  $\mathbf{v}_\alpha \geq 0$ , and if  $k$  is such that  $[\mathbf{v}_\alpha]_k > 0$ , then  $k \in J$  (since these conditions are verified by both  $\mathbf{v}_{\alpha_1}^*$  and  $\mathbf{v}_{\alpha_2}^*$ ). We therefore have  $\mathbf{v}_\alpha^T (\mathbf{S}(\mathbf{v}_\alpha - \boldsymbol{\omega}) + \alpha \mathbf{d}) = 0$ , so that for all  $\mathbf{v} \geq 0$ , the optimality condition  $(\mathbf{v} - \mathbf{v}_\alpha)^T \nabla D_\alpha(\mathbf{v}_\alpha) \geq 0$  holds, i.e.,  $\mathbf{v}_\alpha$  is a solution to (5.1), and  $J_\alpha = J$ .  $\square$

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

$$(B.10) \quad [\mathbf{v}_\alpha^*]_{J_\alpha} = \mathbf{G}([\mathbf{S} \boldsymbol{\omega}]_{J_\alpha} - \alpha \mathbf{d}_{J_\alpha}),$$

and  $\mathbf{G}$  is a generalized inverse of  $\mathbf{S}_{J_\alpha, J_\alpha}$  (i.e.,  $\mathbf{S}_{J_\alpha, J_\alpha} \mathbf{G} \mathbf{S}_{J_\alpha, J_\alpha} = \mathbf{S}_{J_\alpha, J_\alpha}$ ); see, for instance, [2]. Combined with Proposition 7.1, condition (7.1) thus implies that the maps  $\alpha \mapsto \boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_\alpha^*$  and  $\alpha \mapsto \mathbf{d}^T \mathbf{v}_\alpha^*$  are piecewise linear; in addition, since the indices of the strictly positive components of  $\mathbf{v}_\alpha^*$  always belongs to  $J_\alpha$ , any change in the sparsity pattern only involves null components of  $\mathbf{v}_\alpha^*$ , so that these two maps are also continuous. We then introduce  $\zeta(\alpha) = \boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_\alpha^* / \mathbf{d}^T \mathbf{v}_\alpha^*$ ; on the interval  $[0, \alpha_0)$ , the function  $\zeta(\cdot)$  is continuous (since, on this interval,  $\alpha \mapsto \boldsymbol{\omega}^T \mathbf{S} \mathbf{v}_\alpha^*$  and  $\alpha \mapsto \mathbf{d}^T \mathbf{v}_\alpha^*$  are continuous, and  $\mathbf{d}^T \mathbf{v}_\alpha^* > 0$ ). From (B.10), and since  $[\mathbf{v}_\alpha^*]_{J_\alpha^c} = 0$ , we have

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

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

$$(B.11) \quad ([\mathbf{S} \boldsymbol{\omega}]_{J_\alpha}^T \mathbf{G} \mathbf{d}_{J_\alpha})^2 \leq ([\mathbf{S} \boldsymbol{\omega}]_{J_\alpha}^T \mathbf{G} [\mathbf{S} \boldsymbol{\omega}]_{J_\alpha}) (\mathbf{d}_{J_\alpha}^T \mathbf{G} \mathbf{d}_{J_\alpha}).$$

If  $\mathbf{G}$  is symmetric and positive-semidefinite, then inequality (B.11) corresponds to the Cauchy–Schwarz inequality and is therefore verified. Since the number of kinks is finite, we can thus conclude that  $\zeta(\cdot)$  is increasing on  $[0, \alpha_0)$ .  $\square$

*Proof of Lemma 7.1.* Let  $\mathbf{v}_\alpha$  be such that  $[\mathbf{v}_\alpha]_{J^c} = 0$  and  $[\mathbf{v}_\alpha]_J = (\mathbf{S}_{J,J})^{-1}([\mathbf{S} \boldsymbol{\omega}]_J - \alpha \mathbf{d}_J)$ . Following (7.2), from the condition  $[\mathbf{S}(\mathbf{v}_\alpha - \boldsymbol{\omega}) + \alpha \mathbf{d}]_{J^c} \geq 0$ , we define  $\alpha_+$  as the smallest  $\alpha$  satisfying the constraint  $\alpha [\mathbf{M} \mathbf{d}_J - \mathbf{d}_{J^c}]_l \leq [\mathbf{M} [\mathbf{S} \boldsymbol{\omega}]_J - [\mathbf{S} \boldsymbol{\omega}]_{J^c}]_l$ , for all  $l \in \{1, \dots, \text{card}(J^c)\}$ . By definition (and in view of Remark 7.1), this constraint

is satisfied by  $\alpha_p$ ; the components  $l$  such that  $[\mathbf{M}\mathbf{d}_J - \mathbf{d}_{J^c}]_l \geq 0$  therefore carry no information. The problem thus consists in searching for the smallest  $\alpha$  such that

