The representer theorem for Hilbert spaces: a necessary and sufficient condition

Francesco Dinuzzo and Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Spemannstrasse 38,72076 Tübingen
Germany
[fdinuzzo@tuebingen.mpg.de, bs@tuebingen.mpg.de]

Abstract

The representer theorem is a property that lies at the foundation of regularization theory and kernel methods. A class of regularization functionals is said to admit a linear representer theorem if every member of the class admits minimizers that lie in the finite dimensional subspace spanned by the representers of the data. A recent characterization states that certain classes of regularization functionals with differentiable regularization term admit a linear representer theorem for any choice of the data if and only if the regularization term is a radial nondecreasing function. In this paper, we extend such result by weakening the assumptions on the regularization term. In particular, the main result of this paper implies that, for a sufficiently large family of regularization functionals, radial nondecreasing functions are the only lower semicontinuous regularization terms that guarantee existence of a representer theorem for any choice of the data.

1 Introduction

Regularization [1] is a popular and well-studied methodology to address ill-posed estimation problems [2] and learning from examples [3]. In this paper, we focus on regularization problems defined over a real Hilbert space $\mathcal{H}$. A Hilbert space is a vector space endowed with a inner product and a norm that is complete[1]. Such setting is general enough to take into account a broad family of finite-dimensional regularization techniques such as regularized least squares or support vector machines (SVM) for classification or regression, kernel principal component analysis, as well as a variety of methods based on regularization over reproducing kernel Hilbert spaces (RKHS).

The focus of our study is the general problem of minimizing an extended real-valued regularization functional $J : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ of the form

$$J(w) = f(L_1 w, \dots, L_\ell w) + \Omega(w), \tag{1}$$

where $L_1, \dots, L_\ell$ are bounded linear functionals on $\mathcal{H}$. The functional J is the sum of an *error term* f, which typically depends on empirical data, and a *regularization term* Ω that enforces certain desirable properties on the solution. By allowing the error term f to take the value $+\infty$, problems with hard constraints on the values $L_i w$ (for instance, interpolation problems) are included in the framework. Moreover, by allowing Ω to take the value $+\infty$, regularization problems of the Ivanov type are also taken into account.

In machine learning, the most common class of regularization problems concerns a situation where a set of data pairs (x_i, y_i) is available, $\mathcal{H}$ is a space of real-valued functions, and the objective functional to be minimized is of the form

$$J(w) = c\left((x_1, y_1, w(x_1)), \cdots, (x_\ell, y_\ell, w(x_\ell))\right) + \Omega(w).$$

[1] Meaning that Cauchy sequences are convergent.

It is easy to see that this setting is a particular case of (1), where the dependence on the data pairs (x_i, y_i) can be absorbed into the definition of f, and L_i are point-wise evaluation functionals, i.e. such that $L_i w = w(x_i)$. Several popular techniques can be cast in such regularization framework.

Example 1 (Regularized least squares)**.** *Also known as ridge regression when $\mathcal{H}$ is finite-dimensional. Corresponds to the choice*

$$c\left((x_1, y_1, w(x_1)), \cdots, (x_\ell, y_\ell, w(x_\ell)\right) = \gamma \sum_{i=1}^{\ell} (y_i - w(x_i))^2,$$

and $\Omega(w) = \|w\|^2$, where the complexity parameter $\gamma \geq 0$ controls the trade-off between fitting of training data and regularity of the solution.

Example 2 (Support vector machine)**.** *Given binary labels $y_i = \pm 1$, the SVM classifier (without bias) can be interpreted as a regularization method corresponding to the choice*

$$c\left((x_1, y_1, w(x_1)), \cdots, (x_\ell, y_\ell, w(x_\ell)\right) = \gamma \sum_{i=1}^{\ell} \max\{0, 1 - y_i w(x_i)\},$$

and $\Omega(w) = \|w\|^2$. The hard-margin SVM can be recovered by letting $\gamma \to +\infty$.

