A Unifying View of Representer Theorems

Andreas Argyriou ARGYRIOUA@ECP.FR
École Centrale Paris, Center for Visual Computing

Francesco Dinuzzo FRANCESD@IE.IBM.COM
IBM Research, Dublin *and* Max Planck Institute for Intelligent Systems, Tübingen

Abstract

It is known that the solution of regularization and interpolation problems with Hilbertian penalties can be expressed as a linear combination of the data. This very useful property, called the *representer theorem*, has been widely studied and applied to machine learning problems. Analogous optimality conditions have appeared in other contexts, notably in matrix regularization. In this paper we propose a *unified view*, which generalizes the concept of representer theorems and extends necessary and sufficient conditions for such theorems to hold. Our main result shows a close connection between representer theorems and certain classes of regularization penalties, which we call *orthomonotone functions*. This result not only subsumes previous representer theorems as special cases but also yields a new class of optimality conditions, which goes beyond the classical linear combination of the data. Moreover, orthomonotonicity provides a useful *criterion* for testing whether a representer theorem holds for a specific regularization problem.

1. Introduction

One of the dominant approaches in machine learning and statistics is to formulate a learning problem as an optimization problem to be solved. In particular, *regularization* has been widely used for learning or estimating functions or models from input and output data, particularly in supervised and semisupervised learning.

Regularization in a Hilbert space $\mathcal{H}$ frames the problem of learning from data as a minimization of the type

$$\min\{f(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) + \gamma\, \Omega(w) : w \in \mathcal{H}\}\ . \quad (1)$$

The objective function is the sum of an *error term* f which depends on prescribed data[1] $w_1, \ldots, w_m \in \mathcal{H}$, and a *regularization penalty* Ω, which favors certain desirable properties of the solution. An optimal solution of problem (1) yields the desired function or vector, depending on the context of the original learning problem.

It is known that, for a certain class of regularization and interpolation problems, one of the optimal solutions of (1) can be expressed as a linear combination of the data. More specifically, this is the case when the penalty Ω is a Hilbertian norm (or a nondecreasing function of that). This property, known as the *representer theorem*, has proven very useful because it renders many high or infinite dimensional regularization problems amenable to practical computation. This "classical" representer theorem was formulated in various guises in (Girosi, 1998; Kimeldorf & Wahba, 1970; Schölkopf et al., 2001) and has been the topic of extensive further study (Argyriou et al., 2009; De Vito et al., 2004; Dinuzzo & Schölkopf, 2012; Dinuzzo et al., 2007; Gnecco & Sanguineti, 2010; Mukherjee & Wu, 2006; Steinwart, 2003; Yu et al., 2013). In machine learning, the representer theorem is the main factor that enables application of the so-called "kernel trick" and underpins all of the widely used *kernel methods* (Schölkopf & Smola, 2002), such as support vector machines, regularization networks, etc.

Besides the classical result, more recently new types of representer theorems have been proven and studied. For example, it has been realized that analogous optimality conditions apply to the learning of vector-valued functions (Micchelli & Pontil, 2005), ℓ_2-regularized multitask learning (Evgeniou et al., 2005) and structured prediction (Lafferty et al., 2004). Further developments occurred with the advent of *matrix regularization* problems used for multitask learning or collaborative filtering. Thus it has been shown that a type of representer theorem holds when the penalty Ω is a spectral function of matrices

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

[1] We use the term "data" in a more general sense than input vectors in Euclidean space – see Section 3 for examples.

(Amit et al., 2007; Argyriou et al., 2009; 2010) or operators (Abernethy et al., 2009). Very recently these results have been extended to matricizations of tensors as well (Signoretto et al., 2013). Other related results have appeared in the contexts of domain adaptation (Kulis et al., 2011), dimensionality reduction (Jain et al., 2010) and metric learning (Jain et al., 2012).

