SIAM J. MATH. DATA SCI.
Vol. 3, No. 1, pp. 201–224

Multikernel Regression with Sparsity Constraint*

Shayan Aziznejad† and Michael Unser†

Abstract. In this paper, we provide a Banach-space formulation of supervised learning with generalized total-variation (gTV) regularization. We identify the class of kernel functions that are admissible in this framework. Then, we propose a variation of supervised learning in a continuous-domain hybrid search space with gTV regularization. We show that the solution admits a multikernel expansion with adaptive positions. In this representation, the number of active kernels is upper-bounded by the number of data points while the gTV regularization imposes an ℓ_1 penalty on the kernel coefficients. Finally, we illustrate numerically the outcome of our theory.

Key words. representer theorem, regularization theory, multiple-kernel learning, generalized LASSO, generalized total variation

AMS subject classifications. 46E22, 46E27, 47A52, 62G08, 68T05

DOI. 10.1137/20M1318882

1. Introduction.

The determination of an unknown function from a series of samples is a classical problem in machine learning. It falls under the category of "supervised learning," for which there exists a rich literature (see [9, 35, 68] for classical textbooks, as well as [13, 20, 34, 59] for more recent ones). The goal of supervised learning is to recover a target function $f : \mathbb{R}^d \to \mathbb{R}$ from its M noisy samples $y_m = f(\boldsymbol{x}_m) + \epsilon_m$, $m = 1, 2, \ldots, M$. The disturbance terms ϵ_m are typically assumed to be independent and identically distributed (i.i.d.) samples of a zero-mean probability law (e.g., additive Gaussian noise) while the input vectors $\boldsymbol{x}_m$ are assumed to be in either the random or fixed design [34, section 1.9].

A general way to formulate supervised learning is through the minimization problem

$$\min_f \Bigg(\underbrace{\sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m)}_{\mathrm{I}} + \lambda \underbrace{\mathcal{R}(f)}_{\mathrm{II}} \Bigg), \tag{1.1}$$

where the cost function is made of two terms. The first one (data fidelity) measures how well f fits the given training dataset while the second one (regularization) imposes the prior knowledge about the function model. The parameter $\lambda \in \mathbb{R}^+$ balances the terms.

1.1. RKHS in machine learning.

The simplest form of (1.1) is the least-squares problem with Tikhonov regularization

*Received by the editors February 13, 2020; accepted for publication (in revised form) December 3, 2020; published electronically February 8, 2021.
https://doi.org/10.1137/20M1318882
Funding: This work was funded by the Swiss National Science Foundation under grant 200020_184646/1.
†Biomedical Imaging Group, EPFL, Lausanne, Switzerland (shayan.aziznejad@epfl.ch, michael.unser@epfl.ch).

$$\min_{f\in\mathcal{H}_{\mathrm{L}}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} |f(\boldsymbol{x}_m) - y_m|^2 + \lambda \|\mathrm{L}\{f\}\|_{L_2}^2 \right), \tag{1.2}$$

where L is the regularization operator and $\mathcal{H}_{\mathrm{L}}(\mathbb{R}^d)$, known as the native space of L, is the space of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathrm{L}\{f\} \in L_2(\mathbb{R}^d)$ (see (2.1) for the definition of the L_p spaces). It is a classical quadratic minimization problem that has a closed-form solution [63]. An important assumption in this formulation is the continuity of the sampling functionals $\delta_{\boldsymbol{x}_m} = \delta(\cdot - \boldsymbol{x}_m) : f \mapsto f(\boldsymbol{x}_m)$ for $m = 1, 2, \ldots, M$. This is equivalent to $\mathcal{H}_{\mathrm{L}}(\mathbb{R}^d)$ being a reproducing-kernel Hilbert space (RKHS) [1, 15, 68], which is a key concept in supervised learning [8, 59].

The Hilbert space $\mathcal{H}(\mathbb{R}^d)$ consisting of functions from $\mathbb{R}^d$ to $\mathbb{R}$ is called an RKHS if there exists a bivariate symmetric and positive-definite function $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that, for all $\boldsymbol{x} \in \mathbb{R}^d$, $\mathrm{k}(\boldsymbol{x}, \cdot) \in \mathcal{H}(\mathbb{R}^d)$ and $f(\boldsymbol{x}) = \langle \mathrm{k}(\boldsymbol{x}, \cdot), f(\cdot) \rangle_{\mathcal{H}}$ [1]. The function $\mathrm{k}(\cdot, \cdot)$ is unique and is called the reproducing kernel of $\mathcal{H}(\mathbb{R}^d)$.

The supervised learning over the RKHS $\mathcal{H}(\mathbb{R}^d)$ can be formulated through the minimization

$$\min_{f\in\mathcal{H}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \|f\|_{\mathcal{H}}^2 \right). \tag{1.3}$$

The kernel representer theorem states that the solution of (1.3) admits the form

$$f(\cdot) = \sum_{m=1}^{M} a_m \mathrm{k}(\cdot, \boldsymbol{x}_m) \tag{1.4}$$

for some appropriate weights $a_m \in \mathbb{R}$, where $m = 1, 2, \ldots, M$ [36, 49]. The expansion (1.4) is the key element of kernel-based schemes in machine learning [50, 52, 67] and, in particular, support-vector machines (SVMs) [24, 59]. Moreover, optimal rates have been derived for learning using the expansion (1.4) in several setups [12, 42, 62], particularly for Gaussian kernels [23]. Computing the RKHS norm of a function f of the form (1.4) results in $\|f\|_{\mathcal{H}}^2 = \boldsymbol{a}^T \mathbf{G} \boldsymbol{a}$, where $\mathbf{G} \in \mathbb{R}^{M\times M}$ is a symmetric and positive-definite matrix with $[\mathbf{G}]_{m,n} = \mathrm{k}(\boldsymbol{x}_m, \boldsymbol{x}_n)$. It is called the Gram matrix of the kernel $\mathrm{k}(\cdot, \cdot)$. The practical outcome of this observation is that the infinite-dimensional problem (1.3) over the space of functions $\mathcal{H}(\mathbb{R}^d)$ becomes equivalent to the finite-dimensional problem [49]

$$\min_{\boldsymbol{a}\in\mathbb{R}^M} \left(\sum_{m=1}^{M} \mathrm{E}([\mathbf{G}\boldsymbol{a}]_m, y_m) + \lambda \boldsymbol{a}^T \mathbf{G} \boldsymbol{a} \right), \tag{1.5}$$

which is of size M and can be computed numerically.

1.2. Toward sparse kernel expansions. In the solution form (1.4), the kernels are shifted to the location of the data samples. This is elegant but can become cumbersome when the number of samples M grows large. Several schemes have been developed to reduce the number of active kernels. One proposed approach is to use a sparsity-enforcing loss such as the ϵ-insensitive norm of SVM regression [57, 58, 60]. Another approach is to replace the quadratic

regularization $\boldsymbol{a}^T\mathbf{G}\boldsymbol{a}$ in the reduced finite-dimensional problem (1.5) by a sparsity-promoting penalty such as $\|\boldsymbol{a}\|_{\ell_1} = \sum_{m=1}^{M} |a_m|$. This results in (1.5) becoming

$$\min_{\boldsymbol{a}\in\mathbb{R}^M} \left(\sum_{m=1}^{M} \mathrm{E}([\mathbf{G}\boldsymbol{a}]_m, y_m) + \lambda \|\boldsymbol{a}\|_{\ell_1} \right), \tag{1.6}$$

which is called the generalized least absolute shrinkage and selection operator (LASSO) [45]. The properties of this estimator have been studied from both a statistical [53] and approximation-theoretical point of view [69].

In this paper, we consider a Banach-space formulation of supervised learning. We choose the generalized total-variation (gTV) norm as the regularization term in order to promote sparsity in the continuous domain. The effect of gTV regularization has been extensively studied in the context of linear inverse problems [28, 41, 64]. For an invertible operator L (see Definition 3.1), the gTV norm is defined as

$$\mathrm{gTV}(f) = \|\mathrm{L}\{f\}\|_{\mathcal{M}}, \tag{1.7}$$

where $\mathcal{M}(\mathbb{R}^d)$ is the space of bounded Radon measures (see (2.2) for a precise definition) and $\|\cdot\|_{\mathcal{M}}$ is the total-variation norm in the sense of measures [47].

One can formulate supervised learning with gTV regularization through the minimization

$$\min_{f\in\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \|\mathrm{L}\{f\}\|_{\mathcal{M}} \right), \tag{1.8}$$

where $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$ is the native Banach space of the operator $\mathrm{L} : \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d)$ equipped with the gTV norm (see Definition 3.2). The fact that $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$ is a Banach space (i.e., a complete normed space) follows from the invertibility of L (see Theorem 3.3). A consequence of the general representer theorem of [64] is that there is always a solution of (1.8) that admits a linear kernel expansion of the form

$$f(\cdot) = \sum_{l=1}^{M_0} a_l \mathrm{k}(\cdot, \boldsymbol{z}_l) \tag{1.9}$$

for some unknown integer $M_0 \leq M$, nonzero kernel weights $a_l \in \mathbb{R}$, and some distinct adaptive kernel positions $\boldsymbol{z}_l \in \mathbb{R}^d$ [33]. There, the function $\mathrm{k}(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the shift-invariant kernel associated to the Green's function of the operator L. In other words, we have that $\mathrm{k}(\boldsymbol{x}, \boldsymbol{y}) = \rho_{\mathrm{L}}(\boldsymbol{x} - \boldsymbol{y})$, where $\rho_{\mathrm{L}} = \mathrm{L}^{-1}\{\delta\}$.

There exist works on supervised learning over Banach spaces, especially via the concept of reproducing-kernel Banach spaces (RKBS) [25, 71, 72]. However, there are several differences between RKBS and our proposed scheme of learning with gTV regularization. Firstly, as highlighted in [72], the RKBS representer theorem yields a nonlinear kernel expansion for the optimal solution. Secondly, its kernel positions necessarily coincide with the data points. Last but not least, the Banach spaces in the RKBS theory are restricted to reflexive ones (see section 2.1 for the definition of reflexive Banach spaces), which excludes the case of learning with gTV regularization that is known to enforce sparsity in the continuous domain.

Let us also mention that a formulation with strong link to (1.8) has been presented in [3] for learning a function $f : \mathbb{R}^d \to \mathbb{R}$ from a continuously indexed family of atoms $\{\mathrm{k}_{\boldsymbol{z}}\}_{\boldsymbol{z}\in\mathcal{V}}$, where $\mathcal{V}$ is a compact topological space. Putting it in a similar form as (1.8), the proposed formulation in [3] for supervised learning is equivalent to the minimization

$$\min_{\mu\in\mathcal{M}(\mathcal{V})} \left(\sum_{m=1}^{M} \mathrm{E}\left(\int_{\mathcal{V}} \mathrm{k}_{\boldsymbol{z}}(\boldsymbol{x}_m)\mathrm{d}\mu(\boldsymbol{z}), y_m \right) + \lambda\|\mu\|_{\mathcal{M}} \right), \tag{1.10}$$

where $\mathcal{M}(\mathcal{V})$ is the space of Radon measures over $\mathcal{V}$. The relevant property there is that the minimization of (1.10) introduces an atomic measure $\mu = \sum_{l=1}^{M_0} a_l \delta(\cdot - \boldsymbol{z}_l)$. It hence suggests the parametric form (1.9) with $\mathrm{k}(\cdot, \boldsymbol{z}_l) = \mathrm{k}_{\boldsymbol{z}_l}(\cdot)$ for the learned function.

The minimization problem (1.10) is a synthesis-based formulation for supervised learning where the basis functions are known a priori, contrary to (1.8) which is an analysis-based formalism that relies on regularization theory in Banach spaces. Interestingly, the two formulations are equivalent when the family of atoms in (1.10) coincides with the class of shifted Green's function of the regularization operator L; that is, $\mathrm{k}_{\boldsymbol{z}}(\cdot) = \rho_{\mathrm{L}}(\cdot - \boldsymbol{z})$.