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

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

**Acknowledgments.** The authors would like to thank the two referees and the editor for their valuable comments.

## REFERENCES

- [1] F. BACH, *On the equivalence between kernel quadrature rules and random feature expansions*, J. Mach. Learn. Res., 18 (2017), pp. 1–38.
- [2] A. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses: Theory and Applications*. Springer, New York, 2003.
- [3] A. BERLINET AND C. THOMAS-AGNAN, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer, New York, 2011.
- [4] C. BOUTSIDIS, P. DRINEAS, AND M. MAGDON-ISMAIL, *Near-optimal column-based matrix reconstruction*, SIAM J. Comput., 43 (2014), pp. 687–717.
- [5] S. B. DAMELIN, *A walk through energy, discrepancy, numerical integration and group invariant measures on measurable subsets of Euclidean space*, Numer. Algorithms, 48 (2008), pp. 213–235.
- [6] S. B. DAMELIN, F. J. HICKERNELL, D. L. RAGOZIN, AND X. ZENG, *On energy, discrepancy and group invariant measures on measurable subsets of Euclidean space*, J. Fourier Anal. Appl., 16 (2010), pp. 813–839.
- [7] D. DHEERU AND E. K. TANISKIDOU, *UCI Machine Learning Repository*, University of California, Irvine, CA, 2017.
- [8] P. DRINEAS AND M. W. MAHONEY, *On the Nyström method for approximating a Gram matrix for improved kernel-based learning*, J. Mach. Learn. Res., 6 (2005), pp. 2153–2175.
- [9] B. GAUTHIER AND L. PRONZATO, *Spectral approximation of the IMSE criterion for optimal designs in kernel-based interpolation models*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 805–825.
- [10] B. GAUTHIER AND L. PRONZATO, *Convex relaxation for IMSE optimal design in random-field models*, Comput. Statist. Data Anal., 113 (2017), pp. 375–394.
- [11] A. GITTENS AND M. W. MAHONEY, *Revisiting the Nyström method for improved large-scale machine learning*, J. Mach. Learn. Res., 17 (2016), pp. 1–65.
- [12] J. GUÉLAT AND P. MARCOTTE, *Some comments on Wolfe’s “away step”*, Math. Program., 35 (1986), pp. 110–119.
- [13] T. HASTIE, S. ROSSET, R. TIBSHIRANI, AND JI. ZHU, *The entire regularization path for the support vector machine*, J. Mach. Learn. Res., 5 (2004), pp. 1391–1415.
- [14] T. HASTIE, R. TIBSHIRANI, AND M. WAINWRIGHT, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, Boca Raton, FL, 2015.
- [15] S. KUMAR, M. MOHRI, AND A. TALWALKAR, *Sampling methods for the Nyström method*, J. Mach. Learn. Res., 13 (2012), pp. 981–1006.
- [16] H. NIEDERREITER, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.
- [17] M. R. OSBORNE, B. PRESNELL, AND B. A. TURLACH, *A new approach to variable selection in least squares problems*, IMA J. Numer. Anal., 20 (2000), pp. 389–403.
- [18] L. PRONZATO AND A. PÁZMAN, *Design of Experiments in Nonlinear Models*, Springer, New York, 2013.
- [19] B. SCHÖLKOPF, J. C. PLATT, J. S.-TAYLOR, A. J. SMOLA, AND R. C. WILLIAMSON, *Estimating the support of a high-dimensional distribution*, Neural Comput., 13 (2001), pp. 1443–1471.
- [20] L. SCHWARTZ, *Analyse Hilbertienne*, Hermann, Paris, 1978.
- [21] D. SEJIDINOVIC, B. SRIPERUMBUDUR, A. GRETTON, AND K. FUKUMIZU, *Equivalence of distance-based and RKHS-based statistics in hypothesis testing*, Ann. Statist., 41 (2013), pp. 2263–2291.
- [22] S. SMALE AND D.-X ZHOU, *Learning theory estimates via integral operators and their approximations*, Constr. Approx., 26 (2007), pp. 153–172.

- [23] S. SMALE AND D.-X. ZHOU, *Geometry on probability spaces*, Constr. Approx., 30 (2009), pp. 311–323.
- [24] I. STEINWART AND A. CHRISTMANN, *Support Vector Machines*. Springer, New York, 2008.
- [25] S. WANG AND Z. ZHANG, *Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling*, J. Mach. Learn. Res., 14 (2013), pp. 2729–2769.
- [26] K. ZHANG AND J. T. KWOK, *Clustered Nyström method for large scale manifold learning and dimension reduction*, IEEE Trans. Neural Networks, 21 (2010), pp. 1576–1587.