Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-020-09472-x

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

# A Unifying Representer Theorem for Inverse Problems and Machine Learning

**Michael Unser**[1]

Received: 30 November 2019 / Revised: 9 July 2020 / Accepted: 9 July 2020
© The Author(s) 2020

## Abstract

Regularization addresses the ill-posedness of the training problem in machine learning or the reconstruction of a signal from a limited number of measurements. The method is applicable whenever the problem is formulated as an optimization task. The standard strategy consists in augmenting the original cost functional by an energy that penalizes solutions with undesirable behavior. The effect of regularization is very well understood when the penalty involves a Hilbertian norm. Another popular configuration is the use of an $\ell_1$-norm (or some variant thereof) that favors sparse solutions. In this paper, we propose a higher-level formulation of regularization within the context of Banach spaces. We present a general representer theorem that characterizes the solutions of a remarkably broad class of optimization problems. We then use our theorem to retrieve a number of known results in the literature such as the celebrated representer theorem of machine leaning for RKHS, Tikhonov regularization, representer theorems for sparsity promoting functionals, the recovery of spikes, as well as a few new ones.

**Keywords** Convex optimization · Regularization · Representer theorem · Inverse problem · Machine learning · Banach space

**Mathematics Subject Classification** 46N10 · 47A52 · 65J20 · 68T05

Communicated by Tomaso Poggio.

The research leading to these results has received funding from the Swiss National Science Foundation under Grant 200020-184646/1.

✉ Michael Unser
michael.unser@epfl.ch

[1] Biomedical Imaging Group, Ecole polytechnique fédérale de Lausanne (EPFL), Station 17, CH-1015 Lausanne, Switzerland

Published online: 25 September 2020

## 1 Introduction

A recurrent problem in science and engineering is the reconstruction of a multi-dimensional signal $f : \mathbb{R}^d \to \mathbb{R}$ from a finite number of (possibly noisy) linear measurements $\boldsymbol{y} = (y_m) = \boldsymbol{v}(f) \in \mathbb{R}^M$, where the operator $\boldsymbol{v} = (v_m) : f \mapsto \boldsymbol{v}(f) = (\langle v_1, f\rangle, \ldots, \langle v_M, f\rangle)$ symbolizes the linear measurement process. The machine-learning version of the problem is the determination of a function $f : \mathbb{R}^d \to \mathbb{R}$ from a finite number of samples $y_m = f(\boldsymbol{x}_m) + \epsilon_m$ where $\epsilon_m$ is a small perturbation term; it is a special case of the former with $v_m = \delta(\cdot - \boldsymbol{x}_m)$. Since a function that takes values over the continuum is an infinite-dimensional entity, the reconstruction problem is inherently ill-posed.

The standard remedy is to impose an additional minimum-energy requirement which, in effect, regularizes the solution. A natural choice of regularization is a smoothness norm associated with some function space $\mathcal{X}'$ (typically, a Sobolev space), which results in the prototypical formulation of the problem as

$$S = \arg\min_{f \in \mathcal{X}'} \|f\|_{\mathcal{X}'} \quad \text{s.t.} \quad \langle v_m, f\rangle = y_m, \ m = 1, \ldots, M. \tag{1}$$

An alternative version that is better suited for noisy data is

$$S = \arg\min_{f \in \mathcal{X}'} \sum_{m=1}^{M} |y_m - \langle v_m, f\rangle|^2 + \lambda \|f\|_{\mathcal{X}'}^p \tag{2}$$

with an adequate choice of hyper-parameters $\lambda \in \mathbb{R}^+$ and $p \in [1, \infty)$. We note that the unconstrained form (2) is a generalization of (1): the latter is recovered in the limit by taking $\lambda \to 0$.

The term "representer theorem" is typically used to designate a parametric formula—preferably, a linear expansion in terms of some basis functions—that spans the whole range of solutions, irrespective of the value of the data $\boldsymbol{y} \in \mathbb{R}^M$. Representer theorems are valued by practitioners because they indicate the way in which the initial problem can be recast as a finite-dimensional optimization, making it amenable to numerical computations. The other benefit is that the description of the manifold of possible solutions provides one with a better understanding of the effect of regularization. The best-known example is the representer theorem for reproducing kernel Hilbert spaces (RKHS), which states that the solution of (2) with $\langle v_m, f\rangle = f(\boldsymbol{x}_m)$ and a Hilbertian regularization norm necessarily lives in a subspace of dimension $M$ spanned by kernels centered on the data coordinates $\boldsymbol{x}_m$ [7,31,35,36,43]. This theorem, in its extended version [42], is the foundation for the majority of kernel-based methods for machine learning, including regression, radial-basis functions, and support-vector machines [23,44,48]. There is also a whole line of generalizations of the concept that involves reproducing kernel Banach spaces (RKBS) [55–57]. More recently, motivated by the success of $\ell_1$ and total-variation regularization for compressed sensing [11,14,19], researchers have derived alternative representer theorems in order to explain the sparsifying effect of such penalties and their robustness to missing data [8,26,28,52]. A representer theorem for measures has also been invoked

to justify the use of the total-variation norm for the super-resolution localization of spikes [12,17,21,37] (see Sect. 4.1 for details).

In this paper, we present a unifying treatment of regularization by considering the problem from the abstract perspective of optimization in Banach spaces. Our motivation there is essentially twofold: (1) to get a better "geometrical" understanding of the effect of regularization, and (2) to state a generic representer theorem that applies to a wide variety of objects describable as elements of some native Banach space. The supporting theory is developed in Sect. 2. Our formulation takes advantage of the notion of Banach conjugates which is explained in Sect. 2.1. We then immediately proceed with the presentation of our key result: a generalized representer theorem (Theorem 2) that is valid for arbitrary convex data terms and Banach spaces in general, including the non-reflexive ones. The proof that is developed Sect. 2.2 is rather soft (or "high-level"), as it relies exclusively on the powerful machinery of duality mappings and the Hahn–Banach theorem—in other words, there is no need for Gâteaux derivatives nor subdifferentials, which are often invoked in such contexts. The resulting form of the solution in Theorem 2 is enlightening because it separates out the effect of the measurement operator from that of the regularization topology. Specifically, the measurement functionals $\nu_1, \ldots, \nu_M$ in (1) or (2) specify a linear solution manifold that is then isometrically mapped into the primary space via the conjugate map $\mathrm{J}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}'$, which may or may not be linear, depending on whether the regularization norm is Hilbertian or not.

The theory is then complemented with concrete examples of usage of Theorem 2 to illustrate the power of the approach as well as its broad range of applicability. Section 3 is devoted to the scenario where the regularization norm is strictly convex, which ensures that the solution of the underlying minimization problem is unique. We make the link with the existing literature by deriving of a number of classical results: Schölkopf's generalized representer theorem for RKHS (Sect. 3.1), the closed-form solution of continuous-domain Tikhonov regularization with a Hilbertian norm (Sect. 3.2), and the connection with the theory of reproducing kernel Banach spaces (Sect. 3.3). In addition, we present a novel representer theorem for $\ell_p$-norm regularization (Sect. 3.4). Then, in Sect. 4, we turn our attention to sparsity promoting regularization which is more challenging because the underlying Banach spaces are typically non-reflexive and non-convex. The enabling ingredient there is a recent result by Boyer *et al.* [8], which allows one to express the extreme points of the solution set in Theorem 2 as a linear combination of a few basic atoms that are selected adaptively (Theorem 3). This result, in its simplest incarnation with $\mathcal{X}' = \ell_1(\mathbb{Z})$, supports the well-documented sparsifying effect of $\ell_1$-norm minimization, which is central to the theory of compressed sensing. By switching to a continuum, we obtain the representer theorem for $\mathcal{X}' = \mathcal{M}(\Omega)$—the space of signed Radon measures on a compact domain $\Omega$ (Sect. 4.1), which is relevant to super-resolution localization. We then also derive a representer theorem for generalized total-variation (Sect. 4.2)—in the spirit of [53]—that justifies the use of sparse kernel expansions for machine learning, in line with the generalized LASSO [39].

## 2 Mathematical Formulation

### 2.1 Banach Spaces and Duality Mappings

The notion of Banach space—basically, a vector space equipped with a norm—is remarkably general. Indeed, the elements (or points) of a Banach space can be vectors (e.g., $\boldsymbol{v} \in \mathbb{R}^N$), functions (e.g., $f \in L_2(\mathbb{R}^d)$), sequences (e.g., $u[\cdot] \in \ell_1(\mathbb{Z})$), continuous linear functionals (e.g., $f \in \mathcal{X}'$ where $\mathcal{X}'$ is the dual of some primary Banach space), vector-valued functions (e.g., $\boldsymbol{f} = (f_1, \ldots, f_N)$ with $f_n \in L_2(\mathbb{R}^d)$), matrices ($e.g.$, $\mathbf{X} \in \mathbb{R}^{N\times N}$), and, even, bounded linear operators from a Banach space $\mathcal{U}$ (domain) to another Banach space $\mathcal{V}$ (range) [e.g., $\mathrm{X} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$] [32].