To conclude this section, we discuss the connection between (1.8) and generalized LASSO. One readily verifies that the gTV norm enforces an ℓ_1 penalty on the kernel coefficients a_l. More precisely, the expansion (1.9) translates the original problem (1.8) into the discrete minimization

$$\min_{\boldsymbol{a}\in\mathbb{R}^M, \mathrm{Z}\in\mathbb{R}^{d\times M_0}} \left(\sum_{m=1}^{M} \mathrm{E}([\mathbf{G}_{\mathrm{Z}}\boldsymbol{a}]_m, y_m) + \lambda\|\boldsymbol{a}\|_{\ell_1} \right), \tag{1.11}$$

where $\mathrm{Z} = \left(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{M_0}\right)$ is the kernel-position matrix and $\mathbf{G}_{\mathrm{Z}} \in \mathbb{R}^{M\times M_0}$ is a matrix with $[\mathbf{G}_{\mathrm{Z}}]_{m,l} = \mathrm{k}(\boldsymbol{x}_m, \boldsymbol{z}_l)$. The reduced problem (1.11) can be seen as an extended version of the generalized LASSO in (1.6). The fundamental difference is that the minimization is through the positions as well.

1.3. Multikernel schemes. The solution forms (1.4) and (1.9) heavily depend on the kernel function $\mathrm{k}(\cdot,\cdot)$. Hence, choosing the proper kernel is a challenging task that requires careful consideration. One can use a cross-validation scheme in order to compare the performance of several kernel estimators and select the best one for the desired application [32]. Another approach is to learn a new kernel function $\mathrm{k}_{\boldsymbol{\mu}} = \sum_{n=1}^{N} \mu_n \mathrm{k}_n$ from a family of given kernels $\mathrm{k}_1, \mathrm{k}_2, \ldots, \mathrm{k}_N$ [5, 40, 44, 43]. This transforms the original problem (1.5) into the joint optimization

$$\min_{\boldsymbol{\mu}\in\mathbb{R}^N, \boldsymbol{a}\in\mathbb{R}^M} \left(\sum_{m=1}^{M} \mathrm{E}([\mathbf{G}_{\boldsymbol{\mu}}\boldsymbol{a}]_m, y_m) + \lambda \boldsymbol{a}^T \mathbf{G}_{\boldsymbol{\mu}} \boldsymbol{a} + \mathrm{R}(\boldsymbol{\mu}) \right), \tag{1.12}$$

where $\mathbf{G}_{\boldsymbol{\mu}}$ is the Gram matrix of the learned kernel $\mathrm{k}_{\boldsymbol{\mu}}$ and $\mathrm{R}(\cdot)$ regularizes the coefficient vector $\boldsymbol{\mu}$, for example, like in $\mathrm{R}(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_{\ell_p} = (\sum_{n=1}^{N} |\mu_n|^p)^{\frac{1}{p}}$ for $1 \le p \le 2$ [4, 6, 29, 37, 38]. The learned function will then take the generic form

$$f(\cdot) = \sum_{n=1}^{N} \sum_{m=1}^{M} \mu_n a_m \mathrm{k}(\cdot, \boldsymbol{x}_m). \tag{1.13}$$

1.4. Our contribution. In this paper, we provide a Banach-space framework for supervised learning with gTV regularization. We study the topological structures of the search space of this problem, and we characterize the class of admissible regularization operators together with their associated kernel functions.

We also propose a multikernel extension of supervised learning with gTV regularization. To that end, we consider the minimization

$$\min_{\substack{f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d), \\ f = \sum_{n=1}^{N} f_n}} \left(\sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \sum_{n=1}^{N} \|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}} \right). \tag{1.14}$$

In this formulation, the target function f is decomposed into N additive components, where the smoothness of each component has been expressed by its corresponding regularization operator. Our main result, which follows from Theorem 4.1, is the existence of a solution of (1.14) that yields a multikernel expansion of the target function and that takes the form

$$f(\cdot) = \sum_{n=1}^{N} \sum_{l=1}^{M_n} a_{n,l} \mathrm{k}_n(\cdot, \boldsymbol{z}_{n,l}), \quad \|\boldsymbol{a}\|_{\ell_0} \leq M, \tag{1.15}$$

where $\|\boldsymbol{a}\|_{\ell_0}$ is called the ℓ_0 norm of $\boldsymbol{a}$ and is equal to the number of nonzero elements of $\boldsymbol{a}$, and k_n is the shift-invariant kernel associated to the operator L_n. Moreover, the total number of nonzero coefficients is upper-bounded by the number M of data points and, hence, is not growing with the number N of components. We also illustrate numerically the effect of using multiple kernels.

1.5. Roadmap. The paper is organized as follows: We present some mathematical preliminaries in section 2. In section 3, we study the Banach-space structure of the native spaces, and we characterize the class of admissible kernels. We propose and prove our main result in section 4. Finally, we provide further discussions and illustrations in section 5.

2. Preliminaries. In this section, we recall relevant mathematical concepts such as the function spaces that we use throughout the paper along with properties of linear operators that are defined over those spaces.

2.1. Function spaces. All the derivatives of a rapidly decaying function decay faster than the inverse of any polynomial at infinity. Then, a smooth and slowly growing function is an element of $\mathcal{C}^\infty(\mathbb{R}^d)$ such that all of its derivatives have asymptotic growth controlled by a polynomial. Finally, a heavy-tailed function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies $f(\boldsymbol{x}) \geq C(1 + \|\boldsymbol{x}\|)^\alpha$ for some finite constants $C, \alpha > 0$.

For $p \in [1, \infty)$, we denote by $L_p(\mathbb{R}^d)$ the Banach space of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ with finite L_p norm, i.e.,

$$L_p(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ measurable} : \|f\|_{L_p} \triangleq \left(\int_{\mathbb{R}^d} |f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{p}} < +\infty \right\}. \tag{2.1}$$

The Schwartz space of smooth and rapidly decaying functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ is denoted by $\mathcal{S}(\mathbb{R}^d)$. Its topological dual is $\mathcal{S}'(\mathbb{R}^d)$, the space of tempered distributions [30]. We remark that any smooth and slowly growing function $f : \mathbb{R}^d \to \mathbb{R}$ specifies the continuous linear functional $\varphi \mapsto \int_{\mathbb{R}^d} f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ over $\mathcal{S}(\mathbb{R}^d)$ and, hence, is an element of $\mathcal{S}'(\mathbb{R}^d)$.

The space of continuous functions over $\mathbb{R}^d$ that vanish at infinity is $\mathcal{C}_0(\mathbb{R}^d)$. It is a Banach space equipped with the supremum norm $\|\cdot\|_\infty$. The space of Schwartz functions $\mathcal{S}(\mathbb{R}^d)$ is densely embedded in $\mathcal{C}_0(\mathbb{R}^d)$. Hence, the topological dual of $\mathcal{C}_0(\mathbb{R}^d)$ can be defined as

$$\mathcal{M}(\mathbb{R}^d) = \left\{ w \in \mathcal{S}'(\mathbb{R}^d) : \quad \|w\|_{\mathcal{M}} \triangleq \sup_{\substack{\varphi \in \mathcal{S}(\mathbb{R}^d) \\ \|\varphi\|_\infty = 1}} |\langle w, \varphi \rangle| < +\infty \right\}. \tag{2.2}$$

In fact, $\mathcal{M}(\mathbb{R}^d)$ is the Banach space of bounded Radon measures over $\mathbb{R}^d$ equipped with the total-variation norm $\|\cdot\|_{\mathcal{M}}$ [47]. It includes the shifted Dirac impulses $\delta(\cdot - \boldsymbol{x}_0)$ with $\|\delta(\cdot - \boldsymbol{x}_0)\|_{\mathcal{M}} = 1$. Moreover, $L_1(\mathbb{R}^d) \subseteq \mathcal{M}(\mathbb{R}^d)$ with the relation $\|f\|_{L_1} = \|f\|_{\mathcal{M}}$ for all $f \in L_1(\mathbb{R}^d)$. This allows one to interpret $(\mathcal{M}(\mathbb{R}^d), \|\cdot\|_{\mathcal{M}})$ as a generalization of $(L_1(\mathbb{R}^d), \|\cdot\|_{L_1})$.

For a Banach space $\mathcal{X}$, we consider two topologies for its continuous dual space $\mathcal{X}'$. The first one is the strong topology. It is induced from the dual norm in the sense that a sequence $\{w_n\}_{n=0}^\infty \in \mathcal{X}'$ is said to converge in the strong topology to $w^* \in \mathcal{X}'$ if $\lim_{n\to\infty} \|w_n - w^*\|_{\mathcal{X}'} = 0$. The second one is the weak*-topology that comes from the predual space $\mathcal{X}$ in the sense that a sequence $\{w_n\}_{n=0}^\infty$ is said to converge in the weak*-topology to w^* if, for any element $\varphi \in \mathcal{X}$, $\{\langle w_n, \varphi \rangle\}_{n=0}^\infty$ converges to $\langle w^*, \varphi \rangle$.

Finally, let us mention that any Banach space $\mathcal{X}$ is isometrically isomorphic to a closed subspace of its second dual $\mathcal{X}'' = (\mathcal{X}')'$ (see, for example, [46, page 95]). For the sake of simplicity, we make the possible embedding mappings implicit in our framework. This leads to writing the latter proposition simply, via the inclusion $\mathcal{X} \subseteq \mathcal{X}''$. In this regard, a Banach space is reflexive if we have that $\mathcal{X} = \mathcal{X}''$. Typical examples of reflexive Banach spaces are $L_p(\mathbb{R}^d)$ spaces for $p \in (1, \infty)$. By contrast, the space $\mathcal{C}_0(\mathbb{R}^d)$ and, consequently, its dual $\mathcal{M}(\mathbb{R}^d)$ are not reflexive.

2.2. Linear operators. The linear operator $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ is called shift-invariant if, for any function $\varphi \in \mathcal{S}(\mathbb{R}^d)$ and any shift value $\boldsymbol{x}_0 \in \mathbb{R}^d$, we have that

$$\mathrm{L}\{\varphi(\cdot - \boldsymbol{x}_0)\} = \mathrm{L}\{\varphi\}(\cdot - \boldsymbol{x}_0). \tag{2.3}$$

We recall a variant of the celebrated Schwartz kernel theorem for linear and shift-invariant (LSI) operators (see [54] for a "simple" proof of the general case).

Theorem 2.1 (Schwartz kernel theorem). *For any LSI operator* $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$*, there exists a unique distribution* $h \in \mathcal{S}'(\mathbb{R}^d)$*, known as the impulse response of* L*, such that*

$$\forall \varphi \in \mathcal{S}(\mathbb{R}^d) : \mathrm{L}\{\varphi\}(\cdot) = \int_{\mathbb{R}^d} h(\cdot - \boldsymbol{y})\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \tag{2.4}$$

In this paper, we restrict ourselves to the class of continuous LSI operators that have an extended domain and are defined over the space of tempered distributions $\mathcal{S}'(\mathbb{R}^d)$. One can

fully characterize this class in the Fourier domain. The Fourier transform is a well-defined and continuous operator over $\mathcal{S}'(\mathbb{R}^d)$ and is denoted by $\mathcal{F} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$. Consequently, the frequency response of the LSI operator $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ is defined as the Fourier transform of its impulse response

$$\widehat{\mathrm{L}}(\boldsymbol{\omega}) \triangleq \mathcal{F}\{\mathrm{L}\{\delta\}\}(\boldsymbol{\omega}). \tag{2.5}$$

It is known that the frequency response of any continuous LSI operator over $\mathcal{S}'(\mathbb{R}^d)$ is a smooth and slowly growing function [51]. Additionally, any smooth and slowly growing function $\widehat{\mathrm{L}}(\cdot)$ defines an LSI and continuous operator $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ via

$$\mathrm{L}\{f\} = \mathcal{F}^{-1}\{\widehat{\mathrm{L}}\widehat{f}\}. \tag{2.6}$$

Typical examples of such operators are polynomials of derivative in dimension $d = 1$ and polynomials of the Laplacian operator for $d > 1$ [21].