Example 3 (Kernel principal component analysis)**.** *Kernel PCA can be shown to be equivalent to a regularization problem where*

$$c\left((x_1, y_1, w(x_1)), \cdots, (x_\ell, y_\ell, w(x_\ell)\right) = \left\{ \begin{array}{ll} 0, & \frac{1}{\ell}\sum_{i=1}^{\ell}\left(w(x_i) - \frac{1}{\ell}\sum_{j=1}^{\ell} w(x_j)\right)^2 = 1 \\ +\infty, & \text{otherwise} \end{array} \right. ,$$

and Ω is any strictly monotonically increasing function of the norm $\|w\|$, see [4]. In this problem, there are no labels y_i, but the feature extractor function w is constrained to produce vectors with unitary empirical variance.

The possibility of choosing general continuous linear functionals L_i in (1) allows to consider a much broader class of regularization problems. Some examples are the following.

Example 4 (Tikhonov deconvolution)**.** *Given a "input signal" $u : \mathcal{X} \to \mathbb{R}$, assume that the convolution $u * w$ is well-defined for any $w \in \mathcal{H}$, and the point-wise evaluated convolution functionals*

$$L_i w = (u * w)(x_i) = \int_{\mathcal{X}} u(s) w(x_i - s) ds,$$

are continuous. A possible way to recover w from noisy measurements y_i of the "output signal" is to solve regularization problems such as

$$\min_{w \in \mathcal{H}} \left(\gamma \sum_{i=1}^{\ell} \left(y_i - (u * w)(x_i)\right)^2 + \|w\|^2 \right),$$

where the objective functional is of the form (1).

Example 5 (Learning from probability measures)**.** *In certain learning problems, it may be appropriate to represent input data as probability distributions. Given a finite set of probability measures $\mathbb{P}_i$ on a measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of $\mathcal{X}$, introduce the expectations*

$$L_i w = E_{\mathbb{P}_i}(w) = \int_{\mathcal{X}} w(x) d\mathbb{P}_i(x).$$

Then, given output labels y_i, one can learn a input-output relationship by solving regularization problems of the form

$$\min_{w \in \mathcal{H}} \left(c\left((y_1, E_{\mathbb{P}_1}(w)), \cdots, (y_\ell, E_{\mathbb{P}_\ell}(w)\right) + \|w\|^2 \right).$$

If the expectations are bounded linear functionals, such regularization functional is of the form (1).

Example 6 (Ivanov regularization)**.** *By allowing the regularization term Ω to take the value $+\infty$, we can also take into account the whole class of Ivanov-type regularization problems of the form*

$$\min_{w \in \mathcal{H}} f(L_1 w, \ldots, L_\ell w), \quad \textit{subject to} \quad \phi(w) \leq 1,$$

by reformulating them as the minimization of a functional of the type (1), where

$$\Omega(w) = \left\{ \begin{array}{ll} 0, & \phi(w) \leq 1 \\ +\infty, & \text{otherwise} \end{array} \right. .$$

1.1 The representer theorem

Let's now go back to the general formulation (1). By the Riesz representation theorem [5, 6], J can be rewritten as

$$J(w) = f(\langle w, w_1\rangle, \ldots, \langle w, w_\ell\rangle) + \Omega(w),$$

where w_i is the representer of the linear functional L_i with respect to the inner product. Consider the following definition.

Definition 1. *A family $\mathcal{F}$ of regularization functionals of the form (1) is said to admit a* linear representer theorem *if, for any $J \in \mathcal{F}$, and any choice of bounded linear functionals L_i, there exists a minimizer w^* that can be written as a linear combination of the representers:*

$$w^* = \sum_{i=1}^{\ell} c_i w_i.$$

If a linear representer theorem holds, the regularization problem under study can be reduced to a ℓ-dimensional optimization problem on the scalar coefficients c_i, independently of the dimension of $\mathcal{H}$. This property is fundamental in practice: without a finite-dimensional parametrization, it wouldn't be possible to employ numerical optimization techniques to compute a solution. Sufficient conditions under which a family of functionals admits a representer theorem have been widely studied in the literature of statistics, inverse problems, and machine learning. The theorem also provides the foundations of learning techniques such as regularized kernel methods and support vector machines, see [7, 8, 9] and references therein.