**Definition 1** A normed vector space $\mathcal{X}$ is a linear space equipped with a norm, henceforth denoted by $\|\cdot\|_{\mathcal{X}}$. It is called a Banach space if it is complete in the sense that every Cauchy sequence in $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ converges to an element of $\mathcal{X}$. It is said to be *strictly convex* if, for all $v_1, v_2 \in \mathcal{X}$ such that $\|v_1\|_{\mathcal{X}} = \|v_2\|_{\mathcal{X}} = 1$ and $v_1 \neq v_2$, one has that $\|\lambda v_1 + (1-\lambda)v_2\|_{\mathcal{X}} < 1$ for any $\lambda \in (0, 1)$. Finally, a Hilbert space is a Banach space whose norm is induced by an inner product.

We recall that $\mathcal{X}'$ (the continuous dual of $\mathcal{X}$) is the space of linear functionals $u : v \mapsto \langle u, v\rangle \triangleq u(v) \in \mathbb{R}$ that are continuous on $\mathcal{X}$. It is a Banach space equipped with the dual norm

$$\|u\|_{\mathcal{X}'} \triangleq \sup_{v\in\mathcal{X}\setminus\{0\}} \frac{\langle u, v\rangle}{\|v\|_{\mathcal{X}}}. \tag{3}$$

A direct implication of this definition is the generic duality bound

$$|\langle u, v\rangle| \leq \|u\|_{\mathcal{X}'}\|v\|_{\mathcal{X}}, \tag{4}$$

for any $u \in \mathcal{X}$, $v \in \mathcal{X}'$. In fact, (4) can be interpreted as the Banach generalization of the Cauchy-Schwarz inequality for Hilbert spaces. By invoking the Hahn–Banach theorem, one can also prove that the duality bound is sharp for any dual pair $(\mathcal{X}, \mathcal{X}')$ of Banach spaces [41]. This remarkable property inspired Beurling and Livingston to introduce the notion of duality mapping and to identify conditions of uniqueness [4]. We like to view the latter as the generalization of the classical Riesz map $\mathrm{R} : \mathcal{H}' \to \mathcal{H}$ or, rather, its inverse $\mathrm{J}_{\mathcal{H}} = \mathrm{R}^{-1} : \mathcal{H} \to \mathcal{H}'$, which describes the isometric isomorphism between a Hilbert space $\mathcal{H}$ and its continuous dual $\mathcal{H}'$ [38]. The caveat with Banach spaces is that the duality mapping is not necessarily bijective nor even single-valued.

**Definition 2** *(Duality mapping)* Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces. Then, the elements $v^* \in \mathcal{X}'$ and $v \in \mathcal{X}$ form a conjugate pair if they satisfy:

1. Norm preservation: $\|v^*\|_{\mathcal{X}'} = \|v\|_{\mathcal{X}}$, and
2. Sharp duality bound: $\langle v^*, v\rangle_{\mathcal{X}'\times\mathcal{X}} = \|v^*\|_{\mathcal{X}'}\|v\|_{\mathcal{X}}$.

For any given $v \in \mathcal{X}$, the set of admissible conjugates defines the duality mapping

$$\mathcal{J}_{\mathcal{X}}(v) = \{v^* \in \mathcal{X}' : \|v^*\|_{\mathcal{X}'} = \|v\|_{\mathcal{X}} \text{ and } \langle v^*, v\rangle_{\mathcal{X}'\times\mathcal{X}} = \|v^*\|_{\mathcal{X}'}\|v\|_{\mathcal{X}}\},$$

which is a nonempty subset of $\mathcal{X}'$. Whenever the duality mapping is a singleton (for instance, when $\mathcal{X}'$ is strictly convex), one also defines the corresponding duality operator $\mathrm{J}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}'$, which is such that $\mathcal{J}_{\mathcal{X}}(v) = \{v^* = \mathrm{J}_{\mathcal{X}}\{v\}\}$.

We now list the properties of the duality mapping that are relevant for our purpose [see [4], [15, Proposition 4.7 p. 27, Proposition 1.4, p. 43], [45, Theorem 2.53, p. 43]].

**Theorem 1** (Properties of duality mappings) *Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces. Then, the following holds:*

1. *Every $v \in \mathcal{X}$ admits at least one conjugate $v^* \in \mathcal{X}'$.*
2. *$\mathcal{J}_{\mathcal{X}}(\lambda v) = \lambda \mathcal{J}_{\mathcal{X}}(v)$ for any $\lambda \in \mathbb{R}$ (homogeneity).*
3. *For every $v \in \mathcal{X}$, the set $\mathcal{J}_{\mathcal{X}}(v)$ is convex and weak$^*$-closed in $\mathcal{X}'$.*
4. *The duality mapping is single-valued if $\mathcal{X}'$ is strictly convex; the latter condition is also necessary if $\mathcal{X}$ is reflexive.*
5. *When $\mathcal{X}$ is reflexive, the duality map is bijective if and only if both $\mathcal{X}$ and $\mathcal{X}'$ are strictly convex.*

The most favorable scenario is covered by Item 5. In that case, the duality map is invertible with $v = (v^*)^* = \mathrm{J}_{\mathcal{X}'}\mathrm{J}_{\mathcal{X}}\{v\}$; that is, $\mathrm{J}_{\mathcal{X}}^{-1} = \mathrm{J}_{\mathcal{X}'}$, in conformity with the property that $\mathcal{X}'' = \mathcal{X}$.

We now prove that the duality map is linear if and only if $\mathcal{X} = \mathcal{H}$ is a Hilbert space. In that case, the unitary operator $\mathrm{J}_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}'$ is precisely the inverse of the Riesz map $\mathrm{R} : \mathcal{H}' \to \mathcal{H}$.

**Proposition 1** *Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces such that $\mathcal{X}'$ is strictly convex. Then, the duality map $\mathrm{J}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}'$, $v \mapsto \mathrm{J}_{\mathcal{X}}\{v\} = v^*$ is linear if and only if $\mathcal{X}$ is a Hilbert space.*

***Proof*** First, we recall that all Hilbert spaces are strictly convex. Consequently, the indirect part of the statement is Riesz' celebrated representation theorem, which identifies the canonical linear isometry $\mathrm{J}_{\mathcal{X}} = \mathrm{R}^{-1}$ between a Hilbert space and its dual [41]. As for the converse implication, we show that the underlying inner product is

$$\langle u, v\rangle_{\mathcal{X}} = \tfrac{1}{2}\langle \mathrm{J}_{\mathcal{X}}\{u\}, v\rangle_{\mathcal{X}'\times\mathcal{X}} + \tfrac{1}{2}\langle \mathrm{J}_{\mathcal{X}}\{v\}, u\rangle_{\mathcal{X}'\times\mathcal{X}}. \tag{5}$$

Its bilinearity follows from the bilinearity of the duality product and the linearity of $\mathrm{J}_{\mathcal{X}}$, while the symmetry in $u$ and $v$ is obvious. Finally, the definition of the conjugate yields

$$\langle v, v\rangle_{\mathcal{X}} = \langle \mathrm{J}_{\mathcal{X}}\{v\}, v\rangle_{\mathcal{X}'\times\mathcal{X}} = \langle v^*, v\rangle_{\mathcal{X}'\times\mathcal{X}} = \|v\|_{\mathcal{X}}^2, \tag{6}$$

which confirms that the bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ is positive-semi-definite. Hence, it is the inner product that induces the $\|\cdot\|_{\mathcal{X}}$-norm. □

As an example, we provide the expression of the (unique) Banach conjugate $v^* = \mathrm{J}_{\mathcal{X}}\{v\} \in L_q(\mathbb{R}^d)$ of a function $v \in L_p(\mathbb{R}^d)\backslash\{0\}$ with $1 < p < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$:

$$v^*(\boldsymbol{x}) = \frac{|v(\boldsymbol{x})|^{p-1}}{\|v\|_{L_p}^{p-2}} \mathrm{sign}\big(f(\boldsymbol{x})\big). \tag{7}$$

This formula is intimately connected to Hölder's inequality. In particular, the $L_2$ conjugation map with $p = q = 2$ is an identity.