3. Banach-space kernels. In this section, we introduce our Banach-space framework of learning with gTV regularization. We start by defining the class of kernel-admissible operators.

Definition 3.1. *The linear operator* $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *is called kernel-admissible (or simply admissible) if*

(i) *it is shift-invariant;*[1]
(ii) *it is an isomorphism over* $\mathcal{S}'(\mathbb{R}^d)$*, meaning that it is continuous and invertible, its inverse being the continuous operator* $\mathrm{L}^{-1} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$*;*
(iii) *the sampling functional* $\delta_{\boldsymbol{x}_0} : f \mapsto f(\boldsymbol{x}_0)$ *is weak*-continuous in the topology of its native space (see Definition 3.2 and Theorem 3.3).*

The restriction to LSI operators is not crucial to our framework. However, it lends itself to the convenience of an analysis in the Fourier domain. It also allows us to provide necessary and sufficient conditions to characterize the class of admissible operators (see Theorem 3.5). The invertibility assumption, on the other hand, is essential to have decaying kernels, that is, to have $\mathrm{k}(\boldsymbol{x} - \boldsymbol{y}) \to 0$ whenever $\|\boldsymbol{x} - \boldsymbol{y}\| \to \infty$. In fact, it is known that the Green's function of any LSI operator with a nontrivial null space necessarily has a singularity in the Fourier domain at the origin [65]. Finally, the assumption of the (weak*) continuity of the sampling functional is a natural choice in learning theory. The main motivation here is to guarantee the (weak*) lower semicontinuity of the global cost functional in (1.14). This can be used, together with the generalized Weierstrass theorem, to prove the existence of solutions (see Theorem 4.1). Let us note that the definition of weak*-continuity depends on the Banach structure of the native space. In what follows, we first properly define native spaces and then specify their underlying Banach structures.

Definition 3.2. *The native space of the LSI isomorphism* $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *is the preimage of* L *over the space of bounded Radon measures, that is, the space* $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \mathrm{L}^{-1}\{\mathcal{M}(\mathbb{R}^d)\}$.

[1] Although the notion of shift-invariant operators in (2.3) is defined for operators acting on Schwartz functions, one can extend it by duality to those whose domain is $\mathcal{S}'(\mathbb{R}^d)$. For more details on extension by duality, we refer to [65, section 3.3.2].

Theorem 3.3 summarizes the important properties of the native spaces. Its proof is available in Appendix A.

Theorem 3.3. *Let* $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *be an LSI isomorphism over* $\mathcal{S}'(\mathbb{R}^d)$. *Then, its native space is a topological vector space with the following properties:*

(i) *It is a Banach space equipped with the gTV norm*

$$\mathrm{gTV}(f) = \|f\|_{\mathcal{M}_\mathrm{L}} \triangleq \|\mathrm{L}\{f\}\|_{\mathcal{M}}. \tag{3.1}$$

(ii) *The restriction of* L *to its native space results in the isomorphism* $\mathrm{L} : \mathcal{M}_\mathrm{L}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d)$.

(iii) *The adjoint operator* L^* *is well-defined over* $\mathcal{C}_0(\mathbb{R}^d)$, *and its image is the Banach space* $\mathcal{C}_\mathrm{L}(\mathbb{R}^d)$ *with the norm* $\|f\|_{\mathcal{C}_\mathrm{L}} \triangleq \|\mathrm{L}^{-1*}\{f\}\|_\infty$.

(iv) *The space* $\mathcal{C}_\mathrm{L}(\mathbb{R}^d)$ *is the predual of* $\mathcal{M}_\mathrm{L}(\mathbb{R}^d)$, *meaning that* $(\mathcal{C}_\mathrm{L}(\mathbb{R}^d))' = \mathcal{M}_\mathrm{L}(\mathbb{R}^d)$.

(v) *The space of Schwartz functions is embedded in the native space. Moreover, the native space itself is densely embedded in the space of tempered distributions. The embedding hierarchy is indicated as*

$$\mathcal{S}(\mathbb{R}^d) \hookrightarrow \mathcal{M}_\mathrm{L}(\mathbb{R}^d) \overset{d.}{\hookrightarrow} \mathcal{S}'(\mathbb{R}^d). \tag{3.2}$$

We have summarized the Banach spaces and the mappings between them in Figure 1. Due to Theorem 3.3, the weak*-continuity of the sampling functional (condition (iii) in Definition 3.1) is equivalent to the inclusion of the shifted Dirac impulses in the predual of the native space. In other words, for all $\boldsymbol{x}_0 \in \mathbb{R}^d$, one should have that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{C}_\mathrm{L}(\mathbb{R}^d)$.

We now define the shift-invariant kernel associated to an admissible operator.

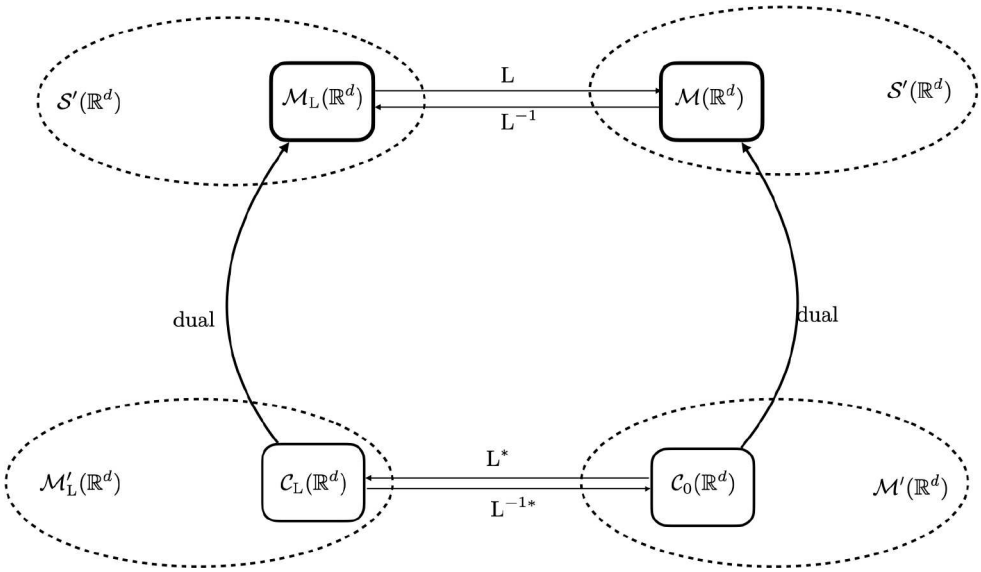


Figure 1. *A schematic diagram that illustrates the Banach spaces of interest.*

Definition 3.4. *The shift-invariant kernel associated to the admissible operator* $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *is the bivariate function* $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *with* $\mathrm{k}(\boldsymbol{x}, \boldsymbol{y}) = \rho_{\mathrm{L}}(\boldsymbol{x} - \boldsymbol{y})$, *where* $\rho_{\mathrm{L}} = \mathrm{L}^{-1}\{\delta\}$ *is the Green's function of* L.

In Theorem 3.5, we provide the necessary and sufficient conditions that characterize the class of admissible LSI operators. The proof can be found in Appendix B.

Theorem 3.5. *Let* L *be an admissible operator. Then, its associated Green's function* $\rho_{\mathrm{L}} = \mathrm{L}^{-1}\{\delta\} : \mathbb{R}^d \to \mathbb{R}$ *satisfies the following properties:*

(i) *It is a continuous function that vanishes at infinity. In other words,* $\rho_{\mathrm{L}} \in \mathcal{C}_0(\mathbb{R}^d)$.

(ii) *Its Fourier transform* $\widehat{\rho_{\mathrm{L}}}(\boldsymbol{\omega})$ *is a smooth, nonvanishing, slowly growing, and heavy-tailed function of* $\boldsymbol{\omega}$.

Additionally, any function $\rho : \mathbb{R}^d \to \mathbb{R}$ *that satisfies these properties can be appointed to an admissible operator* $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *defined as*

$$\mathrm{L}\{f\} = \mathcal{F}^{-1}\left\{\frac{\widehat{f}(\boldsymbol{\omega})}{\widehat{\rho}(\boldsymbol{\omega})}\right\}. \tag{3.3}$$

Remark 1. We have stated in section 2.2 a well-known result by Schwartz that determines the general family of LSI operators (not necessarily invertible) over $\mathcal{S}'(\mathbb{R}^d)$. In Theorem 3.5, particularly via condition (ii), we are excluding the noninvertible members of this family. Hence, condition (ii) fully characterizes the class of linear isomorphisms over $\mathcal{S}'(\mathbb{R}^d)$.

Using Theorem 3.5, we now draw a connection to the well-known class of reproducing kernels, which are constrained to be symmetric (because of their positive-definiteness).

Corollary 3.6. *Any symmetric admissible kernel (in the sense of Theorem 3.5) is a shift-invariant reproducing kernel up to multiplication by a sign factor.*

Proof. Let $\mathrm{k}(\cdot, \cdot)$ be a symmetric and shift-invariant admissible kernel. Then, the corresponding Green's function ρ_{L} is also a symmetric function, and, hence, its Fourier transform $\widehat{\rho}_{\mathrm{L}}(\boldsymbol{\omega})$ is a real function that is also smooth and nonvanishing. Hence, the sign of $\widehat{\rho}_{\mathrm{L}}(\boldsymbol{\omega})$ is constant everywhere. By multiplying with a sign factor, we can then assume that $\widehat{\rho}_{\mathrm{L}}(\boldsymbol{\omega})$ is positive everywhere. Now by invoking Bochner's theorem (see, for example, [65, Appendix B]), we deduce that ρ_{L} is a positive-definite function which, together with the symmetric assumption, implies that $\mathrm{k}(\cdot, \cdot)$ is indeed a reproducing kernel. ∎

The practical implication of Theorem 3.5 is that it yields Fourier-domain criteria to determine the admissibility of an operator L. In particular, and due to the Riemann–Lebesgue lemma, if $\widehat{\rho_{\mathrm{L}}}$ is an absolutely integrable function, then condition (i) holds.

As the last part of this section, we use this characterization to introduce some families of admissible kernels. Our first example is made of superexponential kernels defined as

$$\mathrm{k}_\alpha(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_\alpha^\alpha), \quad \alpha \in (0, 2), \tag{3.4}$$

where $\|\boldsymbol{x}\|_\alpha = (\sum_{i=1}^d |x_i|^\alpha)^{\frac{1}{\alpha}}$ for any $\boldsymbol{x} = (x_i) \in \mathbb{R}^d$. These functions are known to be positive-definite [65, Appendix B]. Their inverse Fourier transforms (the so-called α-stable distributions) are heavy-tailed and infinitely smooth, with algebraically decaying derivatives

of any order [48, Chapter 5]. Hence, they satisfy the conditions of Theorem 3.5. Note that the classical Gaussian kernels are excluded because their frequency responses are not heavy-tailed. However, one can get arbitrarily close by letting α tend to its critical value 2. Moreover, there are arguments in regularized RKHS that support the use of Gaussian kernels. For example, in [31, 55, 70], the Gaussian RKHS has been implicitly characterized by using the Taylor expansion of the corresponding regularization operator. Further, [61] uses the notion of holomorphic functions to explicitly characterize Gaussian RKHS. We conjecture that the present Banach-space formulation can be extended to cover Gaussian kernels as well. However, this requires one to consider a space larger than $\mathcal{S}'(\mathbb{R})$.

Our second example is made of Bessel potentials used in kernel estimation [2]. For a positive real number $s > d$, we consider the operator $(\mathrm{I} - \Delta)^{\frac{s}{2}} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$, where Δ is the Laplacian operator. The Bessel potentials are the Green's function of such operators. They correspond to the shift-invariant kernels

$$G_s(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{F}^{-1}\left\{ \frac{1}{(1 + \|\boldsymbol{\omega}\|_2^2)^{\frac{s}{2}}} \right\}(\boldsymbol{x} - \boldsymbol{y}). \tag{3.5}$$

Clearly, the function $\frac{1}{(1+\|\boldsymbol{\omega}\|_2^2)^{\frac{s}{2}}}$ is in $L_1(\mathbb{R}^d)$ for $s > d$. By invoking the Riemann–Lebesgue lemma, we deduce that its inverse Fourier transform is a continuous function that vanishes at infinity. Hence, the kernel function $G_s(\cdot, \cdot)$ satisfies property (i) of Theorem 3.5. Moreover, from the Fourier-domain definition (3.5) of $G_s(\cdot, \cdot)$, it can be seen that property (ii) also holds. Together, we deduce the admissibility of these kernels. We remark that the Bessel potential kernels are rotation-invariant as well.