Some variants of the classical theorem were shown in the contexts of semisupervised learning (Belkin et al., 2006), semiparametric representer theorems and kernel PCA (Schölkopf et al., 2001). Moreover, there have appeared alternative approaches which lie outside the scope of this paper, such as a Bayesian variant of the classical theorem (Pillai et al., 2007), the theory of reproducing kernel Banach spaces (Zhang & Zhang, 2012) and an algorithmic theorem for matrices (Warmuth et al., 2012). Clearly therefore, representer theorems are important and ubiquitous tools in regularization and underly a wide range of frequently used machine learning methodologies.

In this paper, we address the topic of representer theorems from a new and more abstract viewpoint. One of our contributions is to provide a *unifying framework* which subsumes the results that have already appeared in the literature. These include the classical, vector valued, structured prediction, multitask, tensor, semisupervised, semiparametric, dimensionality reduction, domain adaptation, metric learning results etc. In particular, we show that these theorems are only examples from a larger family. Each theorem in this family corresponds to a class of regularization penalties which are characterized by an *orthomonotonicity* property that we introduce. Another implication of our results is that we can now put the study of representer theorems on a *formal basis* and provide calculus rules and recipes for deriving new results. Most commonly used kernel methods (support vector machines, kernel ridge regression etc.), as well as methods for multitask learning, collaborative filtering and metric learning, fall within our framework. As an illustration of the theory, we demonstrate that regularization problems with a generalized family of matrix penalties, as well as similar problems on the positive semidefinite cone, admit appropriate representer theorems. In many practical situations this implies that the number of degrees of freedom and hence the complexity of solving the learning problem decreases significantly. [2]

2. Mathematical Preliminaries

In this section, we introduce the notation and the mathematical concepts necessary for our framework and the main results of Section 3.

[2] The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7 2007-2013) under grant agreement No. 246556.

2.1. Notational conventions

Let $\mathcal{H}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. We use $\mathscr{L}(\mathcal{H})$ to denote the set of linear operators from $\mathcal{H}$ to itself. We denote the identity operator by $Id \in \mathscr{L}(\mathcal{H})$ and the set of linear subspaces of $\mathcal{H}$ by $\mathcal{V}(\mathcal{H})$.

Also let $\mathbb{N}_m$ denote the set of integers $\{1, \dots, m\}$, $\mathbf{M}_{d,n}$ the set of real $d \times n$ matrices and $\mathbf{M}_n$ the set of real $n \times n$ matrices. Moreover, let $\mathbf{S}_+^n$ denote the set of $n \times n$ positive semidefinite matrices and $\mathbf{S}_{++}^n$ the set of positive definite ones. We denote the t-th column of a matrix $W \in \mathbf{M}_{d,n}$ by w_t. We use the following notation for operations on sets $A, B \subseteq \mathcal{H}$: $A + B := \{a + b : a \in A, b \in B\}$, $A - B := \{a - b : a \in A, b \in B\}$ $\lambda A := \{\lambda a : a \in A\}$, for every $\lambda \in \mathbb{R}$.

In the following, we will be working with *subspace-valued* maps $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. This choice is natural, since representer theorems are statements that solutions of certain optimization problems belong to certain subspaces. For more details, see Section 3 and our general definition of representer theorems. Given two subspace-valued maps S_1 and S_2, their sum $S_1 + S_2$ maps every $x \in \mathcal{H}$ to the sum of subspaces $S_1(x) + S_2(x)$.

2.2. Quasilinear Subspace-Valued Maps

To extend the concept of representers, we first introduce a variant of linearity appropriate for subspace-valued maps.[3]

Definition 2.1. *We call the map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$* quasilinear *if*

$$S(\alpha x + \beta y) \subseteq \alpha S(x) + \beta S(y)$$

for every $x, y \in \mathcal{H}$, $\alpha, \beta \in \mathbb{R}$.