### 2.2 General Representer Theorem

We now make use of the powerful tool of conjugation to characterize the solution of a broad class of unconstrained optimization problems in Banach space. Let us note that the result also covers the equality constraint of Problem (1) if one selects the barrier functional

$$E_{\mathrm{equal}}(\boldsymbol{y}, \boldsymbol{z}) = \begin{cases} 0, & \boldsymbol{y} = \boldsymbol{z} \\ +\infty, & \text{otherwise.} \end{cases}$$

**Theorem 2** (General Banach representer theorem) *Let us consider the following setting:*

- *A dual pair $(\mathcal{X}, \mathcal{X}')$ of Banach spaces.*
- *The analysis subspace $\mathcal{N}_{\boldsymbol{v}} = \mathrm{span}\{v_m\}_{m=1}^M \subset \mathcal{X}$ with the $v_m$ being linearly independent.*
- *The linear measurement operator $\boldsymbol{v} : \mathcal{X}' \to \mathbb{R}^M : f \mapsto \big(\langle v_1, f\rangle, \ldots, \langle v_M, f\rangle\big)$ (it is weak* continuous on $\mathcal{X}'$ because $v_1, \ldots, v_M \in \mathcal{X}$).*
- *The loss functional $E : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}^+ \cup \{+\infty\}$ that is proper, weak lower semicontinuous and convex in its second argument.*
- *Some arbitrary strictly increasing and convex function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$.*

*Then, for any fixed $\boldsymbol{y} \in \mathbb{R}^M$, the solution set of the generic optimization problem*

$$S = \arg\min_{f \in \mathcal{X}'} E\big(\boldsymbol{y}, \boldsymbol{v}(f)\big) + \psi\left(\|f\|_{\mathcal{X}'}\right) \tag{8}$$

*is nonempty, convex, and weak*-compact. If $E$ is strictly convex (or if it imposes the equality $\boldsymbol{y} = \boldsymbol{v}(f)$), then any solution $f_0 \in S \subset \mathcal{X}'$ is a $(\mathcal{X}', \mathcal{X})$-conjugate of a common*

$$v_0 = \sum_{m=1}^{M} a_m v_m \in \mathcal{N}_{\boldsymbol{v}} \subset \mathcal{X} \tag{9}$$

*with a suitable set of weights $\boldsymbol{a} \in \mathbb{R}^M$, i.e., $S \subseteq \mathcal{J}_{\mathcal{X}}(v_0)$. Moreover, if $\mathcal{X}$ is strictly convex and $f \mapsto \psi(\|f\|_{\mathcal{X}'})$ is strictly convex, then the solution is unique with $f_0 =$*

$\mathrm{J}_{\mathcal{X}}\{v_0\} \in \mathcal{X}'$ *(Banach conjugate of* $v_0$*). In particular, if* $\mathcal{X}$ *is a Hilbert space, then* $f_0 = \sum_{m=1}^{M} a_m v_m^*$*, where* $v_m^*$ *is the Riesz conjugate of* $v_m$*.*

The condition of unicity requires the strict convexity of both $\psi : \mathbb{R}^+ \to \mathbb{R}$ and $f \mapsto \|f\|_{\mathcal{X}'}$. This applies to Banach spaces such as $\mathcal{X}' = \big(L_q(\mathbb{R}^d)\big)' = L_p(\mathbb{R}^d)$ (up to some isometric isomorphism) with $1 < p < \infty$ and the canonical choice of regularization $R(f) = \lambda \|f\|_{L_p}^p$ with $\psi(t) = \lambda |t|^p$ being strictly convex. While the solution of (8) also exists for Banach spaces such as $\mathcal{M}(\mathbb{R}^d) = \big(C_0(\mathbb{R}^d)\big)'$ or $L_\infty(\mathbb{R}^d) = \big(L_1(\mathbb{R}^d)\big)'$, the uniqueness is usually lost in such non-reflexive scenarios (see Sect. 4).

***Proof*** The proof uses standard arguments in convex analysis together with a dual reformulation of the problem inspired from the interpretation of best interpolation given by Carl de Boor in [6].

*(i) Existence and Reformulation as a Generalized Interpolation Problem*
First, we recall that the basic properties of (weak lower semi-) continuity and coercivity[1] are preserved though functional composition. The functional $f \mapsto \|f\|_{\mathcal{X}'}$ is convex, (norm-)continuous and coercive on $\mathcal{X}'$ from the definition of a norm. Since $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is strictly increasing and convex, it is necessarily continuous and coercive. This ensures that $f \mapsto \psi(\|f\|_{\mathcal{X}'})$ is endowed with the same three basic properties. The linear measurement operator $\boldsymbol{\nu} : \mathcal{X}' \to \mathbb{R}^N$ is continuous on $\mathcal{X}'$ by assumption (*i.e.,* $\nu_m \in \mathcal{X} \Rightarrow \nu_m \in \mathcal{X}''$ because of the canonical embedding of a Banach space in its bidual) and trivially convex. Since $\boldsymbol{z} \mapsto E\big(\boldsymbol{y}, \boldsymbol{z}\big)$ is lower semicontinuous on $\mathbb{R}^d$ and convex, this implies by composition the lower semicontinuity and convexity of $f \mapsto E\big(\boldsymbol{y}, \boldsymbol{\nu}(f)\big)$. Consequently, the functional $f \mapsto F(f) = E\big(\boldsymbol{y}, \boldsymbol{\nu}(f)\big) + \psi(\|f\|_{\mathcal{X}'})$ is (weakly) lower semicontinuous, convex, and coercive on $\mathcal{X}'$, which guarantees the existence of the solution (as well as the convexity and closedness of the solution set) by a standard argument in convex analysis [22]—see [28, Proposition 8] for the non-reflexive case. Moreover, unicity is ensured when $f \mapsto F(f)$ is strictly convex which happens to be the case when both $\boldsymbol{z} \mapsto E\big(\boldsymbol{y}, \boldsymbol{z}\big)$ and $f \mapsto \psi(\|f\|_{\mathcal{X}'})$ are strictly convex.
For the general (not necessarily unique) scenario, we take advantage of the strict convexity of $E(\boldsymbol{y}, \cdot)$ to show that all minimizers of $F(f)$ share a common measurement vector $\boldsymbol{z}_0 = \boldsymbol{\nu}(f_0) \in \mathbb{R}^M$. To that end, we pick any two distinct solutions $f_i \in S, i = 1, 2$ with corresponding measurements $\boldsymbol{z}_i = \boldsymbol{\nu}(f_i)$ and regularization cost $r_i = \lambda \psi(\|f_i\|_{\mathcal{X}'})$. The convexity of $S$ implies that, for any $\alpha \in (0, 1)$, $f = \alpha f_1 + (1 - \alpha) f_2 \in S$ with $\boldsymbol{z} = \boldsymbol{\nu}(f) = \alpha \boldsymbol{z}_1 + (1 - \alpha)\boldsymbol{z}_2$ and $F(f) = F(f_i), i = 1, 2$. Let us now assume that $\boldsymbol{z}_1 \neq \boldsymbol{z}_2$. Then, by invoking the strictly convexity of $\boldsymbol{z} \mapsto E(\boldsymbol{y}, \boldsymbol{z})$ and the convexity of $f \mapsto \lambda \psi(\|f\|_{\mathcal{X}'})$, we get

[1] The functional $F : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a Banach space, is said to be coercive if $F(f) \to \infty$ as $\|f\|_{\mathcal{X}} \to \infty$.

that

$$F(f) = E\big(\boldsymbol{y}, \alpha z_1 + (1-\alpha)z_2\big) + \lambda\psi\big(\|\alpha f_1 + (1-\alpha) f_2\|_{\mathcal{X}'}\big)$$
$$< \underbrace{\alpha E\,(\boldsymbol{y}, z_1) + (1-\alpha)E\,(\boldsymbol{y}, z_2) + \alpha r_1 + (1-\alpha)r_2}_{\alpha F(f_1)+(1-\alpha)F(f_2)=F(f)},$$

which is a contradiction. It follows that $z_i = \boldsymbol{\nu}(f_i) = z_0$ which, in turn, implies that the optimal regularization cost $r_i = r_0$ is the same for all $f_i \in S$. Although $z_0 = \boldsymbol{\nu}(f_0) \in \mathbb{R}^M$ is usually not known before hand, this property provides us with a convenient parametric characterization of the solution set as

$$S_z = \arg\min_{f\in\mathcal{X}'} \|f\|_{\mathcal{X}'} \text{ s.t. } \boldsymbol{\nu}(f) = z, \tag{10}$$

where $z$ ranges over $\mathbb{R}^M$. In this reformulation, we also exploit the property that the minimization of $\|f\|_{\mathcal{X}'}$ is equivalent to that of $\psi(\|f\|_{\mathcal{X}'})$ because the mapping between the two quantities is monotone.