Our final example is a general class of separable shift-invariant kernels of the form

$$\mathrm{k}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} \rho_{\mathrm{L}}(x_i - y_i), \tag{3.6}$$

where $\mathrm{L} : \mathcal{S}'(\mathbb{R}) \to \mathcal{S}'(\mathbb{R})$ is a stable rational operator whose frequency response is of the form $\widehat{\mathrm{L}}(\omega) = \frac{P(\omega)}{Q(\omega)}$, where P and Q are polynomials with no real roots such that $\deg(P) \geq \deg(Q)+2$. Since $\widehat{\mathrm{L}}(\omega)$ is real, we conclude that that the tail of $\widehat{\mathrm{L}}(\omega)^{-1} = \frac{Q(\omega)}{P(\omega)}$ behaves like ω^{-2} and is absolutely integrable which, together with the Riemann–Lebesgue lemma, implies that $\rho_{\mathrm{L}} \in \mathcal{C}_0(\mathbb{R})$. The other conditions of Theorem 3.5 can be readily shown to be true so that any separable kernel of the form (3.6) is admissible to our theory.

It is worth mentioning that one can rotate and dilate any admissible kernel by considering an invertible mixture matrix $\mathbf{A}$ and by defining the transformed kernel as $\mathrm{k}(\mathbf{A}\boldsymbol{x}, \mathbf{A}\boldsymbol{y})$. One readily verifies that the transformed kernel also satisfies the conditions of Theorem 3.5 and, hence, is also admissible. In Figure 2, we have plotted the superexponential and Bessel-potential kernels in dimension $d = 1$ for different sets of parameters. It can be seen that the width and regularity of these kernels can be adjusted through their parameters. This can be exploited in our framework of learning with multiple kernels to benefit from this diversity. We shall illustrate this numerically in section 5.

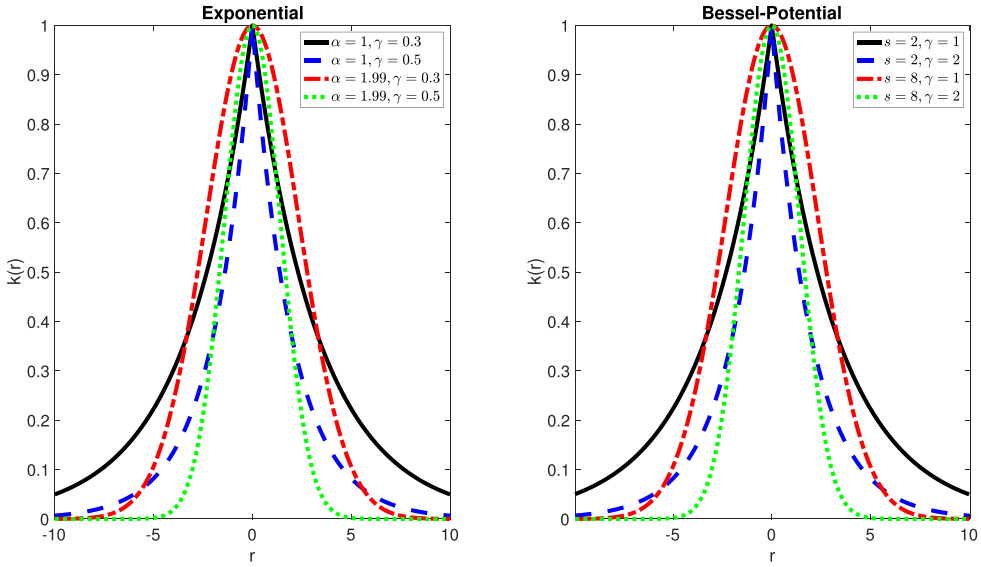


Figure 2. *Superexponential kernels* $\mathrm{k}_\alpha(\boldsymbol{r}) = \exp(-\gamma\|\boldsymbol{r}\|_\alpha^\alpha)$ *(left) and Bessel-potential kernels* $G_s(\gamma\boldsymbol{r})$ *(right), where* $\boldsymbol{r} = (\boldsymbol{x} - \boldsymbol{y})$*. The plots are in the special case* $d = 1$*. The parameters* ($\alpha \in (0, 2)$ *and* $s > 2$) *and* $\gamma > 0$ *adjust smoothness and width of the kernel, respectively.*

4. Multiple-kernel regression.

In this section, we prove our main result: the representer theorem of multiple-kernel regression with gTV regularization. In effect, the gTV norm will force the learned function to use the fewest active kernels.

Theorem 4.1 (multiple-kernel regression with gTV). *Given a training dataset that consists of* M *distinct pairs* $(\boldsymbol{x}_m, y_m)$ *for* $m = 1, 2, \ldots, M$*, we consider the minimization problem*

$$\min_{\substack{f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d),\\ f = \sum_{n=1}^N f_n}} \left(\sum_{m=1}^M \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \sum_{n=1}^N \|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}} \right), \tag{4.1}$$

where $\mathrm{E}(\cdot, y)$ *is a strictly convex nonnegative function and* L_n *is a kernel-admissible operator in the sense of Definition 3.1 for* $n = 1, 2, \ldots, N$*. Then, the solution set of this problem is nonempty, convex, and weak*-compact. For any of its extreme points* $(f_1, f_2, \ldots, f_N)$*, we have the kernel expansions*

$$f_n = \sum_{l=1}^{M_n} a_{n,l} \mathrm{k}_n(\cdot, \boldsymbol{z}_{n,l}), \quad n = 1, 2, \ldots, N \tag{4.2}$$

for its components, where $a_{n,l} \in \mathbb{R}$ *are kernel weights,* $\boldsymbol{z}_{n,l} \in \mathbb{R}^d$ *are adaptive kernel positions, and* $\mathrm{k}_n : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *is the shift-invariant kernel associated to the regularization operator* L_n *for* $n = 1, 2, \ldots, N$*. Moreover, the number of active kernels is upper-bounded by the number of data points, so that* $\sum_{n=1}^N M_n \leq M$*.*

Proof. Our proof is divided in three parts. First, we show the existence of a solution. Then, we show that (4.1) is equivalent to a constrained interpolation problem with fixed

function values, and, from this equivalent form, we deduce the topological properties of the solution set. Finally, we derive the form (4.2) for the extreme points of the solution set.

Let us denote the data-fidelity and regularization terms of the cost functional by $H(\cdot)$ and $R(\cdot)$, respectively, so that we have that

$$H(f_1,\dots,f_N)=\sum_{m=1}^{M}\mathrm{E}(f(\boldsymbol{x}_m),y_m),\quad f=\sum_{n=1}^{N}f_n, \tag{4.3}$$

$$R(f_1,\dots,f_N)=\sum_{n=1}^{N}\|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}}. \tag{4.4}$$

Part 1: Existence. We apply a standard technique in convex analysis. We show that the cost functional is coercive and weakly lower semicontinuous [39]. This also works when the latter property is replaced by weak* lower semicontinuity (see Proposition 8 in [33]).

The cost functional is a weighted sum of the nonnegative data-fidelity term $H(f_1,\dots,f_N)$ and the coercive regularization functional $R(f_1,\dots,f_N)$. This ensures its overall coercivity.

The sampling operator is weak*-continuous by assumption. Its composition with a continuous functional $\mathrm{E}(\cdot,\boldsymbol{y})$ (that follows from its strict convexity) and summation over m yields a cost functional $H(f)$ that is weak* lower semicontinuous as well.

The gTV norms $\|\mathrm{L}_n\cdot\|_{\mathcal{M}}$ are weak* lower semicontinuous on $\mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d)$. This implies that the regularization functional is weak* lower semicontinuous in the product space. Therefore, the overall cost functional $H(f_1,f_2,\dots,f_N)+\lambda R(f_1,\dots,f_N)$ is weak* lower semicontinuous. Together with the coercivity of the cost functional, this proves the existence of a solution.

Part 2: Equivalence to the constrained problem. Considering two solutions $(f_{1,1},\dots,f_{N,1})$ and $(f_{1,2},\dots,f_{N,2})$ of the problem, we denote their reconstructing functions by $f_i=\sum_{n=1}^{N}f_{n,i}$ for $i=1,2$. By contradiction assume that $f_1(\boldsymbol{x}_m)\neq f_2(\boldsymbol{x}_m)$ for some m. Since $\mathrm{E}(\cdot,y)$ is a strictly convex function for any $y\in\mathbb{R}$, we have that

$$H\left(\frac{f_{1,1}+f_{1,2}}{2},\dots,\frac{f_{N,1}+f_{N,2}}{2}\right)<\frac{H(f_{1,1},\dots,f_{N,1})+H(f_{1,2},\dots,f_{N,2})}{2}. \tag{4.5}$$

Similarly, the convexity of $R(\cdot)$ implies the inequality

$$R\left(\frac{f_{1,1}+f_{1,2}}{2},\dots,\frac{f_{N,1}+f_{N,2}}{2}\right)\leq\frac{R(f_{1,1},\dots,f_{N,1})+R(f_{1,2},\dots,f_{N,2})}{2}. \tag{4.6}$$

Together, the inequalities (4.5) and (4.6) imply that $(\frac{f_{1,1}+f_{1,2}}{2},\dots,\frac{f_{N,1}+f_{N,2}}{2})$ has a smaller cost than $(f_{1,i},\dots,f_{N,i})$ for $i=1,2$, which contradicts their optimality. Hence, $f_1(\boldsymbol{x}_m)=f_2(\boldsymbol{x}_m)=z_m$ for $m=1,2,\dots,M$, and one can rewrite the problem as

$$\min_{\substack{f_n\in\mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d),\\ f=\sum_{n=1}^{N}f_n}}\sum_{n=1}^{N}\|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}}\quad\text{s.t.}\quad f(\boldsymbol{x}_m)=z_m,\quad m=1,2,\dots,M. \tag{4.7}$$

Part 3: Identifying the solution set. Let us define $w_n=\mathrm{L}_n\{f_n\}$ for $n=1,\dots,N$ and $\nu_m(w_1,\dots,w_N)=\sum_{n=1}^{N}\langle\delta(\cdot-\boldsymbol{x}_m),\mathrm{L}_n^{-1}\{w_n\}\rangle=\sum_{n=1}^{N}f_n(\boldsymbol{x}_m)$ for $m=1,\dots,M$. We then reformulate (4.7) as

$$\min_{w_1,\dots,w_N\in\mathcal{M}} \sum_{n=1}^{N} \|w_n\|_{\mathcal{M}} \quad \text{s.t.} \quad \nu_m(w_1,\dots,w_N) = z_m, \quad m = 1,2,\dots,M. \tag{4.8}$$

Now, using the vector-valued Fisher–Jerome theorem (Appendix C), we deduce that the solution set of (4.8) is convex and weak*-compact with the extreme points of the form $\mathbf{w} = (w_1,\dots,w_N)$, where w_n takes the form

$$w_n = \sum_{l=1}^{M_n} a_{n,l}\delta(\cdot - \boldsymbol{z}_{n,l}) \tag{4.9}$$

for some $a_{n,l} \in \mathbb{R}$ and $\boldsymbol{z}_{n,l} \in \mathbb{R}^d$. Moreover, the total number of Diracs in $\mathbf{w}$ is upper-bounded by M. This implies that the solution set of (4.7) (and, consequently, the one of (4.1)) is a convex and weak*-compact set due to the linearity and isomorphism of L_n. Correspondingly, the extreme points of the original problem (4.1) take the form of $(f_1, f_2, \dots, f_N)$, where $f_n = \mathrm{L}_n^{-1}\{w_n\}$ has a kernel expansion with M_n kernels at adaptive positions subject to the constraint $\sum_{n=1}^{N} M_n \le M$. ∎

The practical outcome of Theorem 4.1 is that any extreme point of (4.1) maps into a solution of the form

$$f(\cdot) = \sum_{n=1}^{N}\sum_{l=1}^{M_n} a_{n,l}\mathrm{k}_n(\cdot, \boldsymbol{z}_{n,l}) \tag{4.10}$$

for the learned function. The solution form (4.10) has the following important properties:

- The number of active kernels is upper-bounded by the number of samples M. This justifies the use of multiple kernels since the flexibility of the model will be increased while the problem remains well-posed.
- The gTV norm enforces an ℓ_1 penalty on the kernel coefficients. Practically, this will result in an ℓ_1-minimization problem that is reminiscent of the generalized LASSO.
- The kernel positions are adaptive and will be chosen such that the solution becomes sparse. In other words, the adaptiveness of the kernel positions, together with the ℓ_1 regularization on the kernel coefficients, favors solutions with a small number of nonzero terms in the expansion (4.1).