Proposition 2.1. *Let S denote a quasilinear subspace-valued map. Then*

$$S(\alpha x) = \alpha S(x) \tag{2}$$

and

$$\alpha S(x) \subseteq S(\alpha x + \beta y) + \beta S(y) \tag{3}$$

for every $x, y \in \mathcal{H}$, $\alpha, \beta \in \mathbb{R}$.

Definition 2.2. *We call the map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$* idempotent *if*

$$S(S(x)) = S(x), \qquad \forall x \in \mathcal{H}.$$

Lemma 2.1. *Let S be a quasilinear and idempotent subspace-valued map. Then, for every $m \in \mathbb{N}$ and every set $\{x_i : i \in \mathbb{N}_m\} \subseteq \mathcal{H}$, it holds that*

$$S\left(\sum_{i=1}^m S(x_i)\right) \subseteq \sum_{i=1}^m S(x_i).$$

[3] Proofs of the lemmas appearing throughout the paper can be found in the supplement.

Thus addition of subspaces can be used to generate subspaces invariant under S.

In addition to the quasilinearity and idempotence assumptions, we require that sums of images under S are closed. This ensures that *orthogonal projection* on such subspaces is feasible, which is a crucial step in the proof of representer theorems. For simplicity, to satisfy this property we assume that all images under S are finite dimensional. Another assumption necessary for the proof of our main result is that any point belongs to its image under S. Summarizing, we collect all of the above assumptions in the following definition.

Definition 2.3. *Let $r \in \mathbb{N}$. We call the subspace-valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ r-*regular quasilinear *if it is quasilinear, idempotent and if, for all $x \in \mathcal{H}$, $S(x)$ has dimensionality at most r and contains x.*

The simplest example of regular quasilinear subspace-valued map S is the map that associates a given vector to its own linear span, which thus has dimensionality one.

Example 2.1. *Suppose that S maps*

$$x \mapsto \operatorname{span}\{x\} \ .$$

Then S is 1-regular quasilinear.

More generally, we can map each point to a subspace by applying a set of linear transformations and taking the linear subspace spanned by the resulting vectors.

Example 2.2. *Let $r \in \mathbb{N}$ and suppose that S maps*

$$x \mapsto \operatorname{span}\{T_i x : i \in \mathbb{N}_r\} \ ,$$

where $T_i \in \mathscr{L}(\mathcal{H})$ for all $i \in \mathbb{N}_r$. Then S is quasilinear. This map is r-regular quasilinear if

- $Id \in \operatorname{span}\{T_i : i \in \mathbb{N}_r\}$,
- $T_j T_\ell \in \operatorname{span}\{T_i : i \in \mathbb{N}_r\} \quad \forall j, \ell \in \mathbb{N}_r$.

Remark 2.1. *It may not hold that S maps any linear subspace of $\mathcal{H}$ to a linear subspace (as illustrated in Example 2.2 when, say, $r = 2$, and $T_1 x, T_2 x, T_1 y, T_2 y$ are linearly independent for some $x, y \in \mathcal{H}$). Even when this condition holds, the image of a linear subspace may be a different subspace (consider $S(x) = \mathcal{H}$, $\forall x \neq 0$, and $\operatorname{span}\{x\}$).*

A special case of Example 2.2 is the following, defined for a space of matrices. As we shall see, this example is relevant to representer theorems for multitask learning.

Example 2.3. *Let $\mathcal{H} = \mathbf{M}_{d,n}$ equipped with the standard inner product, and suppose that S maps*

$$X \mapsto \{XC : C \in \mathbf{M}_n\} \ .$$

Then S is $\min\{n^2, dn\}$-regular quasilinear.