*(ii) Explicit Resolution of the Generalized Interpolation Problem* (10)

The linear independence of the functionals $\nu_m$ ensures that any $\nu \in \mathcal{N}_{\boldsymbol{\nu}}$ has the unique expansion $\nu = \sum_{m=1}^M a_m \nu_m$. Based on this representation, we define the linear functional

$$\nu \mapsto \zeta(\nu) = \sum_{m=1}^M a_m z_m$$

with $z = z_0$ fixed. By construction, $\zeta$ is continuous $\big(\mathcal{N}_{\boldsymbol{\nu}}, \|\cdot\|_{\mathcal{X}}\big) \xrightarrow{\text{c.}} \mathbb{R}$ with $|\zeta(\nu)| \le \|\zeta\|\ \|\nu\|_{\mathcal{X}}$, where $\|\zeta\| = \sup_{\nu\in\mathcal{N}_{\boldsymbol{\nu}}:\ \|\nu\|_{\mathcal{X}}=1} \zeta(\nu) < \infty$. Moreover, the Hahn–Banach theorem ensures the existence of a continuous, norm-preserving extension of $\zeta$ to the whole Banach space $\mathcal{X}$; that is, an element $f_0 \in \mathcal{X}'$ such that

$$\|f_0\|_{\mathcal{X}'} = \sup_{g\in\mathcal{X}:\ \|g\|_{\mathcal{X}}=1} \langle f_0, g\rangle = \|\zeta\|.$$

The connection between the above statement and the generalized interpolation problem (10) is that the complete set of continuous extensions of $\zeta$ to $\mathcal{X} \supset \mathcal{N}_{\boldsymbol{\nu}}$ is given by

$$U = \{f \in \mathcal{X}' : \langle f, \nu\rangle = \zeta(\nu) \text{ for all } \nu \in \mathcal{N}_{\boldsymbol{\nu}}\}$$

with the property that

$$f_0 \in \arg\inf_{f\in U} \|f\|_{\mathcal{X}'} = S_{z_0} \quad\Leftrightarrow\quad \|f_0\|_{\mathcal{X}'} = \|\zeta\|. \tag{11}$$

The next fundamental observation is that $\mathcal{N}_{\boldsymbol{\nu}} = \big(\mathcal{N}_{\boldsymbol{\nu}}'\big)'$ because both spaces are of finite dimension $M$ and, hence, reflexive. Consequently, for any $\nu_0 \in \mathcal{J}_{\mathcal{X}}(\zeta) \subseteq$

$\left(\mathcal{N}_{\boldsymbol{\nu}}'\right)' = \mathcal{N}_{\boldsymbol{\nu}}$, we have that $\|\nu_0\|_{\mathcal{X}} = \|\zeta\|$ and $\zeta(\nu_0) = \|\nu_0\|_{\mathcal{X}}^2$, as well as $\|\nu_0\|_{\mathcal{X}} = \|f_0\|_{\mathcal{X}'}$ for all $f_0 \in S_{z_0}$ because of (11). Since $f_0 \in U \subset \mathcal{X}'$ and $\nu_0 \in \mathcal{N}_{\boldsymbol{\nu}} \subset \mathcal{X}$, this yields

$$\langle f_0, \nu_0 \rangle = \lambda(\nu_0) = \|f_0\|_{\mathcal{X}'} \|\nu_0\|_{\mathcal{X}},$$

which implies that $f_0 \in \mathcal{J}_{\mathcal{X}}(\nu_0)$ with $\mathcal{J}_{\mathcal{X}}$ the duality mapping from $\mathcal{X}$ to $\mathcal{X}'$.
*(iii) Structure of the Solution Set*
We have just shown that $S_{z_0} \subseteq \mathcal{J}_{\mathcal{X}}(\nu_0)$ for any extremal element $\nu_0 \in \{g \in \mathcal{N}_{\boldsymbol{\nu}} : \zeta(g) = \|\zeta\| \, \|g\|_{\mathcal{X}}, \|g\|_{\mathcal{X}} = \|\zeta\|\}$. We now deduce that $S_{z_0}$ is weak*-compact since it is included in the closed ball in $\mathcal{X}'$ of radius $\|f_0\|_{\mathcal{X}'} < \infty$, which is itself weak*-compact, by the Banach–Alaoglu theorem.
When $\mathcal{X}'$ is strictly convex, the situation is simpler because the duality mapping from $\mathcal{X}$ to $\mathcal{X}'$ is single-valued and the solution $f_0 \in \mathcal{X}'$ is unique. Moreover, the latter conjugate map is linear if and only if $\mathcal{X}$ is a Hilbert space, by Proposition 1. □

Note that the existence of the conjugate of $\nu_0 \in \mathcal{N}_{\boldsymbol{\nu}} \subset \mathcal{X}$ is essential to the argumentation. This is the reason why the problem is formulated with $f \in \mathcal{X}'$ subject to the hypothesis that $\nu_1, \ldots, \nu_M \in \mathcal{X}$ (weak* continuity). These considerations are inconsequential in the simpler reflexive scenario where the role of the two spaces is interchangeable since $\mathcal{X} = \mathcal{X}''$. The hypothesis of linear independence of the $\nu_m$ in Theorem 2 is only made for convenience. When it does not hold, one can adapt the proof by picking a basis of $\mathcal{N}_{\boldsymbol{\nu}}$ of reduced dimension $M' < M$, which then leads to a corresponding reduction in the number $M'$ of degrees of freedom of the solution.

In the sequel, as we shall apply Theorem 2 to concrete scenarios, we shall implicitly interpret $f \in \mathcal{X}'$ in (8) as a function (or, eventually, a vector) rather than a continuous linear functional on $\mathcal{X}$ (the abstract definition of an element of the dual space). This is acceptable provided that the defining space $\mathcal{X}'$ is isometrically embedded in some classical function spaces such as $L_p(\mathbb{R}^d)$ because of the bijective mapping (isometric isomorphism) that relates the two types of entities; for instance, there is a unique element of $f \in L_p(\mathbb{R})$ with $p$ the conjugate exponent of $q \in [1, \infty)$ such that the linear functional $\zeta \in \left(L_q(\mathbb{R}^d)\right)'$ can be specified as $\zeta(g) = \langle f, g \rangle = \int_{\mathbb{R}^d} f(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ and vice versa. This allows us to identify $\zeta = \zeta_f$ as $f \in L_p(\mathbb{R}^d)$, while it also gives a precise meaning to identities such as $L_p(\mathbb{R}^d) = \left(L_q(\mathbb{R}^d)\right)'$.

## 3 Strictly Convex Regularization

The solution of the optimization problem in Theorem 2 is unique whenever the Banach space $\mathcal{X}$ (or $\mathcal{X}'$) is reflexive and strictly convex. This is the setting that has been studied the most in the literature. We now illustrate the unifying character of Theorem 2 by using it to retrieve the key results in this area; that is, the classical kernel methods for machine learning in RKHS (Sect. 3.1), the resolution of linear inverse problems with Tikhonov regularization (Sect. 3.2), and the link with reproducing kernel Banach

spaces (Sect. 3.3). In addition, we make use of the conjugate map to present a novel perspective on $\ell_p$ regularization for $p > 1$ in Sect. 3.4.

### 3.1 Kernel/RKHS Methods in Machine Learning

Here, the search space $\mathcal{X}'$ is a reproducing kernel Hilbert space on $\mathbb{R}^d$ denoted by $\mathcal{H}$ with $\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}}$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the underlying inner product. The predual space is $\mathcal{X} = \mathcal{H}'$ which agrees with $\mathcal{X}' = \mathcal{H}'' = \mathcal{H}$ (reflexive scenario). The RKHS property [3] is equivalent to the existence of a (unique) positive-definite kernel $r_{\mathcal{H}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (the reproducing kernel of $\mathcal{H}$) such that

$$(i)\quad r_{\mathcal{H}}(\cdot, \boldsymbol{x}_m) \in \mathcal{H} \tag{12}$$

$$(ii)\quad f(\boldsymbol{x}_m) = \langle f, r_{\mathcal{H}}(\cdot, \boldsymbol{x}_m)\rangle_{\mathcal{H}} \tag{13}$$

for all $f \in \mathcal{H}$ and any $\boldsymbol{x}_m \in \mathbb{R}^d$.