Representer theorems are of particular interest when $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) [10]. Given a non-empty set $\mathcal{X}$, a RKHS is a space of functions $w : \mathcal{X} \to \mathbb{R}$ such that point-wise evaluation functionals are bounded, namely, for any $x \in \mathcal{X}$, there exists a non-negative real number C_x such that

$$|w(x)| \leq C_x \|w\|, \quad \forall w \in \mathcal{H}.$$

It can be shown that a RKHS can be uniquely associated to a positive-semidefinite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (called *reproducing kernel*), such that the so-called *reproducing property* holds:

$$w(x) = \langle w, K_x\rangle, \qquad \forall\, (x, w) \in \mathcal{X} \times \mathcal{H},$$

where the *kernel sections* K_x are defined as

$$K_x(y) = K(x, y), \qquad \forall y \in \mathcal{X}.$$

The reproducing property states that the representers of point-wise evaluation functionals coincide with the kernel sections. Starting from the reproducing property, it is also easy to show that the representer of any bounded linear functional L is given by a function $K_L \in \mathcal{H}$ such that

$$K_L(x) = LK_x, \qquad \forall x \in \mathcal{X}.$$

Therefore, in a RKHS, the representer of any bounded linear functional can be obtained explicitly in terms of the reproducing kernel.

If the regularization functional (1) admits minimizers, and the regularization term Ω is a nondecreasing function of the norm, i.e.

$$\Omega(w) = h(\|w\|), \quad \text{with } h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}, \text{ nondecreasing}, \tag{2}$$

the linear representer theorem follows easily from the Pythagorean identity. A proof that the condition (2) is sufficient appeared in [11] in the case where $\mathcal{H}$ is a RKHS and L_i are point-wise evaluation functionals. Earlier instances of representer theorems can be found in [12, 13, 14]. More recently, the question of whether condition (2) is also necessary for the existence of linear representer theorems has been investigated [15]. In particular, [15] shows that, if Ω is differentiable (and certain technical existence conditions hold), then (2) is a necessary and sufficient condition for certain classes of regularization functionals to admit a representer theorem. The proof of [15] heavily exploits differentiability of Ω, but the authors conjecture that the hypothesis can be relaxed. In the following, we indeed show that (2) is necessary and sufficient for the family of regularization functionals of the form (1) to admit a linear representer theorem, by merely assuming that Ω is lower semicontinuous and satisfies basic conditions for the existence of minimizers. The proof is based on a characterization of radial nondecreasing functions defined on a Hilbert space.

2 A characterization of radial nondecreasing functions

In this section, we present a characterization of radial nondecreasing functions defined over Hilbert spaces. We will make use of the following definition.

Definition 2. *A subset $\mathcal{S}$ of a Hilbert space $\mathcal{H}$ is called* star-shaped *with respect to a point $z \in \mathcal{H}$ if*

$$(1-\lambda)z + \lambda x \in \mathcal{S}, \quad \forall x \in \mathcal{S}, \quad \forall \lambda \in [0,1].$$

It is easy to verify that a convex set is star-shaped with respect to any point of the set, whereas a star-shaped set does not have to be convex.

The following Theorem provides a geometric characterization of radial nondecreasing functions defined on a Hilbert space that generalizes the analogous result of [15] for differentiable functions.