To conclude this section, let us mention that the existence of the kernel locations $\mathbf{z}_{n,l}$ in (4.10) is guaranteed by our representer theorem. However, unlike in RKHS methods, these locations do not necessarily coincide with the data points. The adaptiveness comes from the fact that the kernel positions become part of the reduced finite-dimensional optimization problem (see (1.11) for the single-kernel scenario). Hence, an optimization scheme is required in order to "learn" these unknown parameters along with the kernel weights.

5. Discussion and illustration.

In this section, we provide some further discussions together with a numerical example that illustrates important aspects of our framework.

5.1. Optimization scheme.

Finding the kernel positions in general can be very challenging. Once the positions are fixed, one can find the kernel weights efficiently using classical

ℓ_1-minimization techniques [7, 14, 27]. In low dimensions, one can use grid-based algorithms and reach a solution where the positions of the kernels are quantified [17, 33]. It is also possible to adapt the algorithms developed for finding Dirac locations in super resolution in order to find the kernel positions [10, 11, 19, 26]. However, for high-dimensional problems, this is an open numerical challenge and requires further considerations. A possible avenue of research would be to use first-order primal-dual splitting methods for convex-nonconvex problems [66] and take advantage of the convexity of the problem with respect to the kernel weights.

5.2. Numerical example. In this section, we provide a numerical example in the case $d = 1$. We would like to emphasize that the computational aspects of our framework (e.g., the derivation of efficient algorithms in high dimensions) is left to future works. The sole purpose of our example is to illustrate the use of Theorem 4.1 and highlight two important features, namely, adaptivity and sparsity.

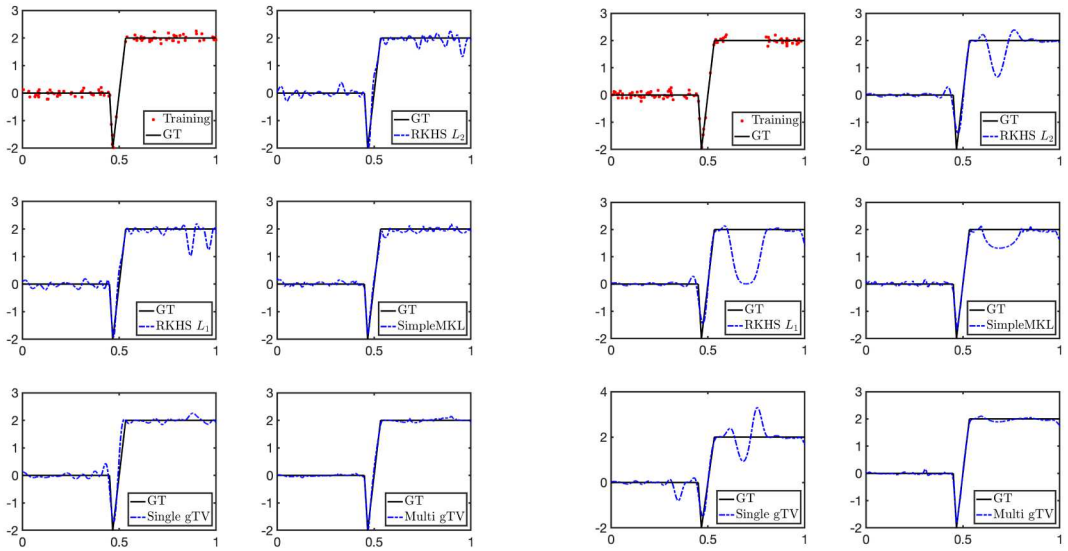
In our example, we compare the performance of five kernel estimators:

1. **RKHS $\mathbf{L_2}$:** RKHS regularization (1.5).
2. **RKHS $\mathbf{L_1}$:** Generalized LASSO (1.6).
3. **SimpleMKL:** Multiple-kernel learning (MKL) using the SimpleMKL algorithm [44].
4. **Single gTV:** Single-kernel gTV regularized learning (1.11).
5. **Multi gTV:** Learning with multiple kernels and gTV regularization (4.1).

To avoid the difficulty of optimizing over the data centers in the gTV-based methods, which would result in a nonconvex problem, we use a convex proxy in which a redundant set of centers is placed on a grid and the excess ones are suppressed with the help of ℓ_1-minimization. With this grid-based approach, the search for the kernel positions is reduced to a large-scale ℓ_1-minimization problem for which robust algorithms are known to exist—specifically, we have used a fast iterative shrinkage-thresholding algorithm (FISTA) [7] in our example. This scheme will obviously only work when the input dimension is very low, such as $d = 1$ in the present example.

We consider the reconstruction of a function from its noisy samples in two scenarios: full data versus missing data. The results are depicted in Figure 3, while we refer to Appendix D for the full implementation details. As we can see in Figure 3(a), due to the presence of a nonsmooth region in the target function, the single-kernel methods are forced to use narrow kernels with a small width which creates undesirable oscillations in the smoother regions. By contrast, our multikernel scheme uses both narrow and wide kernels, hence providing the reconstruction with the least fluctuation. In the presence of missing data, we observe in Figure 3(b) that the reconstructed function of RKHS-based methods exhibits an undesirable dip. This is due to the fact that, in the RKHS-based methods, the kernel functions are located on the data points and their width is too short to fill the gap in the data. By contrast, the kernel locations are adaptive in our scheme, which yields a decent reconstruction in this case as well. Finally, we have plotted the 100 largest kernel coefficients of each expansion in the full-data experiment in Figure 4. This plot highlights that the gTV-based methods are providing the sparsest representation for the target function, as expected.

The above visual observations are also supported quantitatively in Table 1, where we report the mean-squared error (MSE) error and sparsity (number of coefficients that are larger than one tenth of the maximum coefficient) of each method in the two scenarios.



(a) Full data (b) Missing data

Figure 3. *Performance of the kernel estimators in two scenarios: full data (left) or missing data (right). Solid line: ground-truth (GT) function. Dash-dotted line: reconstructed functions. Dots: noisy data points.*

5.3. Uniqueness of the solution. An interesting question is to explore the cases where perfect recovery is theoretically guaranteed. This is an open research topic on its own for which a rich literature exists [22, 18, 16]. Nevertheless, we analyze in Proposition 5.1 a very simple scenario for which we can prove uniqueness.

Proposition 5.1. *Let* $\mathrm{k}_1, \ldots, \mathrm{k}_N$ *be a collection of* N *symmetric admissible kernels that are normalized so that* $\mathrm{k}_n(\boldsymbol{x}, \boldsymbol{x}) = 1$ *for* $n = 1, \ldots, N$ *and* $\boldsymbol{x} \in \mathbb{R}^d$*. Consider the minimization*

$$\min_{\substack{f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d), \\ f = \sum_{n=1}^N f_n}} \sum_{n=1}^N \|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}} \quad s.t. \quad f(\boldsymbol{x}_m) = f_0(\boldsymbol{x}_m), \quad m = 1, \ldots, M, \tag{5.1}$$

where $f_0(\cdot) = a_0 \mathrm{k}_{n_0}(\cdot, \boldsymbol{z}_0)$ *for some* $n_0 \in \{1, \ldots, N\}$*,* $a_0 \in \mathbb{R}$*, and* $\boldsymbol{z}_0 \in \mathbb{R}^d$*. Assume that the set of data points* $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ *contains* $\boldsymbol{z}_0$ *and is such that the* M *by* N *matrix* $\mathbf{K} = [\mathrm{k}_n(\boldsymbol{x}_m, \boldsymbol{z}_0)]$ *has full column rank. Then,* f_0 *is the unique solution of* (5.1)*.*

Proof. From the proof of Theorem 4.1, we know that the solution set of (5.1) is the convex hull of functions of the form (4.10). We now show that the solution set has only one extreme point (that is, f_0), which is equivalent to the solution being unique.

Let f be an extreme point of the solution set of (5.1) whose form is given in (4.10). Since $\boldsymbol{z}_0$ is among the data points, we deduce that

$$a_0 = a_0 \mathrm{k}_{n_0}(\boldsymbol{z}_0, \boldsymbol{z}_0) = f_0(\boldsymbol{z}_0) = f(\boldsymbol{z}_0) = \sum_{n=1}^N \sum_{l=1}^{M_n} a_{n,l} \mathrm{k}_n(\boldsymbol{z}_0, \boldsymbol{z}_{n,l}).$$

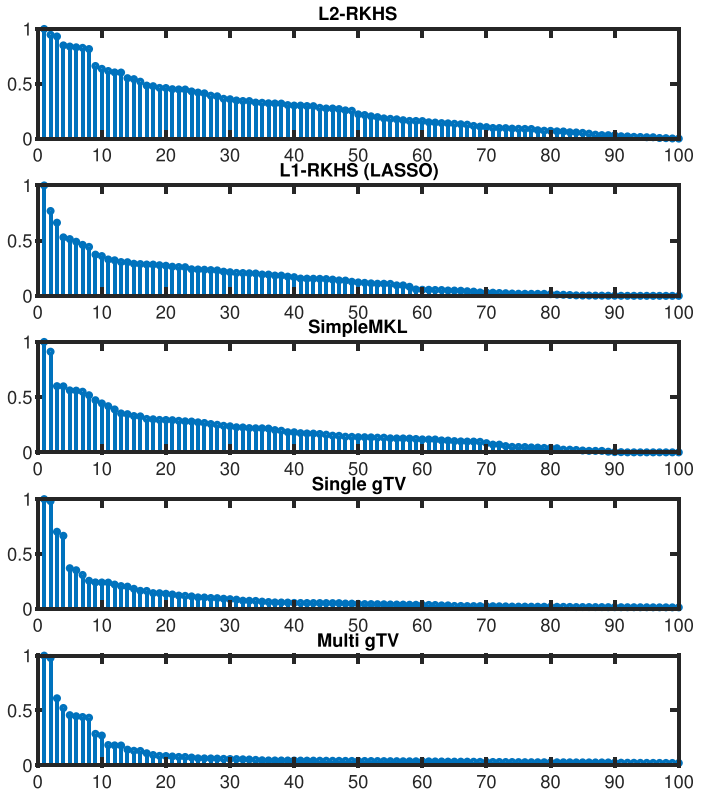


Figure 4. 100 *largest coefficients of each expansion in the full-data case.*

Table 1
MSE and sparsity of the kernel estimators. The results are averaged over 10 *runs.*

Quantity	Dataset	L2-RKHS	L1-RKHS	SimpleMKL	Single gTV	Multi gTV
Sparsity	Full data	64.7	44.1	54.4	32.5	**20.0**
	Missing data	66.1	39.3	56.0	32.9	**31.1**
MSE (dB)	Full data	-17.2	-16.1	-15.2	-16.7	**-18.1**
	Missing data	-2.6	-2.7	-10.9	-3.9	**-17.3**