In the context of machine learning, the loss function $E$ is usually chosen to be additive with $E(\boldsymbol{y}, \boldsymbol{z}) = \sum_{m=1}^{M} E_m\big(y_m, z_m\big)$ [29,43]. Given a series of data points $\big(\boldsymbol{x}_m, y_m\big)$, $m = 1, \ldots, M$ with $\boldsymbol{x}_m \in \mathbb{R}^d$, the learning problem is then to estimate a function $f_0 : \mathbb{R}^d \to \mathbb{R}$ such that

$$f_0 = \arg\min_{f\in\mathcal{H}} \left( \sum_{m=1}^{M} E_m\big(y_m, f(\boldsymbol{x}_m)\big) + \lambda \|f\|_{\mathcal{H}}^2 \right) \tag{14}$$

where $\lambda \in \mathbb{R}^+$ is an adjustable regularization parameter. In functional terms, the reproducing kernel represents the Schwartz kernel [27,46] of the Riesz map $\mathrm{R} : \mathcal{H}' \to \mathcal{H} : \nu \mapsto \nu^* = \int_{\mathbb{R}^d} r_{\mathcal{H}}(\cdot, \boldsymbol{y})\nu(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$ so that $\nu_m^*(\boldsymbol{x}) = \mathrm{R}\{\delta(\cdot - \boldsymbol{x}_m)\}(\boldsymbol{x}) = r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_m)$. The application of Theorem 2 with $\mathcal{X}' = \mathcal{H}$ then immediately yields the parametric form of the solution

$$f_0(\boldsymbol{x}) = \sum_{m=1}^{M} a_m r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_m), \tag{15}$$

which is a linear kernel expansion. The optimality of such kernel expansions is precisely the result stated in Schölkopf's representer theorem for RKHS [42]. Moreover, by invoking the reproducing kernel property (13) with $f = r_{\mathcal{H}}(\cdot, \boldsymbol{x}_n) \in \mathcal{H}$, one readily finds that $\|f_0\|_{\mathcal{H}}^2 = \boldsymbol{a}^T \mathbf{G}\boldsymbol{a}$, where the Gram matrix $\mathbf{G} \in \mathbb{R}^{M\times M}$ is specified by $[\mathbf{G}]_{m,n} = r_{\mathcal{H}}(\boldsymbol{x}_m, \boldsymbol{x}_n)$. By injecting the parametric form of the solution into the cost functional in (14), we then end up with the equivalent finite-dimensional minimization task

$$\boldsymbol{a}_0 = \arg\min_{\boldsymbol{a}\in\mathbb{R}^M} \Big( E\big(\boldsymbol{y}, \mathbf{G}\boldsymbol{a}\big) + \lambda \boldsymbol{a}^T \mathbf{G}\boldsymbol{a} \Big), \tag{16}$$

which yields the exact solution of the original infinite-dimensional optimization problem. In short, (16) is the optimal discretization of the functional optimization problem

(15), which is then readily transcribable into a numerical implementation using standard (finite-dimensional) techniques.

### 3.2 Tikhonov Regularization

Tikhonov regularization is a classical approach for dealing with ill-posed linear inverse problems [30,50]. The goal there is to recover a function $f : \mathbb{R}^d \to \mathbb{R}$ from a noisy or imprecise series of linear measurements $y_m = \langle \nu_m, f \rangle + \epsilon_m$, where $\epsilon_m$ is the disturbance term. By using the same functional framework as in Sect. 3.1 with $\nu_1, \ldots, \nu_M \in \mathcal{H}' = \mathcal{X}$, and $\mathcal{X}' = \mathcal{H}'' = \mathcal{H}$, one formulates the recovery problem as

$$f_0 = \arg\min_{f \in \mathcal{H}} \left( \sum_{m=1}^{M} |y_m - \langle \nu_m, f \rangle|^2 + \lambda \| f \|_{\mathcal{H}}^2 \right). \tag{17}$$

The application of Theorem 2 then yields a solution that takes the parametric form

$$f_0 = \sum_{m=1}^{M} a_m \varphi_m \tag{18}$$

with $\varphi_m = \mathrm{R}\{\nu_m\}$, where R is the Riesz map $\mathcal{H}' = \mathcal{X} \to \mathcal{H} = \mathcal{X}'$. The next fundamental observation is that the bilinear form $(\nu_m, \nu_n) \mapsto \langle \nu_m, \mathrm{R}\{\nu_n\} \rangle$ is actually the inner product for the dual space $\mathcal{H}'$ leading to $\langle \nu_m, \varphi_n \rangle = \langle \nu_m, \nu_n \rangle_{\mathcal{H}'}$. In fact, by using the property that $\nu_m$ and $\varphi_m = \nu_m^*$ are Hilbert conjugates, we have that

$$\langle \nu_m, \varphi_n \rangle = \langle \nu_m, \nu_n \rangle_{\mathcal{H}'} = \langle \nu_m^*, \nu_n^* \rangle_{\mathcal{H}} = \langle \varphi_m, \varphi_n \rangle_{\mathcal{H}} \tag{19}$$

which, somewhat remarkably, shows that the underlying system matrix is equal to the Gram matrix of the basis $\{\varphi_m\}$.

Therefore, by injecting (18) into the cost functional in (17), we are able to reformulate the initial optimization problem as the finite-dimensional minimization

$$\boldsymbol{a}_0 = \arg\min_{\boldsymbol{a} \in \mathbb{R}^M} \left( \| \boldsymbol{y} - \mathbf{H}\boldsymbol{a} \|^2 + \lambda \boldsymbol{a}^T \mathbf{H} \boldsymbol{a} \right), \tag{20}$$

where the system/Gram matrix $\mathbf{H} \in \mathbb{R}^{M \times M}$ with $[\mathbf{H}]_{m,n} = \langle \nu_m, \varphi_n \rangle = \langle \varphi_m, \varphi_n \rangle_{\mathcal{H}}$ is symmetric positive-definite. By differentiating the quadratic form in (20) with respect to $\boldsymbol{a}$ and setting the gradient to zero, we readily derive the very pleasing closed-form solution

$$\boldsymbol{a}_0 = (\mathbf{H}\mathbf{H} + \lambda \mathbf{H})^{-1} \mathbf{H} \boldsymbol{y} = (\mathbf{H} + \lambda \mathbf{I})^{-1} \boldsymbol{y} \tag{21}$$

under the implicit assumption that $\mathbf{H}$ is invertible. We note that the latter is equivalent to the linear independence of the $\varphi_m$ (resp., the linear independence of the $\nu_m$ due to the Riesz pairing).

### 3.3 Reproducing Kernel Banach Spaces

The concept of reproducing kernel Banach space, which is the natural generalization of RKHS, was introduced and investigated by Zhang and Xu in [55,56]. Similar to the Hilbertian case, one can identify the RKBS property as follows.

**Definition 3** A strictly convex and reflexive Banach space $\mathcal{B}$ of functions on $\mathbb{R}^d$ is called a reproducing kernel Banach space (RKBS) if $\delta(\cdot - \boldsymbol{x}) \in \mathcal{B}'$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Then, the unique representer $r_{\mathcal{B}}(\cdot, \boldsymbol{x}) = \mathrm{J}_{\mathcal{B}'}\{\delta(\cdot - \boldsymbol{x})\} \in \mathcal{B}$ when indexed by $\boldsymbol{x}$ is called the reproducing kernel of the Banach space.

It is then of interest to consider the Banach variant of (14) that involves a slightly more general regularization term: Given the data points $\left(\boldsymbol{x}_m, y_m\right)_{m=1}^M, m = 1, \ldots, M$, we want to find the unique solution of the optimization problem

$$f_0 = \arg\min_{f \in \mathcal{B}} \left( \sum_{m=1}^{M} E\left(y_m, f(\boldsymbol{x}_m)\right) + \psi(\|f\|_{\mathcal{B}}) \right) \tag{22}$$

where the loss function $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is convex in its second argument and the regularization strength modulated by the function $\psi : \mathbb{R} \to \mathbb{R}^+$, which is convex and strictly increasing. Since the space $\mathcal{B}$ is reflexive by assumption, the optimization problem falls into the framework of Theorem 2 with $\mathcal{X} = B'$ and $\mathcal{X}' = \mathcal{B}'' = \mathcal{B}$ and $v_m = \delta(\cdot - \boldsymbol{x}_m) \in \mathcal{B}', m = 1, \ldots, M$, where the latter inclusion is guaranteed by the RKBS property. We thereby obtain the parametric form of the solution as

$$f_0 = \mathrm{J}_{\mathcal{B}'} \left\{ \sum_{m=1}^{M} a_m \delta(\cdot - \boldsymbol{x}_m) \right\} = \mathrm{J}_{\mathcal{B}'} \left\{ \sum_{m=1}^{M} a_m r_{\mathcal{B}}^*(\cdot, \boldsymbol{x}_m) \right\} \tag{23}$$

with appropriate coefficients $(a_m) \in \mathbb{R}^M$, where the expression on the right-hand side has been included in order to make the connection with the Banach reproducing kernel, as in [56,57]. Due to the homogeneity and invertibility of the duality mapping (see Theorem 1), we have that $\mathrm{J}_{\mathcal{B}'}\left\{a_m r_{\mathcal{B}}^*(\cdot, \boldsymbol{x}_m)\right\} = a_m r_{\mathcal{B}}(\cdot, \boldsymbol{x}_m)$. This implies that (23) yields a *linear expansion* in terms of kernels *if and only* if $M = 1$ or if the duality map $\mathrm{J}_{\mathcal{B}} : \mathcal{B}' \to \mathcal{B}$ is linear. We note that the latter condition together with Definition 3 is equivalent to $\mathcal{B} = \mathcal{H}$ being a RKHS (by Proposition 1), which brings us back to the classical setting of Sect. 3.1. The same argumentation is also extendable to the vector-valued setting which has been considered by various authors both for RKHS and RKBS settings [1,33,58]. We also like to point out that our analysis is compatible with some recent results of Combettes et al. [16], where the corresponding conditions of optimality are stated using subdifferentials.