Theorem 1. *Let $\mathcal{H}$ denote a Hilbert space such that $\dim \mathcal{H} \geq 2$, and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ a lower semicontinuous function. Then, (2) holds if and only if*

$$\Omega(x+y) \geq \max\{\Omega(x), \Omega(y)\}, \qquad \forall x, y \in \mathcal{H} : \langle x, y \rangle = 0. \tag{3}$$

Proof. Assume that (2) holds. Then, for any pair of orthogonal vectors $x, y \in \mathcal{H}$, we have

$$\begin{aligned} \Omega(x+y) &= h\left(\|x+y\|\right) = h\left(\sqrt{\|x\|^2 + \|y\|^2}\right) \geq \max\{h\left(\|x\|\right), h\left(\|y\|\right)\} \\ &= \max\{\Omega(x), \Omega(y)\}. \end{aligned}$$

Conversely, assume that condition (3) holds. Since $\dim \mathcal{H} \geq 2$, by fixing a generic vector $x \in \mathcal{X} \setminus \{0\}$ and a number $\lambda \in [0,1]$, there exists a vector y such that $\|y\| = 1$ and

$$\lambda = 1 - \cos^2\theta,$$

where

$$\cos\theta = \frac{\langle x, y \rangle}{\|x\|\|y\|}.$$

In view of (3), we have

$$\begin{aligned} \Omega(x) &= \Omega(x - \langle x, y \rangle y + \langle x, y \rangle y) \\ &\geq \Omega(x - \langle x, y \rangle y) = \Omega\left(x - \cos^2\theta x + \cos^2\theta x - \langle x, y \rangle y\right) \\ &\geq \Omega\left(\lambda x\right). \end{aligned}$$

Since the last inequality trivially holds also when $x = 0$, we conclude that

$$\Omega(x) \geq \Omega(\lambda x), \qquad \forall x \in \mathcal{H}, \quad \forall \lambda \in [0,1], \tag{4}$$

so that Ω is nondecreasing along all the rays passing through the origin. In particular, the minimum of Ω is attained at $x = 0$.

Now, for any $c \geq \Omega(0)$, consider the sublevel sets

$$\mathcal{S}_c = \{x \in \mathcal{H} : \Omega(x) \leq c\}.$$

From (4), it follows that $\mathcal{S}_c$ is not empty and star-shaped with respect to the origin. In addition, since Ω is lower semicontinuous, $\mathcal{S}_c$ is also closed. We now show that $\mathcal{S}_c$ is either a closed ball centered at the origin, or the whole space. To this end, we show that, for any $x \in \mathcal{S}_c$, the whole ball

$$\mathcal{B} = \{y \in \mathcal{H} : \|y\| \leq \|x\|\},$$

is contained in $\mathcal{S}_c$. First, take any $y \in \text{int}(\mathcal{B}) \setminus \text{span}\{x\}$, where int denotes the interior. Then, y has norm strictly less than $\|x\|$, that is

$$0 < \|y\| < \|x\|,$$

and is not aligned with x, i.e.

$$y \neq \lambda x, \quad \forall \lambda \in \mathbb{R}.$$

Let $\theta \in \mathbb{R}$ denote the angle between x and y. Now, construct a sequence of points x_k as follows:

$$\begin{cases} x_0 = y, \\ x_{k+1} = x_k + a_k u_k, \end{cases}$$

where

$$a_k = \|x_k\| \tan\left(\frac{\theta}{n}\right), \qquad n \in \mathbb{N}$$

and u_k is the unique unitary vector that is orthogonal to x_k, belongs to the two-dimensional subspace span$\{x, y\}$, and is such that $\langle u_k, x \rangle > 0$, that is

$$u_k \in \text{span}\{x, y\}, \qquad \|u_k\| = 1, \qquad \langle u_k, x_k \rangle = 0, \qquad \langle u_k, x \rangle > 0.$$

See Figure 1 for a geometrical illustration of the sequence x_k.