Hence, by using the triangle inequality, we obtain that

$$|a_0| = \left| \sum_{n=1}^{N} \sum_{l=1}^{M_n} a_{n,l} \mathrm{k}_n(\boldsymbol{z}_0, \boldsymbol{z}_{n,l}) \right| \leq \sum_{n=1}^{N} \sum_{l=1}^{M_n} |a_{n,l}| |\mathrm{k}_n(\boldsymbol{z}_0, \boldsymbol{z}_{n,l})| \leq \sum_{n=1}^{N} \sum_{l=1}^{M_n} |a_{n,l}|,$$

where the last inequality comes from the fact that, for any positive-definite kernel k, we have that $\mathrm{k}(\boldsymbol{x}, \boldsymbol{y})^2 \leq \mathrm{k}(\boldsymbol{x}, \boldsymbol{x})\mathrm{k}(\boldsymbol{y}, \boldsymbol{y}) = 1$. Note that the positive-definiteness here is guaranteed by Corollary 3.6. This Cauchy–Schwarz-type inequality is saturated if and only if $\boldsymbol{x} = \boldsymbol{y}$. Together with the optimality of f, we deduce that $\boldsymbol{z}_0 = \boldsymbol{z}_{n,l}$ for all $n = 1, \ldots, N$ and $l = 1, \ldots, M_n$. Hence, we can rewrite the constraints as

$$\sum_{n=1}^{N} \tilde{a}_n \mathrm{k}_n(\boldsymbol{x}_m, \boldsymbol{z}_0) = a_0 \mathrm{k}_{n_0}(\boldsymbol{x}_m, \boldsymbol{z}_0), \quad m = 1, \dots, M,$$

where $\tilde{a}_n = \sum_{l=1}^{M_n} a_{n,l}$. In matrix form, this becomes $\mathbf{K}\tilde{\boldsymbol{a}} = \mathbf{K}a_0\mathbf{e}_{n_0}$, where $\mathbf{e}_{n_0} \in \mathbb{R}^N$ is the n_0th element of the canonical basis of $\mathbb{R}^N$. Finally, by using the full column rank assumption, we deduce that $\tilde{\boldsymbol{a}} = a_0\mathbf{e}_{n_0}$, which completes the proof. ∎

6. Conclusion. In this paper, we have provided a theoretical foundation for multiple-kernel regression with gTV regularization. We have studied the Banach structure of our search space and identified the class of kernel functions that are admissible. Then, we have derived a representer theorem that shows that the learned function can be written as a linear combination of kernels with adaptive centers. Our representer theorem also provides an upper bound to the number of active elements, which allows us to use as many kernels as convenient. We have illustrated numerically the effect of using multiple kernels with a sparsity constraint. Further research directions could be the development of efficient methods in high dimensions to approximate the kernel positions and an extension of the current theory to make Gaussian kernels admissible.

Appendix A. Proof of Theorem 3.3.

Proof. (i) The linearity and invertibility of L implies that the native space together with the gTV norm is a bona fide Banach space.

(ii) The restriction of L over its native space is injective (inherited from L) and is continuous due to the definition of the gTV norm. For all $w \in \mathcal{M}(\mathbb{R}^d)$, the relation $\mathrm{L}\{\mathrm{L}^{-1}\{w\}\} = w$ implies that it is surjective as well and that its inverse is the restriction of L^{-1} over $\mathcal{M}(\mathbb{R}^d)$ which continuously maps $\mathcal{M}(\mathbb{R}^d) \to \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$. This ensures that $\mathrm{L} : \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d)$ is an isomorphism.

(iii) The isomorphism of part (ii) implies the existence of the adjoint operator over $\big(\mathcal{M}(\mathbb{R}^d)\big)'$. By restricting the adjoint operator to $\mathcal{C}_0(\mathbb{R}^d)$, we obtain the operator $\mathrm{L}^* : \mathcal{C}_0(\mathbb{R}^d) \to \mathcal{C}_{\mathrm{L}}(\mathbb{R}^d)$, where the space $\mathcal{C}_{\mathrm{L}}(\mathbb{R}^d)$ is the image of L^* over $\mathcal{C}_0(\mathbb{R}^d)$. This space, equipped with the norm $\|f\|_{\mathcal{C}_{\mathrm{L}}} \triangleq \|\mathrm{L}^{-1*}\{f\}\|_\infty$, is a Banach space due to the linearity and invertibility of L^{-1*}.

(iv) Similarly to part (ii), we readily verify that the adjoint operator $\mathrm{L}^* : \mathcal{C}_0(\mathbb{R}^d) \to \mathcal{C}_{\mathrm{L}}(\mathbb{R}^d)$ is indeed an isomorphism. Therefore, the double-adjoint operator is the isomorphism $\mathrm{L}^{**} : (\mathcal{C}_{\mathrm{L}}(\mathbb{R}^d))' \to \mathcal{M}(\mathbb{R}^d)$. Consequently, the domains of L and L^{**} must be equal, which implies that $\mathcal{C}_{\mathrm{L}}(\mathbb{R}^d)$ is the predual of the native space.

(v) First, we show that the operator L is closed over the space of Schwartz functions. It is known that the impulse response of $\mathrm{L}^* : \mathcal{S} \to \mathcal{S}$ is the flipped version of the one of L [65]. In other words, the application of L^* on a Schwartz function can be expressed by

$$\forall \varphi \in \mathcal{S}(\mathbb{R}^d) : \mathrm{L}^*\{\varphi\}(\cdot) = \int_{\mathbb{R}^d} h(\boldsymbol{x} - \cdot)\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{A.1}$$

where $h \in \mathcal{S}'(\mathbb{R}^d)$ is the impulse response of L, described in (2.4). By the change of variable $\boldsymbol{y} = (-\boldsymbol{x})$, one verifies that, for any $\varphi \in \mathcal{S}(\mathbb{R}^d)$, we have that

$$\mathrm{L}\{\varphi\} = \mathrm{L}^*\{\varphi^\vee\}^\vee, \tag{A.2}$$

where $\varphi^\vee$ is the flipped version of $\varphi \in \mathcal{S}(\mathbb{R}^d)$ with $\varphi^\vee(\boldsymbol{x}) = \varphi(-\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$. In effect, (A.2) shows that $\mathrm{L}\{\varphi\} \in \mathcal{S}(\mathbb{R}^d)$ for any $\varphi \in \mathcal{S}(\mathbb{R}^d)$.

Now, from the inclusions $\mathrm{L}\{\mathcal{S}(\mathbb{R}^d)\} \subseteq \mathcal{S}(\mathbb{R}^d)$ and $\mathcal{S}(\mathbb{R}^d) \subseteq \mathcal{M}(\mathbb{R}^d)$, we deduce that $\mathcal{S}(\mathbb{R}^d) \subseteq \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$. Moreover, $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \subseteq \mathcal{S}'(\mathbb{R}^d)$ by Definition 3.2. This verifies the inclusion $\mathcal{S}(\mathbb{R}^d) \subseteq \mathcal{M}(\mathbb{R}^d) \subseteq \mathcal{S}'(\mathbb{R}^d)$. To complete the proof, we need to show that the identity operators $id_1 : \mathcal{S}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d)$ and $id_2 : \mathcal{M}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ are continuous.

For a converging sequence of Schwartz functions $\varphi_n \overset{\mathcal{S}}{\to} \varphi$, the continuity of L implies that $\mathrm{L}\{\varphi_n\} \overset{\mathcal{S}}{\to} \mathrm{L}\{\varphi\}$. Since $\mathcal{S}(\mathbb{R}^d)$ is continuously embedded in $\mathcal{M}(\mathbb{R}^d)$, we have that $\mathrm{L}\{\varphi_n\} \overset{\mathcal{M}}{\to} \mathrm{L}\{\varphi\}$ and, consequently, that $\varphi_n \overset{\mathcal{M}_{\mathrm{L}}}{\to} \varphi$. This proves that the embedding is continuous, which is denoted by $\mathcal{S}(\mathbb{R}^d) \hookrightarrow \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$. Moreover, since the space $\mathcal{M}(\mathbb{R}^d)$ is continuously embedded in $\mathcal{S}'(\mathbb{R}^d)$, the convergence $\mathrm{L}\{\varphi_n\} \overset{\mathcal{M}}{\to} \mathrm{L}\{\varphi\}$ implies that $\mathrm{L}\{\varphi_n\} \overset{\mathcal{S}'}{\to} \mathrm{L}\{\varphi\}$. This proves that $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \overset{d.}{\hookrightarrow} \mathcal{S}'(\mathbb{R}^d)$. The latter continuous embedding is also dense due to the denseness of $\mathcal{S}(\mathbb{R}^d)$ in $\mathcal{S}'(\mathbb{R}^d)$ and the inclusion $\mathcal{S}(\mathbb{R}^d) \subseteq \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$. ∎

Appendix B. Proof of Theorem 3.5.

Proof. Assume that L is a kernel-admissible operator. The weak*-continuity of the sampling functional implies that the shifted Dirac impulses $\delta(\cdot - \boldsymbol{x}_0)$ should be included in the predual space $\mathcal{C}_{\mathrm{L}}(\mathbb{R}^d)$. Therefore, $\mathrm{L}^{-1*}\{\delta(\cdot - \boldsymbol{x}_0)\}$ should be in $\mathcal{C}_0(\mathbb{R}^d)$. Since the Green's functions of L and L^* are flipped versions of each other, we deduce that $\rho_{\mathrm{L}} = \mathrm{L}^{-1}\{\delta(\cdot - \boldsymbol{x}_0)\} \in \mathcal{C}_0(\mathbb{R}^d)$. For the second property, we recall that the continuity of $\mathrm{L}^{-1} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ implies the smoothness and slow growth of the Fourier transform of its frequency response. Hence, $\widehat{\rho_{\mathrm{L}}}(\omega)$ is smooth and slowly growing. Similarly, the continuity of L implies that $\frac{1}{\widehat{\rho_{\mathrm{L}}}(\omega)}$ is a smooth and slowly growing function as well. Thus, $\widehat{\rho_{\mathrm{L}}}(\omega)$ is nonvanishing and heavy-tailed.

For the converse, assume that the function ρ satisfies properties (i) and (ii) in Theorem 3.5. First, note that, if $f, g : \mathbb{R}^d \to \mathbb{R}$ are smooth and slowly growing functions and, moreover, g is nonzero and heavy-tailed, then

$$\frac{\partial}{\partial x_i}\left(\frac{f}{g}\right) = \frac{\frac{\partial f}{\partial x_i} g - \frac{\partial g}{\partial x_i} f}{g^2} \tag{B.1}$$

is a quotient whose numerator is a smooth and slowly growing function and whose denominator g^2 is a nonzero, heavy-tailed, smooth, and slowly growing function. Hence, the quotient itself is a smooth function whose growth is bounded by a polynomial. Based on this observation, one can deduce from induction that all the arbitrary-order derivatives of $\frac{1}{\widehat{\rho}(\boldsymbol{\omega})}$ can be expressed by a quotient with a slowly growing nominator and a heavy-tailed denominator. This shows that $\frac{1}{\widehat{\rho}(\boldsymbol{\omega})}$ is a smooth and slowly growing function as well. These properties ensure the existence of continuous LSI operators $\mathrm{L}, \tilde{\mathrm{L}} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ with the frequency responses $\frac{1}{\widehat{\rho}(\omega)}$ and $\widehat{\rho}(\omega)$, respectively. The one-to-one correspondence between an operator and its frequency response then yields that $\tilde{\mathrm{L}} = \mathrm{L}^{-1}$, from which we conclude that L is an isomorphism over $\mathcal{S}'(\mathbb{R}^d)$. Moreover, due to property (i), we know that the Green's function of L is in $\mathcal{C}_0(\mathbb{R}^d)$. Hence, the Green's function of L^* is also in $\mathcal{C}_0(\mathbb{R}^d)$ so that, for any $\boldsymbol{x}_0 \in \mathbb{R}^d$, we have that

$$\mathrm{L}^{-1*}\{\delta(\cdot - \boldsymbol{x}_0)\} = \mathrm{L}^{-1*}\{\delta\}(\cdot - \boldsymbol{x}_0) \in \mathcal{C}_0(\mathbb{R}^d). \tag{B.2}$$

In other words, $\delta(\cdot - \boldsymbol{x}_0) \in \mathrm{L}^*(\mathcal{C}_0(\mathbb{R}^d)) = \mathcal{C}_{\mathrm{L}}(\mathbb{R}^d)$, which shows that the sampling functionals are weak*-continuous. ∎

Appendix C. Vector-valued Fisher–Jerome theorem. Here, we propose and prove a generalization of the Fisher–Jerome theorem [28] for a vector of bounded Radon measures. The result is not deducible from the original theorem, but its proof is an adaptation of the scalar case (Theorem 7 in [64]). We denote the space of bounded Radon vector measures $(w_1, \ldots, w_N)$ by $\mathcal{M}(\mathbb{R}^d; \mathbb{R}^N)$, where each component $w_n \in \mathcal{M}(\mathbb{R}^d)$ is a bounded Radon measure. The total-variation norm of the vector $\mathbf{w} = (w_1, \ldots, w_N) \in \mathcal{M}(\mathbb{R}^d; \mathbb{R}^N)$ is defined by $\|\mathbf{w}\|_{\mathcal{M}} = \sum_{n=1}^N \|w_n\|_{\mathcal{M}}$.