3. Characterization of the General Representer Theorem

Our focus of interest is the variational problem of minimizing, over a Hilbert space, a regularization functional of the form

$$J(w) = f(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) + \gamma\, \Omega(w) \ . \qquad (4)$$

The functional J is the sum of an *error term* $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, which depends on prescribed data $w_1, \ldots, w_m \in \mathcal{H}$, and a *regularization term* $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, which enforces certain desirable properties on the solution, scaled by a *regularization parameter* $\gamma > 0$. We allow both f and Ω to take the value $+\infty$, so that interpolation problems and regularization problems of the Ivanov type can also be taken into account.

Since the same functional J might be decomposed into a form like (4) in multiple ways, we fix $m \in \mathbb{N}$ and use the tuple $(f, \Omega, \gamma, w_1, \ldots, w_m)$ to describe such a regularization functional.

Example 3.1 (Interpolation in a Hilbert space)**.** *Let $w_1, \ldots, w_m \in \mathcal{H}$ and $y_1, \ldots, y_m \in \mathbb{R}$ be prescribed data and*

$$f(z) = \begin{cases} 0 & \text{if } z = y \\ +\infty & \text{otherwise} \end{cases} , \qquad \forall z \in \mathbb{R}^m .$$

Then the interpolation problem

$$\min\{\Omega(w) \ : \ w \in \mathcal{H},\ \langle w, w_i\rangle = y_i \ \ \forall i \in \mathbb{N}_m\} \ .$$

is equivalent to the problem of minimizing (4) over $\mathcal{H}$.

Example 3.2 (Ivanov regularization)**.** *Ivanov regularization amounts to solving a problem of the form*

$$\min\{f(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) \ : \ w \in \mathcal{H},\ \omega(w) \leq 1\} \ ,$$

where $\omega : \mathcal{H} \to \mathbb{R}$ is a prescribed constraining function. Defining

$$\Omega(w) = \begin{cases} 0 & \text{if } \omega(w) \leq 1 \\ +\infty & \text{otherwise} \end{cases} ,$$

this problem can be rewritten as the minimization of a functional of the form (4).

Example 3.3 (Regularization in an RKHS)**.** *Reproducing Kernel Hilbert Spaces (RKHS) are Hilbert spaces $\mathcal{H}$ of functions $w : \mathcal{X} \to \mathbb{R}$ defined over a nonempty set $\mathcal{X}$ such that all point-wise evaluation functionals are bounded, that is, for all $x \in \mathcal{X}$ there exists a constant $C_x < +\infty$ such that*

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

It can be shown that RKHS exhibit the so-called reproducing property $w(x) = \langle w, K_x \rangle, \forall\, (x, w) \in \mathcal{X} \times \mathcal{H}$, *where the representers* $K_x \in \mathcal{H}$ *are expressible as sections of a symmetric and positive semidefinite kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *such that* $K_x(y) = K(x, y)$, $\forall y \in \mathcal{X}$. *The reproducing property of an RKHS allows for rewriting any regularization functional of the form*

$$J(w) = f(w(x_1), \ldots, w(x_m)) + \gamma\, \Omega(w)$$

in the standard form (4)*, where the representers* w_i *coincide with the kernel sections* K_{x_i}.

Example 3.4 (Regularization with averaged data)**.** *In some estimation problems, it may be appropriate to assume that the measured output data are obtained by averaging a function* w *(to be estimated) with respect to suitable probability measures. Let* $\mathbb{P}_1, \ldots, \mathbb{P}_m$ *denote probability measures on the measurable space* $(\mathcal{X}, \mathcal{A})$*, where* $\mathcal{A}$ *is a* σ*-algebra of subsets of* $\mathcal{X}$*, and let* $\mathcal{H}$ *denote a Hilbert space of functions* $w : \mathcal{X} \to \mathbb{R}$*. If, for every* $i \in \mathbb{N}_m$*, the expectation*