By orthogonality, we have

$$\|x_{k+1}\|^2 = \|x_k\|^2 + a_k^2 = \|x_k\|^2 \left(1 + \tan^2\left(\frac{\theta}{n}\right)\right) = \|y\|^2 \left(1 + \tan^2\left(\frac{\theta}{n}\right)\right)^{k+1}. \tag{5}$$

In addition, the angle between x_{k+1} and x_k is given by

$$\theta_k = \arctan\left(\frac{a_k}{\|x_k\|}\right) = \frac{\theta}{n},$$

so that the total angle between y and x_n is given by

$$\sum_{k=0}^{n-1} \theta_k = \theta.$$

Since all the points x_k belong to the subspace spanned by x and y, and the angle between x and x_n is zero, we have that x_n is positively aligned with x, that is

$$x_n = \lambda x, \qquad \lambda \geq 0.$$

Now, we show that n can be chosen in such a way that $\lambda \leq 1$. Indeed, from (5) we have

$$\lambda^2 = \left(\frac{\|x_n\|}{\|x\|}\right)^2 = \left(\frac{\|y\|}{\|x\|}\right)^2 \left(1 + \tan^2\left(\frac{\theta}{n}\right)\right)^n,$$

and it can be verified that

$$\lim_{n \to +\infty} \left(1 + \tan^2\left(\frac{\theta}{n}\right)\right)^n = 1,$$

therefore $\lambda \leq 1$ for a sufficiently large n. Now, write the difference vector in the form

$$\lambda x - y = \sum_{k=0}^{n-1} (x_{k+1} - x_k),$$

and observe that

$$\langle x_{k+1} - x_k, x_k \rangle = 0.$$

By using (4) and proceeding by induction, we have

$$c \geq \Omega(\lambda x) = \Omega\left(x_n - x_{n-1} + x_{n-1}\right) \geq \Omega(x_{n-1}) \geq \cdots \geq \Omega(x_0) = \Omega(y),$$

so that $y \in \mathcal{S}_c$. Since $\mathcal{S}_c$ is closed and the closure of int$(\mathcal{B}) \setminus$ span$\{x\}$ is the whole ball $\mathcal{B}$, every point $y \in \mathcal{B}$ is also included in $\mathcal{S}_c$. This proves that $\mathcal{S}_c$ is either a closed ball centered at the origin, or the whole space $\mathcal{H}$.

Finally, for any pair of points such that $\|x\| = \|y\|$, we have $x \in \mathcal{S}_{\Omega(y)}$, and $y \in \mathcal{S}_{\Omega(x)}$, so that

$$\Omega(x) = \Omega(y).$$

$\square$

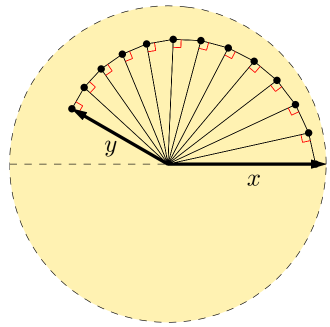

Figure 1: The sequence x_k constructed in the proof of Theorem 1 is associated with a geometrical construction known as *spiral of Theodorus*. Starting from any y in the interior of the ball (excluding points aligned with x), a point of the type λx (with $0 \leq \lambda \leq 1$) can be reached by using a finite number of right triangles.

3 Representer theorem: a necessary and sufficient condition

In this section, we prove that condition (2) is necessary and sufficient for suitable families of regularization functionals of the type (1) to admit a linear representer theorem.

Theorem 2. *Let $\mathcal{H}$ denote a Hilbert space of dimension at least 2. Let $\mathcal{F}$ denote a family of functionals $J : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ of the form (1) that admit minimizers, and assume that $\mathcal{F}$ contains a set of functionals of the form*

$$J_p^\gamma(w) = \gamma f\left(\langle w, p\rangle\right) + \Omega\left(w\right), \qquad \forall p \in \mathcal{H}, \quad \forall \gamma \in \mathbb{R}_+, \tag{6}$$

where $f(z)$ is uniquely minimized at $z = 1$. Then, for any lower semicontinuous Ω, the family $\mathcal{F}$ admits a linear representer theorem if and only if (2) holds.