### 3.4 Toward Compressed Sensing: $\ell_p$-Norm Regularization

A classical problem in signal processing is to recover an unknown discrete signal $\boldsymbol{s} \in \mathbb{R}^N$ from a set of corrupted linear measurements $y_m = \boldsymbol{h}_m^T \boldsymbol{s} + \epsilon_m$,

$m = 1, \ldots, M$. The measurement vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M \in \mathbb{R}^N$ specify the system matrix $\mathbf{H} = [\boldsymbol{h}_1\ \boldsymbol{h}_2\ \cdots\ \mathbf{h}_M]^T \in \mathbb{R}^{M\times N}$. When $M$ (the number of measurements) is less than $N$ (the size of the unknown signal $\boldsymbol{s}$), the reconstruction problem is *a priori* ill-posed, and strongly so when $M \ll N$ (compressed-sensing scenario). However, if the original signal is known to be sparse (*i.e.,* $\|\boldsymbol{s}\|_0 \leq K_0$ with $K_0 < 2M$) and the system matrix $\mathbf{H}$ satisfies some "incoherence" properties, then the theory of compressed sensing provides general guarantees for a stable recovery [14,19,26]. The computational strategy then is to impose an $\ell_p$ regularization (with $p$ small to favor sparsity) on the solution and to formulate the reconstruction problem as

$$\boldsymbol{s} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \left( E\big(\boldsymbol{y}, \mathbf{H}\boldsymbol{x}\big) + \lambda \|\boldsymbol{x}\|_{\ell_p}^p \right) \tag{24}$$

with $\|\boldsymbol{x}\|_{\ell_p} \triangleq \left(\sum_{n=1}^N |x_n|^p\right)^{1/p}$. The traditional choice for compressed sensing is $p = 1$, which is the smallest exponent that still results in a convex optimization problem.

We now show how we can use Theorem 2 to characterize the effect of such a regularization for $p \in (1, \infty)$. The corresponding Banach space is $\mathcal{X}' = (\mathbb{R}^N, \|\cdot\|_{\ell_p})$ whose predual is $\mathcal{X} = (\mathbb{R}^N, \|\cdot\|_{\ell_q})$ with $\frac{1}{p} + \frac{1}{q} = 1$. Moreover, the underlying norms are strictly convex for $p > 1$, which guarantees that the solution is unique, irrespective of $M$ and $\mathbf{H}$. By introducing the dual signal $\boldsymbol{v}_0 = \mathbf{H}^T\mathbf{a} \in \mathcal{X}$ and by using the known form of the corresponding Banach $q$-to-$p$ duality map $\mathrm{J}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}'$, we then readily deduce that the solution can be represented as

$$[\boldsymbol{s}]_n = \frac{\big|[\mathbf{H}^T\boldsymbol{a}]_n)\big|^{q-1}}{\|\mathbf{H}^T\boldsymbol{a}\|_{\ell_q}^{q-2}} \mathrm{sign}\big([\mathbf{H}^T\boldsymbol{a}]_n\big) \tag{25}$$

for a suitable value of the (dual) parameter vector $\boldsymbol{a} \in \mathbb{R}^M$. While the exact value of $\boldsymbol{a}$ is data dependent, (25) provides us with the description of the solution manifold of intrinsic dimension $M$. Another way to put it is that the fact that $\boldsymbol{s}$ minimizes (24) induces a nonlinear pairing between the data vector $\mathbf{y} \in \mathbb{R}^M$ and the dual variable $\boldsymbol{a} \in \mathbb{R}^M$ in (25). In particular, for $p = 2$, we have that $\boldsymbol{s} = \mathbf{H}^T\boldsymbol{a} = \sum_{m=1}^M \boldsymbol{h}_m a_m$, which confirms the well-known result that $\boldsymbol{s} \in \mathrm{span}\{\mathbf{h}_m\}$. The latter also explains why classical quadratic/Tikhonov regularization performs poorly when $M$ is much smaller than $N$.

## 4 Sparsity-Promoting Regularization

The limit case[2] of the previous scenario is $p = 1$ (CS) for which the norm is no longer strictly convex. To deal with such cases where the solution is potentially non-unique, we first recall the Krein–Milman theorem [41, p. 75], which allows us to describe the

[2] Our analysis is not applicable to $p < 1$ because the corresponding metric no longer fulfills the properties of a norm; in other words, $\ell_p(\mathbb{Z})$ fails to be a Banach space for $p \in (0, 1)$.

weak*-compact solution set $S$ in Theorem 2 as the convex hull of its extreme points. We then invoke a recent result by Boyer et al. that yields the following characterization of the extremal points of Problem (8).

**Theorem 3** *All extremal points $f_{0,\text{ext}}$ of the solution set $S$ of Problem (8) can be expressed as*

$$f_{0,\text{ext}} = \sum_{k=1}^{K_0} a_k e_k \tag{26}$$

*for some $1 \leq K_0 \leq M$ where the $e_k$ are some extremal points of the unit "regularization" ball $B_{\mathcal{X}'} = \{f \in \mathcal{X}' : \|f\|_{\mathcal{X}'} \leq 1\}$ and $(a_k) \in \mathbb{R}^{K_0}$ is a vector of appropriate weights.*

The above is a direct corollary of [8, Theorem 1 with $j = 0$] applied to an extreme point of the equivalent generalized interpolation problem (10). We also note that the existence of a minimizer $f_0 \in S$ of the form (26) has been established independently by Bredies and Carioni [9] in a framework that is even more general than the one considered here. The latter property is also directly deducible from the reduced problem (10) and a classical result by Singer [47, Lemma 1.3, p. 169]. It remains that the existence of a global minimizer of the form (26) is not as strong a result as Theorem 3, which tells us the characterization applies for all extremal points of $S$. Moreover, it should be pointed out that the result in Theorem 3 is not particularly informative for strictly convex spaces such as $\ell_p(\mathbb{Z})$ or $L_p(\mathbb{R}^d)$ with $p \in (1, \infty)$ for which all unit vectors (*i.e.*, $e \in \mathcal{X}'$ with $\|e\|_{\mathcal{X}'} = 1$) are extremal points of the unit ball. Indeed, since the corresponding solution is unique (by Theorem 2), we trivially have that $f_0 = \|f_0\|_{\mathcal{X}'} e_1$ with $K_0 = 1$ and $e_1 = f_0/\|f_0\|_{\mathcal{X}'}$.

By contrast, the characterization in Theorem 3 is highly relevant for the non-strictly convex space $\mathcal{X}' = \ell_1(\mathbb{Z})$ whose extreme vectors are intrinsically sparse, i.e., $e_k = (\pm\delta[n - n_k])_{n\in\mathbb{Z}}$ for some fixed offset $n_k \in \mathbb{Z}$. Here, $\delta[\cdot]$ denotes the Kronecker impulse which is such that $\delta[0] = 1$ and $\delta[n] = 0$ for $n \neq 0$. Hence, the outcome is that the use of the $\ell_1$ penalty [e.g., (24) with $p = 1$] has a tendency to induce sparse solutions with $\|f\|_0 = K_0 \leq M$, which is the flavor of the representer theorem(s) in [52]. Two other practically relevant examples that fall in the non-strictly convex category are considered next.

### 4.1 Super-Resolution Localization of Spikes

The space of continuous functions over a compact domain $\Omega \subset \mathbb{R}^d$ equipped with the supremum (or $L_\infty$) norm is a classical Banach space denoted by

$$C(\Omega) = \{f : \Omega \to \mathbb{R} : \|f\|_\infty \triangleq \sup_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})| < \infty\}. \tag{27}$$

Its continuous dual

$$\mathcal{M}(\Omega) = \{f : C(\Omega) \to \mathbb{R} : \|f\|_{\mathcal{M}} \triangleq \sup_{\varphi \in C(\Omega):\, \|\varphi\|_\infty \leq 1} \langle f, \varphi \rangle < \infty\} \tag{28}$$

is the Banach space of bounded (signed) Radon measures on $\Omega$ (by the Riesz–Markov representation theorem [40]). Moreover, it is well known that the extreme points of the unit ball in $\mathcal{M}(\Omega)$ are point measures (a.k.a. Dirac impulses) of the form $e_k = \pm\delta(\cdot - \boldsymbol{x}_k)$ for some $\boldsymbol{x}_k \in \Omega$, with the property that

$$\varphi \mapsto \langle \delta(\cdot - \boldsymbol{x}_k), \varphi \rangle = \varphi(\boldsymbol{x}_k) \tag{29}$$

for any $\varphi \in C(\Omega)$. For a series of (independent) analysis functions $\nu_1, \ldots, \nu_M \in C(\Omega)$ (e.g., Fourier exponentials), we can invoke Theorems 2 and 3 with $\mathcal{X}' = \mathcal{M}(\Omega)$ to deduce that the extreme points of the problem

$$S = \arg \min_{f \in \mathcal{M}(\Omega)} \big(E\big(\boldsymbol{y}, \boldsymbol{\nu}(f)\big) + \lambda \|f\|_{\mathcal{M}}\big) \tag{30}$$

are inherently sparse. This means that there necessarily exists at least one minimizer of the form

$$f_0 = \sum_{k=1}^{K_0} a_k \delta(\cdot - \boldsymbol{x}_k) \tag{31}$$

with $K_0 \leq M$, $(a_k) \in \mathbb{R}^{K_0}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{K_0} \in \Omega$. The fact that (30) admits a global solution whose representation is given by (31) is a result that can be traced back to the work of Fisher and Jerome in [25, Theorem 1]. This optimality result is the foundation for a recent variational method for super-resolution localization that was investigated by a number of authors [10,12,24]. Besides the development of grid-free optimization schemes, researchers have worked out the conditions on $\boldsymbol{x}_k$ and $\nu_m$ under which (30) can provide a perfect recovery of spike trains of the form given by (31) with a small $K_0$ [13,17,37]. The remarkable finding is that there are many configurations for which super-resolution recovery is guaranteed, with an accuracy that only depends on the signal-to-noise ratio and the minimal spacing between neighboring spikes.