Theorem C.1 (vector-valued Fisher–Jerome). *Let $\mathcal{B} = \mathcal{M}(\mathbb{R}^d; \mathbb{R}^N) \bigoplus \mathcal{N}$, where $\mathcal{N}$ is an N_0-dimensional normed space, and assume that $\mathrm{F} : \mathcal{B} \to \mathbb{R}^M$ is a linear and weak*-continuous functional ($M \geq N_0$) such that*

$$\exists B > 0 : \quad \forall \mathbf{p} \in \mathcal{N} \backslash \{\mathbf{0}\}, \quad B \leq \frac{\|\mathrm{F}(0, \mathbf{p})\|_2}{\|\mathbf{p}\|_{\mathcal{N}}} \tag{C.1}$$

and that the minimization problem

$$\mathcal{V} = \underset{(\mathbf{w}, \mathbf{p}) \in \mathcal{B}}{\arg\min} \|\mathbf{w}\|_{\mathcal{M}} \quad s.t. \quad \mathrm{F}(\mathbf{w}, \mathbf{p}) \in \mathcal{C} \tag{C.2}$$

is feasible for a convex and compact set $\mathcal{C} \subseteq \mathbb{R}^M$. Then, $\mathcal{V}$ is a nonempty, convex, weak-compact subset of $\mathcal{B}$ while the components of its extreme points $(w_1, w_2, \ldots, w_N, \mathbf{p})$ are all of the form*

$$w_n = \sum_{l=1}^{M_n} a_{n,l} \delta(\cdot - \boldsymbol{z}_{n,l}), \quad n = 1, 2, \ldots, N, \tag{C.3}$$

where $a_{n,l} \in \mathbb{R}$ and $\boldsymbol{z}_{n,l} \in \mathbb{R}^d$. Moreover, $\sum_{n=1}^N M_n \leq M$, and the minimum $\mathcal{M}$-norm obtained for the problem is equal to $\sum_{n=1}^N \sum_{l=1}^{M_n} |a_{n,l}|$.

Proof. The proof is in two parts. First, we show that the solution set is nonempty, weak*-compact, and convex. Then, we explore the form of its extreme points to complete the theorem.

Structure of the solution set. Consider a point in the feasible set, and denote it by $(\mathbf{w}_0, \mathbf{p}_0)$. Then, (C.2) is equivalent to the minimization

$$\mathcal{V} = \underset{(\mathbf{w}, \mathbf{p}) \in \mathcal{B}}{\arg\min} \|\mathbf{w}\|_{\mathcal{M}} \quad \text{s.t.} \quad \mathrm{F}(\mathbf{w}, \mathbf{p}) \in \mathcal{C}, \|\mathbf{w}\|_{\mathcal{M}} < \|\mathbf{w}_0\|_{\mathcal{M}}. \tag{C.4}$$

Since $\mathcal{C}$ is compact, $A = \max_{\boldsymbol{x} \in \mathcal{C}} \|\boldsymbol{x}\|_2$ will be a finite constant. Due to the linearity of $F(\cdot, \cdot)$ and due to the triangle inequality, for all $(\mathbf{w}, \mathbf{p})$ in the feasible set we have that

$$\|\mathbf{p}\|_{\mathcal{N}} \leq \frac{1}{B} \|F(0, \mathbf{p})\|_2 = \frac{1}{B} \|F(\mathbf{w}, \mathbf{p}) - F(\mathbf{w}, 0)\|_2 \leq \frac{1}{B}(A + \|w_0\|_{\mathcal{M}}). \tag{C.5}$$

The conclusion is that the feasible set of (C.4) is bounded. It is also weak*-closed due to the weak*-continuity of $F(\cdot,\cdot)$ and the closedness of $\mathcal{C}$. Hence, it is weak*-compact due to the Banach–Alaoglu theorem [46, Theorem 3.15]. The conclusion is that (C.2) is equivalent to the minimization of a weak*-continuous functional over a weak*-compact domain. Moreover, due to the generalized Weierstrass theorem [39, Theorem 7.3.1], its solution set is nonempty. Denote the optimal cost of (C.2) by β. Now, note that the feasible set of (C.2) is the preimage of the linear continuous functional F over the convex set $\mathcal{C}$. So, one can rewrite the solution set $\mathcal{V}$ as

$$\mathcal{V} = F^{-1}(\mathcal{C}) \cap \{\mathbf{w} \in \mathcal{M}(\mathbb{R}^d;\mathbb{R}^N) : \|\mathbf{w}\|_{\mathcal{M}} = \beta\}. \tag{C.6}$$

This implies that $\mathcal{V}$ is bounded (due to (C.5)), weak*-closed, and convex (the intersection of two weak*-closed and convex set). Hence, it is also weak*-compact. Using the Krein–Milman theorem [46, Theorem 3.23], we deduce that $\mathcal{V}$ is the convex hull of its extreme points.

Form of the extreme points. Consider an arbitrary extreme point of $\mathcal{V}$ such as $(\mathbf{w},\mathbf{p})$, where $\mathbf{w} = (w_1, w_2, \dots, w_N)$. We show that it is not possible to have disjoint Borelian sets $E_{n,l} \subseteq \mathbb{R}^d$ such that $\langle w_n, \mathbb{1}_{E_{n,l}}\rangle \neq 0$, where $n = 1, 2, \dots, N$ and $l = 1, 2, \dots, M_n$ with $\sum_{n=1}^N M_n \geq M+1$. We prove the result by contradiction. Assume such disjoint sets exist. Define $v_{n,l} = w_n \mathbb{1}_{E_{n,l}}$, $\mathbf{v}_{n,l} = \boldsymbol{e}_n v_{n,l}$, $\overline{E}_n = (\bigcup_{l=1}^{M_n} E_{n,l})^{\mathrm{c}}$, $\overline{v}_n = w_n \mathbb{1}_{\overline{E}_n}$, and let $\overline{\mathbf{w}} = (\overline{v}_1, \overline{v}_2, \dots, \overline{v}_N)$. It can be seen that $\mathbf{w} = \overline{\mathbf{w}} + \sum_{n=1}^N \sum_{l=1}^{M_n} \mathbf{v}_{n,l}$. Define $\boldsymbol{y}_{n,l} = F(\mathbf{v}_{n,l}, \mathbf{p})$. Since the $\boldsymbol{y}_{n,l}$ are at least $M+1$ vectors in $\mathbb{R}^M$, they are linearly dependent. Consequently, there exist constants $\alpha_{n,l} \in \mathbb{R}$, with at least one of them being nonzero, such that

$$\sum_{n=1}^{N} \sum_{l=1}^{M_n} \alpha_{n,l} \boldsymbol{y}_{n,l} = \mathbf{0}. \tag{C.7}$$

For $n = 1, 2, \dots, N$, define $\mu_n = \sum_{l=1}^{M_n} \alpha_{n,l} v_{n,l}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_N)$. Also, denote $\epsilon_{\max} = \frac{1}{\max_{n,l} |\alpha_{n,l}|} > 0$. For any $\epsilon \in (-\epsilon_{\max}, \epsilon_{\max})$, we have that $1 + \epsilon \alpha_{n,l} > 0$ for all $n = 1, 2, \dots, N$ and $l = 1, 2, \dots, M_n$. We also see that

$$F(\boldsymbol{\mu}, \mathbf{p}) = \sum_{n=1}^{N} \sum_{l=1}^{M_n} \alpha_{n,l} \boldsymbol{y}_{n,l} = \mathbf{0}. \tag{C.8}$$

Now, for any $\epsilon \in (-\epsilon_{\max}, \epsilon_{\max})$, we have that $F(\mathbf{w} + \epsilon\boldsymbol{\mu}, \mathbf{p}) = F(\mathbf{w}, \mathbf{p}) \in \mathcal{C}$ and, therefore, $(\mathbf{w} + \epsilon\boldsymbol{\mu}, \mathbf{p}) \in \mathcal{U}$. Moreover,

$$\mathbf{w} + \epsilon\boldsymbol{\mu} = \overline{\mathbf{w}} + \sum_{n=1}^{N} \sum_{l=1}^{M_n} (1 + \epsilon\alpha_{n,l}) \mathbf{v}_{n,l}. \tag{C.9}$$

Note that the nth element of $\mathbf{w}_{\mathrm{c}}$ has support $E_{n,\mathrm{c}}$. Moreover, the nth element of $v_{n',l}$ has support $E_{n,l}$ for $n' = n$ and has empty support otherwise. Therefore, the nth entries have disjoint supports, which allows us to write that

$$\begin{aligned}\|\mathbf{w}+\epsilon\boldsymbol{\mu}\|_{\mathcal{M}} &= \sum_{n=1}^{N}\left\|\overline{v}_n+\sum_{l=1}^{M_n}(1+\epsilon\alpha_{n,l})\mathbf{v}_{n,l}\right\|_{\mathcal{M}}\\ &= \sum_{n=1}^{N}\|\overline{v}_n\|_{\mathcal{M}}+\sum_{n=1}^{N}\sum_{l=1}^{M_n}(1+\epsilon\alpha_{n,l})\|v_{n,l}\|_{\mathcal{M}}\\ &= \beta+\epsilon\sum_{n=1}^{N}\sum_{l=1}^{M_n}\alpha_{n,l}\|v_{n,l}\|_{\mathcal{M}}. \qquad \text{(C.10)}\end{aligned}$$

For sufficiently small values of ϵ, this gives either $\|\mathbf{w}+\epsilon\boldsymbol{\mu}\|_{\mathcal{M}} < \beta$ or $\|\mathbf{w}-\epsilon\boldsymbol{\mu}\|_{\mathcal{M}} < \beta$. Therefore, $\sum_{n=1}^{N}\sum_{l=1}^{M_n}\alpha_{n,l}\|v_{n,l}\|_{\mathcal{M}} = 0$, which yields that $\|\mathbf{w}+\epsilon\boldsymbol{\mu}\|_{\mathcal{M}} = \|\mathbf{w}-\epsilon\boldsymbol{\mu}\|_{\mathcal{M}} = \beta$. This shows that $(\mathbf{w}+\epsilon\boldsymbol{\mu},\mathbf{p}),(\mathbf{w}-\epsilon\boldsymbol{\mu},\mathbf{p})\in\mathcal{V}$, which contradicts that $(\mathbf{w},\mathbf{p})$ is an extreme point. Therefore $\mathbf{w}$, is nonzero at most in M points, which yields the form of (C.3). Computing the norm of such an extreme point results in

$$\|\mathbf{w}\|_{\mathcal{M}} = \sum_{n=1}^{N}\sum_{l=1}^{M_n}|a_{n,l}|\|\delta(\cdot-x_{n,l})\|_{\mathcal{M}} = \sum_{n=1}^{N}\sum_{l=1}^{M_n}|a_{n,l}|, \qquad \text{(C.11)}$$

which completes the proof. ∎

Appendix D. Implementation details of the numerical example. The ground-truth signal for our experiment is a piecewise linear function with four segments that connects five points, located at $\{(0,0),(0.45,0),(\frac{7}{15},-2),(\frac{8}{15},2),(1,2)\}$. We then sample data from the model $y_m = f(x_m)+\epsilon_m, m=1,\dots,M$, where $\epsilon_m\sim\mathcal{N}(0,\sigma^2)$ is i.i.d. Gaussian with $\sigma=0.1$. We formed two training datasets of size $M=100$. In the first one, x_m are i.i.d. samples of a uniform distribution over $[0,1]$. In the second case, we put a gap in the training dataset by sampling x_m uniformly over $[0,1]\backslash[0.6,0.8]$.