Proof. The first part of the theorem (sufficiency) follows from an orthogonality argument. Take any functional $J \in \mathcal{F}$. Let $\mathcal{R} = \text{span}\{w_1, \ldots, w_\ell\}$ and let $\mathcal{R}^\perp$ denote its orthogonal complement. Any minimizer w^* of J can be uniquely decomposed as

$$w^* = u + v, \qquad u \in \mathcal{R}, \quad v \in \mathcal{R}^\perp.$$

If (2) holds, then we have

$$J(w^*) - J(u) = h(\|w^*\|) - h(\|u\|) \geq 0,$$

so that $u \in \mathcal{R}$ is also a minimizer.

Now, let's prove the second part of the theorem (necessity). First of all, observe that the functional

$$J_0^\gamma(w) = \gamma f(0) + \Omega(w),$$

obtained by setting $p = 0$ in (6), belongs to $\mathcal{F}$. By hypothesis, J_0^γ admits minimizers. In addition, by the representer theorem, the only admissible minimizer of J_0 is the origin, that is

$$\Omega(y) \geq \Omega(0), \qquad \forall y \in \mathcal{H}. \tag{7}$$

Now take any $x \in \mathcal{H} \setminus \{0\}$ and let

$$p = \frac{x}{\|x\|^2}.$$

By the representer theorem, the functional J_p^γ of the form (6) admits a minimizer of the type

$$w = \lambda(\gamma)x.$$

Now, take any $y \in \mathcal{H}$ such that $\langle x, y\rangle = 0$. By using the fact that $f(z)$ is minimized at $z = 1$, and the linear representer theorem, we have

$$\gamma f(1) + \Omega\left(\lambda(\gamma)x\right) \leq \gamma f(\lambda(\gamma)) + \Omega\left(\lambda(\gamma)x\right) = J_p^\gamma(\lambda(\gamma)x) \leq J_p^\gamma(x+y) = \gamma f(1) + \Omega\left(x+y\right).$$

By combining this last inequality with (7), we conclude that

$$\Omega\left(x+y\right) \geq \Omega\left(\lambda(\gamma)x\right), \qquad \forall x, y \in \mathcal{H} : \langle x, y\rangle = 0, \qquad \forall \gamma \in \mathbb{R}_+. \tag{8}$$

Now, there are two cases:

- $\Omega(x+y) = +\infty$
- $\Omega(x+y) = C < +\infty$.

In the first case, we trivially have

$$\Omega(x+y) \geq \Omega(x).$$

In the second case, using (7) and (8), we obtain

$$0 \leq \gamma\left(f(\lambda(\gamma)) - f(1)\right) \leq \Omega(x+y) - \Omega(\lambda(\gamma)x) \leq C - \Omega(0) < +\infty, \qquad \forall \gamma \in \mathbb{R}_+. \tag{9}$$

Let γ_k denote a sequence such that $\lim_{k\to+\infty} \gamma_k = +\infty$, and consider the sequence

$$a_k = \gamma_k\left(f(\lambda(\gamma_k)) - f(1)\right).$$

From (9), it follows that a_k is bounded. Since $z = 1$ is the only minimizer of $f(z)$, the sequence a_k can remain bounded only if

$$\lim_{k\to+\infty} \lambda(\gamma_k) = 1.$$

By taking the limit inferior in (8) for $\gamma \to +\infty$, and using the fact that Ω is lower semicontinuous, we obtain condition (3). It follows that Ω satisfies the hypotheses of Theorem 1, therefore (2) holds. $\square$

The second part of Theorem 2 states that any lower semicontinuous regularization term Ω has to be of the form (2) in order for the family $\mathcal{F}$ to admit a linear representer theorem. Observe that Ω is not required to be differentiable or even continuous. Moreover, it needs not to have bounded lower level sets. For the necessary condition to hold, the family $\mathcal{F}$ has to be broad enough to contain at least a set of regularization functionals of the form (6). The following examples show how to apply the necessary condition of Theorem 2 to classes of regularization problems with standard loss functions.