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

is a bounded linear functional over $\mathcal{H}$*, then one may consider synthesizing a function* w *by minimizing a functional of the form*

$$J(w) = f(E_{\mathbb{P}_1}(w), \ldots, E_{\mathbb{P}_m}(w)) + \gamma\, \Omega(w)$$

which can be rewritten in the form (4) *by introducing suitable representers* w_i*. In particular, if* $\mathcal{H}$ *is an RKHS with a bounded reproducing kernel, the representers of the expectation functionals* $E_{\mathbb{P}_i}$ *are called* kernel mean embeddings *– see, for example, (Muandet et al., 2012; Sriperumbudur et al., 2010) and references therein – and can be explicitly expressed as*

$$w_i(t) = \int_{\mathcal{X}} K(x, t) d\mathbb{P}_i(x), \quad \forall i \in \mathbb{N}_m, t \in \mathcal{X}.$$

Clearly, we are interested only in cases in which the optimization problem

$$\min\{J(w) : w \in \mathcal{H}\}$$

is well defined, that is, a minimizer of J exists. This always holds by construction in machine learning and statistics applications. More generally, existence of a minimizer can be ensured under lower semicontinuity and coercivity conditions on J. We will avoid specifying such precise conditions since they are not relevant to our purposes, instead assuming existence of minimizers for each problem of interest.

The main question we address in this paper is to characterize the functions Ω for which minimizers of the regularization functional (4) admit certain convenient representations. As already mentioned in the introduction, representer theorems have been proven for regularization with the Hilbertian norm, Schatten ℓ_p regularization (Abernethy et al., 2009; Argyriou et al., 2009; 2010) and in some other cases. These theorems state that a minimizer must lie in a subspace which depends on the data points $w_1, \ldots, w_m$. This dependence on the data varies according to the regularization penalty Ω. For example, in the classical representer theorem ($\Omega = \|\cdot\|_{\mathcal{H}}$), the subspace is simply the span of the data points. In the multitask theorem (Ω is a spectral function on matrices), the subspace is generated by the columns of the data matrices.

Our goal is to unify this prior work under one framework and at the same time to extend the applicability of representer theorems to other regularization problems. The key to this is to associate representations of minimizers with the data points in an abstract way, specifically to associate a subspace to each data point. Hence we assume a subspace-valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ and require that the representation for a minimizer of (4) be spanned by the elements of $S(w_i)$, $i \in \mathbb{N}_m$.

Definition 3.1. *Let* $m \in \mathbb{N}$, $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ *be a subspace-valued map and* $J = (f, \Omega, \gamma, w_1, \ldots, w_m)$ *a regularization functional of the form* (4)*. Then* J *is said to admit a* representer theorem *with respect to* S *if there exists a minimizer* $\hat{w}$ *of* J *such that*

$$\hat{w} \in \sum_{i=1}^{m} S(w_i).$$

Definition 3.2. *Let* $m \in \mathbb{N}$, $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ *be a subspace-valued map and* $\mathcal{F}$ *a family of regularization functionals of the form* (4)*. Then* $\mathcal{F}$ *is said to admit a representer theorem with respect to* S *if every* $J \in \mathcal{F}$ *admits a representer theorem with respect to* S.

Our main tool for characterizing regularization functionals that admit representer theorems is the property defined below, which we call *orthomonotonicity*. The connection between orthomonotonicity and representer theorems has appeared in (Argyriou et al., 2009) in the context of regularization with the Hilbertian norm or with orthogonally invariant matrix penalties. In Theorem 3.1, we extend this connection to a broader class of regularization penalties Ω which arise by varying the choice of the map S.

Definition 3.3. *We call the function* $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ orthomonotone *with respect to the map* $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$*, if*

$$\Omega(x + y) \geq \Omega(x), \qquad \forall x \in \mathcal{H},\ y \in S(x)^{\perp}. \tag{5}$$

Note that in this definition the left hand side of (5), or even both sides, may equal $+\infty$.