### 4.2 Sparse Kernel Expansions

Schwartz' space of smooth and rapidly decaying functions on $\mathbb{R}^d$ is denoted by $\mathcal{S}(\mathbb{R}^d)$. Its continuous dual is $\mathcal{S}'(\mathbb{R}^d)$: the space of tempered distributions. In this section, $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{S}'(\mathbb{R}^d)$ is an invertible operator with continuous inverse $\mathrm{L}^{-1} : \mathcal{S}'(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{S}'(\mathbb{R}^d)$. We also assume that the generalized impulse response of $\mathrm{L}^{-1}$ is a bivariate function of slow growth $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. In other words, the inverse operator $\mathrm{L}^{-1}$

has the explicit integral representation

$$\mathrm{L}^{-1}\{\varphi\} = \int_{\mathbb{R}^d} h(\cdot, \boldsymbol{y})\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \tag{32}$$

for any $\varphi \in \mathcal{S}(\mathbb{R}^d)$. In conformity with the nomenclature of [53], the native Banach space for $\big(\mathrm{L}, \mathcal{M}(\mathbb{R}^d)\big)$ is

$$\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \{f \in \mathcal{S}'(\mathbb{R}^d) : \|\mathrm{L}f\|_{\mathcal{M}} \triangleq \sup_{\varphi\in\mathcal{S}(\mathbb{R}^d):\, \|\varphi\|_\infty \leq 1} \langle \mathrm{L}f, \varphi\rangle < \infty\}. \tag{33}$$

It is isometrically isomorphic to $\mathcal{M}(\mathbb{R}^d)$ (the space of bounded Radon measures on $\mathbb{R}^d$). This is to say that the operators L, $\mathrm{L}^{-1}$ have restrictions $\mathrm{L} : \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{M}(\mathbb{R}^d)$ and $\mathrm{L}^{-1} : \mathcal{M}(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$ that are isometries. Consequently, we can apply Theorem 2 to deduce that the generic learning problem

$$S = \arg \min_{f\in\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)} \left( \sum_{m=1}^{M} E_m\big(y_m, f(\boldsymbol{x}_m)\big) + \lambda\|\mathrm{L}f\|_{\mathcal{M}} \right) \tag{34}$$

admits a solution, albeit not necessarily a unique one since the underlying search space $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$—or, equivalently, the parent space $\mathcal{M}(\mathbb{R}^d)$—is neither reflexive nor strictly convex.

In order to refine the above statement with the help of Theorem 3, we first observe that the extreme points of the unit ball in $\mathcal{M}(\mathbb{R}^d)$ take the form $e_k = \pm\delta(\cdot - \boldsymbol{\tau}_k)$ with $\boldsymbol{\tau}_k \in \mathbb{R}^d$, which is consistent with the result in Sect. 4.1 for $\mathcal{M}(\Omega)$. Since the map $\mathrm{L}^{-1} : \mathcal{M}(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$ is isometric, this allows us to identify the extreme points of the unit ball in $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$ as

$$u_k = \mathrm{L}^{-1}\{e_k\} = \pm\mathrm{L}^{-1}\{\delta(\cdot - \boldsymbol{\tau}_k)\} = \pm h(\cdot, \boldsymbol{\tau}_k) \tag{35}$$

where $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the kernel of the operator in (32). Consequently, we can invoke Theorem 3 to prove that the extreme points of Problem (34) are all expressible as

$$f_0(\boldsymbol{x}) = \sum_{k=1}^{K_0} a_k h(\boldsymbol{x}, \boldsymbol{\tau}_k) \tag{36}$$

with parameters $K_0 \leq M$, $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{K_0} \in \mathbb{R}^d$, and $(a_k) \in \mathbb{R}^{K_0}$. Moreover, since $\mathrm{L}\{h(\cdot, \boldsymbol{\tau}_k)\} = \delta(\cdot - \boldsymbol{\tau}_k)$ and $\|\delta(\cdot - \boldsymbol{\tau}_k)\|_{\mathcal{M}} = \|e_k\|_{\mathcal{M}} = 1$, the optimal regularization cost is $\|\mathrm{L}f_0\|_{\mathcal{M}} = \sum_{k=1}^{K_0} |a_k| = \|\boldsymbol{a}\|_{\ell_1}$, which makes an interesting connection with $\ell_1$-norm minimization and the generalized LASSO [39,49]. To sum up, the solution (36) has a kernel expansion that is similar to (15), with the important twist that the kernel centers $\boldsymbol{\tau}_k$ are adaptive, meaning that their location as well as their cardinality $K_0$ is data dependent. In effect, it is the underlying $\ell_1$-norm penalty that helps reducing

the number $K_0$ of active kernels, thereby producing a sparse solution. We should also point out that the form of the solution is compatible with the empirical method of moving and learning the data centers in kernel expansions [see [35, Section IV]] with the important difference that the present proposal is purely variational.

When $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{S}'(\mathbb{R}^d)$ is linear shift-invariant (LSI) with frequency response $\mathcal{F}\{\mathrm{L}\delta\}(\boldsymbol{\omega}) = \widehat{L}(\boldsymbol{\omega})$, then $h(\boldsymbol{x}, \boldsymbol{\tau}) = h_{\mathrm{LSI}}(\boldsymbol{x} - \boldsymbol{\tau})$ with

$$h_{\mathrm{LSI}}(\boldsymbol{x}) = \mathcal{F}^{-1}\left\{\frac{1}{\widehat{L}(\boldsymbol{\omega})}\right\}(\boldsymbol{x}), \tag{37}$$

where the operator $\mathcal{F}^{-1} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ is the generalized inverse Fourier transform.

The overarching message in the optimality result of the present section is that the choice of the regularization operator L in (34) predetermines the parametric form of the kernel in (36). Now, in light of (37), we can choose to specify first a kernel $h_{\mathrm{LSI}} : \mathbb{R}^d \to \mathbb{R}$ and then infer the frequency response of the corresponding regularization operator

$$\widehat{L}(\boldsymbol{\omega}) = \frac{1}{\widehat{h}_{\mathrm{LSI}}(\boldsymbol{\omega})}. \tag{38}$$

Now, the necessary and sufficient condition for the continuity of $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ is that the function $\widehat{L} : \mathbb{R}^d \to \mathbb{R}$ be smooth and slowly growing [46]. A parametric class of kernels that meets this admissibility requirement is the super-exponential family

$$h_{\mathrm{LSI}}(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x}\|^{\alpha}\right) \tag{39}$$

with $\alpha \in (0, 2)$. The limit case with $\alpha = 2$ (Gaussian) is excluded because the corresponding frequency response in (38) (inverse of a Gaussian) fails to be slowly increasing.

## 5 Conclusion

We have shown that the fundamental ingredient in the quest for a representer theorem is the identification and characterization of a dual pair of Banach spaces that is linked to the regularization functional. The main result of the paper is expressed by Theorem 2, which is valid for Banach spaces in general. This characterization of the solution of the general optimization problem (8) is directly exploitable in the reflexive and strictly convex scenario—in which case the solution is also known to be unique—whenever the duality mapping is known. While our formulation also offers interesting insights for certain non-strictly convex and sparsity-promoting norms such as $\|\cdot\|_{\ell_1}$ and its continuous-domain counterpart—the total variation $\|\cdot\|_{\mathcal{M}}$ and generalization thereof—it raises intriguing questions about the unicity of such solutions and the necessity to develop some corresponding numerical optimization schemes.

We have made the link with the existing literature in machine learning (regression) and the resolution of ill-posed inverse problems by considering several concrete cases, including reproducing kernel Hilbert spaces (RKHS) and compressed sensing. The conciseness and self-containedness of the proposed derivations are a good indication of the power of the approach.