- Let $L : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ denote any loss function of the type

  $$L(y,z) = \widetilde{L}(y-z),$$

  such that $\widetilde{L}(t)$ is uniquely minimized at $t = 0$. Then, for any lower semicontinuous regularation term Ω, the family of regularization functionals of the form

  $$J(w) = \gamma \sum_{i=1}^{\ell} L\left(y_i, \langle w, w_i\rangle\right) + \Omega(w),$$

  admits a linear representer theorem if and only if (2) holds. To see that the hypotheses of Theorem 2 are satisfied, it is sufficient to consider the subset of functionals with $\ell = 1$, $y_1 = 1$, and $w_1 = p \in \mathcal{H}$. These functionals can be written in the form (6) with

  $$f(z) = L(1,z).$$

- The class of regularization problems with the hinge (SVM) loss of the form

  $$J(w) = \gamma \sum_{i=1}^{\ell} \max\{0, 1 - y_i\langle w, w_i\rangle\} + \Omega(w),$$

  with Ω lower semicontinuous, admits a linear representer theorem if and only if Ω satisfy (2). For instance, by choosing $\ell = 2$, and

  $$(y_1, w_1) = (1, p), \qquad (y_2, w_2) = (-1, p/2),$$

  we obtain regularization functionals of the form (6) with

  $$f(z) = \max\{0, 1-z\} + \max\{0, 1+z/2\},$$

  and it is easy to verify that f is uniquely minimized at $z = 1$.

4 Conclusions

Sufficiently broad families of regularization functionals defined over a Hilbert space with lower semicontinuous regularization term admit a linear representer theorem if and only if the regularization term is a radial nondecreasing function. More precisely, the main result of this paper (Theorem 2) implies that, for any sufficiently large family of regularization functionals, nondecreasing functions of the norm are the only lower semicontinuous (extended-real valued) regularization terms that guarantee existence of a representer theorem for any choice of the data functionals L_i.

As a concluding remark, it is important to observe that other types of regularization terms are possible if the representer theorem is only required to hold for a restricted subset of the data functionals. Exploring necessary conditions for the existence of representer theorems under different types of restrictions on the data functionals is an interesting future research direction.

5 Acknowledgments

The authors would like to thank Andreas Argyriou for useful discussions.

References

[1] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, Washington, D. C., 1977.

[2] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, USA, 1990.

[3] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39:1–49, 2001.

[4] B. Schölkopf, A. J. Smola, and K-R Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[5] F. Riesz. Sur une espèce de géométrie analytique des systèmes de fonctions sommables. *Comptes rendus de l'Académie des sciences Paris*, 144:1409–1411, 1907.

[6] M. Fréchet. Sur les ensembles de fonctions et les opérations linéaires. *Comptes rendus de l'Académie des sciences Paris*, 144:1414–1416, 1907.

[7] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

[8] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (Adaptive Computation and Machine Learning). MIT Press, 2001.

[9] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[10] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[11] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *In Proceedings of the Annual Conference on Computational Learning Theory*, pages 416–426, 2001.

[12] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[13] D. Cox and F. O' Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics*, 18:1676–1695, 1990.

[14] T. Poggio and F. Girosi. Networks for approximation and learning. In *Proceedings of the IEEE*, volume 78, pages 1481–1497, 1990.

[15] A. Argyriou, C. A. Micchelli, and M. Pontil. When is there a representer theorem? Vector versus matrix regularizers. *Journal of Machine Learning Research*, 10:2507–2529, 2009.