Theorem 3.1. *Let* $r, m \in \mathbb{N}$, $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *and suppose that* $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ *is an* r*-regular quasilinear map. Then the following hold:*

1. *If Ω is orthomonotone w.r.t. S then, for any $w_1, \dots, w_m \in \mathcal{H}$ and any $\gamma > 0$ such that the regularization functional $J = (f, \Omega, \gamma, w_1, \dots, w_m)$ of the form (4) admits a minimizer, J admits a representer theorem w.r.t. S.*

2. *Let $\mathcal{F}$ denote the following family of regularization functionals of form (4)*

$$\mathcal{F} = \{(f, \Omega, \gamma, w_1, \dots, w_m) : w_1, \dots, w_m \in \mathcal{H}, \gamma > 0\} \tag{6}$$

and assume that

   - *f is lower semicontinuous, admits a unique minimizer $\hat{z} \neq 0$, and there exists $\varepsilon > 0$ such that the sublevel set $\{z \in \mathcal{H} : f(z) \leq f(\hat{z}) + \varepsilon\}$ is bounded.*
   - *Ω is lower semicontinuous and is minimized at 0*
   - *$r \leq m$.*

   If $\mathcal{F}$ admits a representer theorem w.r.t. S then Ω is orthomonotone w.r.t. S.

Proof. The first part of the theorem (sufficiency) can be proven by adapting a classical orthogonality argument. Take any $w_1, \dots, w_m \in \mathcal{H}$, $\gamma > 0$. Let $\mathcal{L} = \sum_{i=1}^{m} S(w_i)$ and let $\mathcal{L}^\perp$ denote its orthogonal complement. Due to the regular quasilinearity of S, $\mathcal{L}$ is a finite dimensional subspace that contains $\mathcal{R} = \operatorname{span}\{w_1, \dots, w_m\}$. Therefore any minimizer $\hat{w}$ of the regularization functional $J = (f, \Omega, \gamma, w_1, \dots, w_m)$ can be decomposed as

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

Applying Lemma 2.1 we obtain that $S(u) \subseteq \mathcal{L}$ and hence that $v \in S(u)^\perp$. If Ω is orthomonotone then

$$\begin{aligned} J(\hat{w}) &= f(\langle u+v, w_1\rangle, \dots, \langle u+v, w_m\rangle) + \gamma\, \Omega(u+v) \\ &= f(\langle u, w_1\rangle, \dots, \langle u, w_m\rangle) + \gamma\, \Omega(u+v) \\ &\geq f(\langle u, w_1\rangle, \dots, \langle u, w_m\rangle) + \gamma\, \Omega(u) \\ &= J(u), \end{aligned}$$

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

Now, let us prove the second part of the theorem (necessity). Let us fix arbitrary $x \in \mathcal{H}$ and $y \in S(x)^\perp$. The goal of the proof is to establish orthomonotonicity, namely the inequality

$$\Omega(x+y) \geq \Omega(x). \tag{7}$$

The proof is organized in three cases.

1. First, we observe that for $x = 0$ the inequality follows directly from the hypothesis on Ω.

2. Secondly, observe that if $\Omega(x+y) = +\infty$, inequality (7) is trivially satisfied.

3. It remains to prove (7) in the case when

$$x \neq 0 \quad \text{and} \quad \Omega(x+y) = C < +\infty. \tag{8}$$

Since S is r-regular quasilinear and $r \leq m$, $S(x)$ has dimensionality $\mu \leq m$. Let us choose a set of vectors $\{b_i(x) : i \in \mathbb{N}_m\}$ spanning $S(x)$ in a way such that

$$\langle x, b_i(x)\rangle = \hat{z}_i, \qquad \forall i \in \mathbb{N}_m. \tag{9}$$

Such a set can always be constructed, since $x \in S(x)$, $x \neq 0$ and $\hat{z} \neq 0$, as follows. Pick $u_1, \dots, u_{\mu-1} \in S(x)$, such that $\{u_1, \dots, u_{\mu-1}, x\}$ forms an orthogonal basis of $S(x)$. Without loss of generality, we may assume that $\hat{z}_m \neq 0$. Then we may define $b_i(x) = u_i + \frac{\hat{z}_i}{\|x\|^2} x$ for $1 \leq i \leq \mu - 1$ and $b_i(x) = \frac{\hat{z}_i}{\|x\|^2} x$ for $\mu \leq i \leq m$. Clearly these vectors span $S(x)$ and satisfy (9).