Since the concept of Banach spaces is remarkably general, one can easily conceive of other variations around the common theme of regularization and representer theorems. Potential topics for further research include the use of nonstandard norms, the deployment of hybrid regularization schemes, vector-valued functions or feature maps [1], and the consideration of direct-sum spaces and semi-norms, as in the theory of splines [7,18,20,34,53,54]. In short, there is ample room for additional theoretical and practical investigations, in direct analogy with what has been accomplished during the past few decades in the simpler but more restrictive context of RKHS [1,2]. Interestingly, there also appears to be a link with deep neural/kernel networks, as has been demonstrated recently [5,51].

**Acknowledgements** The research was partially supported by the Swiss National Science Foundation under Grant 200020-184646/1. The author is thankful to Julien Fageot, Shayan Aziznejad, Pakshal Bohra, Quentin Denoyelle, and Philippe Thévenaz for helpful comments on the manuscript.

**Funding** Open access funding provided by EPFL Lausanne.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Alvarez, M.A., Rosasco, L., Lawrence, N.D.: Kernels for vector-valued functions: A review. Foundations and Trends in Machine Learning **4**(3), 195–266 (2012)
2. Argyriou, A., Micchelli, C.A., Pontil, M.: When is there a representer theorem? Vector versus matrix regularizers. Journal of Machine Learning Research **10**(Nov), 2507–2529 (2009)
3. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society **68**(3), 337–404 (1950)
4. Beurling, A., Livingston, A.E.: A theorem on duality mappings in Banach spaces. Arkiv för Matematik **4**(5), 405–411 (1962)
5. Bohn, B., Rieger, C., Griebel, M.: A representer theorem for deep kernel learning. Journal of Machine Learning Research **20**(64), 1–32 (2019)
6. de Boor, C.: On "best" interpolation. Journal of Approximation Theory **16**(1), 28–42 (1976)
7. de Boor, C., Lynch, R.E.: On splines and their minimum properties. Journal of Mathematics and Mechanics **15**(6), 953–969 (1966)
8. Boyer, C., Chambolle, A., De Castro, Y., Duval, V., De Gournay, F., Weiss, P.: On representer theorems and convex regularization. SIAM Journal of Optimization **29**(2), 1260–1281 (2019)
9. Bredies, K., Carioni, M.: Sparsity of solutions for variational inverse problems with finite-dimensional data. Calculus of Variations and Partial Differential Equations **59**(14), 26 (2020)
10. Bredies, K., Pikkarainen, H.K.: Inverse problems in spaces of measures. ESAIM: Control, Optimisation and Calculus of Variations **19**(1), 190–218 (2013)

11. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Review **51**(1), 34–81 (2009)
12. Candès, E.J., Fernandez-Granda, C.: Super-resolution from noisy data. Journal of Fourier Analysis and Applications **19**(6), 1229–1254 (2013)
13. Candès, E.J., Fernandez-Granda, C.: Towards a mathematical theory of super-resolution. Communications on Pure and Applied Mathematics **67**(6), 906–956 (2014)
14. Candès, E.J., Romberg, J.: Sparsity and incoherence in compressive sampling. Inverse Problems **23**(3), 969–985 (2007)
15. Cioranescu, I.: Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems, vol. 62. Springer Science & Business Media, (2012)
16. Combettes, P.L., Salzo, S., Villa, S.: Regularized learning schemes in feature Banach spaces. Analysis and Applications **16**(01), 1–54 (2018)
17. Denoyelle, Q., Duval, V., Peyré, G.: Support recovery for sparse super-resolution of positive measures. Journal of Fourier Analysis and Applications **23**(5), 1153–1194 (2017)
18. Dodu, F., Rabut, C.: Irrotational or divergence-free interpolation. Numerische Mathematik **98**, 477–498 (2004)
19. Donoho, D.L.: Compressed sensing. IEEE Transactions on Information Theory **52**(4), 1289–1306 (2006)
20. Duchon, J.: Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: W. Schempp, K. Zeller (eds.) Constructive Theory of Functions of Several Variables, pp. 85–100. Springer-Verlag, Berlin (1977)
21. Duval, V., Peyré, G.: Exact support recovery for sparse spikes deconvolution. Foundations of Computational Mathematics **15**(5), 1315–1355 (2015)
22. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems, *Classics in Applied Mathematics*, vol. 28. SIAM (1999)
23. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Advances in Computational Mathematics **13**(1), 1–50 (2000)
24. Fernandez-Granda, C.: Super-resolution of point sources via convex programming. Information and Inference **5**, 251–303 (2016)
25. Fisher, S.D., Jerome, J.W.: Spline solutions to $L_1$ extremal problems in one and several variables. Journal of Approximation Theory **13**(1), 73–83 (1975)
26. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Springer (2013)
27. Gelfand, I.M., Vilenkin, N.Y.: Generalized Functions. Vol. 4. Applications of Harmonic Analysis. Academic Press, New York, USA (1964)
28. Gupta, H., Fageot, J., Unser, M.: Continuous-domain solutions of linear inverse problems with Tikhonov *versus* generalized TV regularization. IEEE Transactions on Signal Processing **66**(17), 4670–4684 (2018)
29. Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. Annals of Statistics **36**(3), 1171–1220 (2008)
30. Karayiannis, N.B., Venetsanopoulos, A.N.: Regularization theory in image restoration—The stabilizing functional approach. IEEE Transactions on Acoustics, Speech and Signal Processing **38**(7), 1155–1179 (1990)
31. Kimeldorf, G., Wahba, G.: Some results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications **33**(1), 82–95 (1971)
32. Megginson, R.E.: An Introduction to Banach Space Theory. Springer (1998)
33. Micchelli, C.A., Pontil, M.: On learning vector-valued functions. Neural Computation **17**(1), 177–204 (2005)
34. Mosamam, A., Kent, J.: Semi-reproducing kernel Hilbert spaces, splines and increment kriging. Journal of Nonparametric Statistics **22**(6), 711–722 (2010)
35. Poggio, T., Girosi, F.: Networks for approximation and learning. Proceedings of the IEEE **78**(9), 1481–1497 (1990)
36. Poggio, T., Girosi, F.: Regularization algorithms for learning that are equivalent to multilayer networks. Science **247**(4945), 978–982 (1990)
37. Poon, C., Peyré, G.: Multidimensional sparse super-resolution. SIAM Journal on Mathematical Analysis **51**(1), 1–44 (2019)
38. Riesz, M.: Sur les fonctions conjuguées. Mathematische Zeitschrift **27**, 218–244 (1927)
39. Roth, V.: The generalized LASSO. IEEE Transactions on Neural Networks **15**(1), 16–28 (2004)

40. Rudin, W.: Real and Complex Analysis, 3rd edn. McGraw-Hill, New York (1987)
41. Rudin, W.: Functional Analysis, 2nd edn. McGraw-Hill, New York (1991). McGraw-Hill Series in Higher Mathematics
42. Schölkopf, B., Herbrich, R., Smola, A.J.: A generalized representer theorem. In: D. Helmbold, B. Williamson (eds.) Computational Learning Theory, pp. 416–426. Springer Berlin Heidelberg (2001)
43. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press (2002)
44. Schölkopf, B., Sung, K.K., Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V.: Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Transactions on Signal Processing **45**(11), 2758–2765 (1997)
45. Schuster, T., Kaltenbacher, B., Hofmann, B., Kazimierski, K.S.: Regularization Methods in Banach Spaces, vol. 10. Walter de Gruyter (2012)
46. Schwartz, L.: Théorie des Distributions. Hermann, Paris (1966)
47. Singer, I.: Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces. Springer-Verlag, Berlin (1970)
48. Steinke, F., Schölkopf, B.: Kernels, regularization and differential equations. Pattern Recognition **41**(11), 3271–32286 (2008)
49. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B **58**(1), 265–288 (1996)
50. Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Soviet Mathematics **4**, 1035–1038 (1963)
51. Unser, M.: A representer theorem for deep neural networks. Journal of Machine Learning Research **20**(110), 1–30 (2019)
52. Unser, M., Fageot, J., Gupta, H.: Representer theorems for sparsity-promoting $\ell_1$ regularization. IEEE Transactions on Information Theory **62**(9), 5167–5180 (2016)
53. Unser, M., Fageot, J., Ward, J.P.: Splines are universal solutions of linear inverse problems with generalized-TV regularization. SIAM Review **59**(4), 769–793 (2017)
54. Wahba, G.: Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia, PA (1990)
55. Xu, Y., Ye, Q.: Generalized Mercer kernels and reproducing kernel Banach spaces, vol. 258. Memoirs of the American Mathematical Society (2019)
56. Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel Banach spaces for machine learning. Journal of Machine Learning Research **10**, 2741–2775 (2009)
57. Zhang, H., Zhang, J.: Regularized learning in Banach spaces as an optimization problem: representer theorems. Journal of Global Optimization **54**(2), 235–250 (2012)
58. Zhang, H., Zhang, J.: Vector-valued reproducing kernel Banach spaces with applications to multi-task learning. Journal of Complexity **29**(2), 195–215 (2013)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.