We use Gaussian kernels in the RKHS-based methods and superexponential kernels with $\alpha=1.99$ in the gTV-based methods. We have set $\alpha=1.99$ to have similar (near-Gaussian) kernel shapes in all cases. All methods have access to ten different width parameters from 10 to 10^5 in log scale.

We set the data fidelity to be the quadratic term $\mathrm{E}(x,y)=(x-y)^2$ in all cases except for MKL, since the SimpleMKL toolbox [44] uses the ϵ-insensitive SVM loss. The other methods are implemented using the GlobalBioIm library [56], and the codes are all available online.[2] In the gTV-based methods, we have used the multiresolution strategy of [17] to control the accuracy. More precisely, we start by considering 16 equispaced kernels, and we then use FISTA to solve the convex problem of finding the corresponding kernel coefficients. The solution is propagated as initialization of a finer grid (with 32 kernels), and we continue until we reach to the finest scale with 1,024 kernels.

Finally, to have a fair comparison, we optimize the hyperparameters of each method by following a standard K-fold cross-validation scheme, setting $K=5$ in our example. This includes a tuning of the regularization parameter λ for all methods. In addition, we tune the width of the kernel function in single-kernel schemes so that all methods have access to the

[2] https://github.com/Biomedical-Imaging-Group/Multi-Kernel-Regression-gTV-.

same family of kernel functions. For computing the test error, we consider a very fine grid with stepsize 10^{-4} over $[0, 1]$, and we compute the MSE between the learned function and the ground-truth signal.

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] N. Aronszajn and K. T. Smith, *Theory of Bessel potentials* I, Ann. Inst. Fourier, 11 (1961), pp. 385–475.

[3] F. Bach, *Breaking the curse of dimensionality with convex neural networks*, J. Mach. Learn. Res., 18 (2017), pp. 629–681.

[4] F. R. Bach, *Consistency of the group LASSO and multiple kernel learning*, J. Mach. Learn. Res., 9 (2008), pp. 1179–1225.

[5] F. R. Bach, G. Lanckriet, and M. Jordan, *Multiple kernel learning, conic duality, and the SMO algorithm*, in Proceedings of the Twenty-First International Conference on Machine Learning, 2004, pp. 41–48.

[6] J. Bazerque and G. Giannakis, *Nonparametric basis pursuit via sparse kernel-based learning*, IEEE Signal Process. Mag., 30 (2013), pp. 112–125.

[7] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[8] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer Science & Business Media, New York, 2011.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[10] K. Bredies and H. Pikkarainen, *Inverse problems in spaces of measures*, ESAIM Control Optim. Calc. Var., 19 (2013), pp. 190–218.

[11] E. Candès and C. Fernandez-Granda, *Super-resolution from noisy data*, J. Fourier Anal. Appl., 19 (2013), pp. 1229–1254.

[12] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Found. Comput. Math., 7 (2007), pp. 331–368.

[13] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, Cambridge, 2007.

[14] I. Daubechies, R. DeVore, M. Fornasier, and S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Comm. Pure Appl. Math., 63 (2010), pp. 1–38.

[15] C. De Boor and R. E. Lynch, *On splines and their minimum properties*, J. Math. Mech., 15 (1966), pp. 953–969.

[16] Y. De Castro and F. Gamboa, *Exact reconstruction using Beurling minimal extrapolation*, J. Math. Anal. Appl., 395 (2012), pp. 336–354.

[17] T. Debarre, J. Fageot, H. Gupta, and M. Unser, *B-spline-based exact discretization of continuous-domain inverse problems with generalized TV regularization*, IEEE Trans. Inform. Theory, 65 (2019), pp. 4457–4470.

[18] Q. Denoyelle, V. Duval, and G. Peyré, *Support recovery for sparse super-resolution of positive measures*, J. Fourier Anal. Appl., 23 (2017), pp. 1153–1194.

[19] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, *The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy*, Inverse Problems, 36 (2020), 014001.

[20] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer Science & Business Media, New York, 2013.

[21] J. Duchon, *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, in Constructive Theory of Functions of Several Variables, Springer, New York, 1977, pp. 85–100.

[22] V. Duval and G. Peyré, *Exact support recovery for sparse spikes deconvolution*, Found. Comput. Math., 15 (2015), pp. 1315–1355.

[23] M. Eberts and I. Steinwart, *Optimal regression rates for SVMs using Gaussian kernels*, Electron. J. Stat., 7 (2013), pp. 1–42.

[24] T. Evgeniou, M. Pontil, and T. Poggio, *Regularization networks and support vector machines*, Adv. Comput. Math., 13 (2000), 1.
[25] G. E. Fasshauer, F. J. Hickernell, and Q. Ye, *Solving support vector machines in reproducing kernel Banach spaces with positive definite functions*, Appl. Comput. Harmon. Anal., 38 (2015), pp. 115–139.
[26] C. Fernandez-Granda, *Super-resolution of point sources via convex programming*, Inf. Inference, 5 (2016), pp. 251–303.
[27] M. A. Figueiredo, R. D. Nowak, and S. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE J. Sel. Topics Signal Process., 1 (2007), pp. 586–597.
[28] S. D. Fisher and J. W. Jerome, *Spline solutions to L1 extremal problems in one and several variables*, J. Approx. Theory, 13 (1975), pp. 73–83.
[29] J. Gao, P. Kwan, and D. Shi, *Sparse kernel learning with LASSO and Bayesian inference algorithm*, Neural Netw., 23 (2010), pp. 257–264.
[30] I. M. Gelfand and G. E. Shilov, *Generalized Functions. Vol. 1, Properties and Operations*, Academic Press, New York, 1969.
[31] F. Girosi, M. Jones, and T. Poggio, *Priors, Stabilizers and Basis Functions: From Regularization to Radial, Tensor and Additive Splines*, Technical report, Massachusetts Institute of Technology, 1993.
[32] M. Gönen and E. Alpaydin, *Multiple kernel learning algorithms*, J. Mach. Learn. Res., 12 (2011), pp. 2211–2268.
[33] H. Gupta, J. Fageot, and M. Unser, *Continuous-domain solutions of linear inverse problems with Tikhonov versus generalized TV regularization*, IEEE Trans. Signal Process., 66 (2018), pp. 4670–4684.
[34] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer Science & Business Media, New York, 2006.
[35] T. Hastie, R. Tibshirani, and J. Friedman, *Overview of supervised learning*, in The Elements of Statistical Learning, Springer, New York, 2009, pp. 9–41.
[36] G. Kimeldorf and G. Wahba, *Some results on Tchebycheffian spline functions*, J. Math. Anal. Appl., 33 (1971), pp. 82–95.
[37] M. Kloft, U. Brefeld, P. Laskov, K. Müller, A. Zien, and S. Sonnenburg, *Efficient and accurate ℓ_p-norm multiple kernel learning*, in Advances in Neural Information Processing Systems, 2009, pp. 997–1005.
[38] M. Kloft, U. Rückert, and P. L. Bartlett, *A unifying view of multiple kernel learning*, in Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, New York, 2010, pp. 66–81.
[39] A. Kurdila and M. Zabarankin, *Convex Functional Analysis*, Springer Science & Business Media, New York, 2006.
[40] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, *Learning the kernel matrix with semidefinite programming*, J. Mach. Learn. Res., 5 (2004), pp. 27–72.
[41] E. Mammen and S. van de Geer, *Locally adaptive regression splines*, Ann. Statist., 25 (1997), pp. 387–413.
[42] S. Mendelson and J. Neeman, *Regularization in kernel learning*, Ann. Statist., 38 (2010), pp. 526–565.
[43] C. A. Micchelli and M. Pontil, *Learning the kernel function via regularization*, J. Mach. Learn. Res., 6 (2005), pp. 1099–1125.
[44] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, *SimpleMKL*, J. Mach. Learn. Res., 9 (2008), pp. 2491–2521.
[45] V. Roth, *The generalized LASSO*, IEEE Trans. Neural Netw., 15 (2004), pp. 16–28.
[46] W. Rudin, *Functional Analysis. International Series in Pure and Applied Mathematics*, McGraw-Hill, New York, 1991.
[47] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 2006.
[48] K. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, 1999.
[49] B. Schölkopf, R. Herbrich, and A. Smola, *A generalized representer theorem*, in Computational Learning Theory, Springer, New York, 2001, pp. 416–426.

[50] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2001.
[51] L. Schwartz, *Théorie des distributions*, vol. 2, Hermann, Paris, 1957.
[52] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.
[53] L. Shi, Y.-L. Feng, and D.-X. Zhou, *Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces*, Appl. Comput. Harmon. Anal., 31 (2011), pp. 286–302.
[54] B. Simon, *Distributions and their Hermite expansions*, J. Math. Phys., 12 (1971), pp. 140–148.
[55] A. Smola, B. Schölkopf, and K. Müller, *The connection between regularization operators and support vector kernels*, Neural Netw., 11 (1998), pp. 637–649.
[56] E. Soubies, F. Soulez, M. McCann, T.-a. Pham, L. Donati, T. Debarre, D. Sage, and M. Unser, *Pocket guide to solve inverse problems with GlobalBioIm*, Inverse Problems, 35 (2019), pp. 1–20.
[57] I. Steinwart, *Sparseness of support vector machines*, J. Mach. Learn. Res., 4 (2003), pp. 1071–1105.
[58] I. Steinwart, *Sparseness of support vector machines—Some asymptotically sharp bounds*, in Advances in Neural Information Processing Systems, 2004, pp. 1069–1076.
[59] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Science & Business Media, New York, 2008.
[60] I. Steinwart and A. Christmann, *Sparsity of SVMs that use the epsilon-insensitive loss*, in Advances in Neural Information Processing Systems, 2009, pp. 1569–1576.
[61] I. Steinwart, D. Hush, and C. Scovel, *An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4635–4643.
[62] I. Steinwart, D. R. Hush, and C. Scovel, *Optimal rates for regularized least squares regression*, in Proceedings of the Conference on Learning Theory, 2009, pp. 79–93.
[63] A. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 4 (1963), pp. 1035–1038.
[64] M. Unser, J. Fageot, and J. Ward, *Splines are universal solutions of linear inverse problems with generalized TV regularization*, SIAM Rev., 59 (2017), pp. 769–793.
[65] M. Unser and P. D. Tafti, *An Introduction to Sparse Stochastic Processes*, Cambridge University Press, Cambridge, 2014.
[66] T. Valkonen, *First-Order Primal-Dual Methods for Nonsmooth Non-Convex Optimisation*, preprint, arXiv:1910.00115 [math. OC], 2019.
[67] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
[68] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, PA, 1990.
[69] H.-Y. Wang, Q.-W. Xiao, and D.-X. Zhou, *An approximation theory approach to learning with ℓ_1 regularization*, J. Approx. Theory, 167 (2013), pp. 240–258.
[70] A. Yuille, *The motion coherence theory*, in Proceedings of the International Conference on Computer Vision, 1998, pp. 344–354.
[71] H. Zhang, Y. Xu, and J. Zhang, *Reproducing kernel Banach spaces for machine learning*, J. Mach. Learn. Res., 10 (2009), pp. 2741–2775.
[72] H. Zhang and J. Zhang, *Regularized learning in Banach spaces as an optimization problem: Representer theorems*, J. Global Optim., 54 (2012), pp. 235–250.