For every $\gamma > 0$, the functional $J_x^\gamma = (f, \Omega, \gamma, b_1(x), \dots, b_m(x))$ belongs to $\mathcal{F}$. Therefore, by the representer theorem w.r.t. S, J_x^γ admits a minimizer $w_x^\gamma \in \sum_{i=1}^{m} S(b_i(x)) \subseteq S(x)$ (where the last inclusion follows from the idempotence of S). Let

$$z_x^\gamma = \left(\langle w_x^\gamma, b_1(x)\rangle \ \dots \ \langle w_x^\gamma, b_m(x)\rangle\right).$$

Using the facts that f is minimized at $\hat{z}$, J_x^γ is minimized at w_x^γ and $y \in S(x)^\perp$, we obtain

$$\begin{aligned} f(\hat{z}) + \gamma\, \Omega(w_x^\gamma) &\leq f(z_x^\gamma) + \gamma\, \Omega(w_x^\gamma) = J_x^\gamma(w_x^\gamma) \\ &\leq J_x^\gamma(x+y) = f(\hat{z}) + \gamma\, \Omega(x+y) \end{aligned} \tag{10}$$

for all $\gamma > 0$. Note that $f(\hat{z})$ is finite since $\hat{z}$ is the unique minimizer of f. Hence we conclude that

$$\Omega(x+y) \geq \Omega(w_x^\gamma), \qquad \forall \gamma > 0. \tag{11}$$

In order to conclude the proof, we show that w_x^γ converges to x as $\gamma \to 0^+$. Using (10), (8) and the hypothesis that Ω is minimized at 0, we obtain

$$\begin{aligned} 0 \leq f(z_x^\gamma) - f(\hat{z}) &\leq \gamma\, (\Omega(x+y) - \Omega(w_x^\gamma)) \\ &= \gamma\, (C - \Omega(w_x^\gamma)) \\ &\leq \gamma\, (C - \Omega(0)) \\ &< +\infty, \qquad \forall \gamma > 0. \end{aligned} \tag{12}$$

Now, let γ_k denote a sequence of positive real numbers such that $\lim_{k\to\infty} \gamma_k = 0$. From (12) it follows that

$$\lim_{k\to\infty} f(z_x^{\gamma_k}) = f(\hat{z}).$$

It follows that there exists an index M such that, for all $k \geq M$, $z_x^{\gamma_k}$ belongs to the bounded set $\{z \in \mathcal{H} : f(z) \leq f(\hat{z}) + \varepsilon\}$. Therefore, the sequence $z_x^{\gamma_k}$ is bounded and it has a convergent subsequence. Now, take an arbitrary convergent subsequence of $z_x^{\gamma_k}$ and let $\bar{z}$ denote its limit. Since f is lower-semicontinuous and $\hat{z}$ is its only minimizer, it must be $\bar{z} = \hat{z}$. Hence, the whole sequence $z_x^{\gamma_k}$ converges to $\hat{z}$, namely

$$\lim_{k\to\infty} \langle w_x^{\gamma_k}, b_i(x)\rangle = \hat{z}_i, \qquad \forall i \in \mathbb{N}_m.$$

In view of (9) we have, for every $i \in \mathbb{N}_m$,

$$\lim_{k\to\infty} \langle w_x^{\gamma_k} - x, b_i(x)\rangle = \lim_{k\to\infty} \langle w_x^{\gamma_k}, b_i(x)\rangle - \hat{z}_i = 0$$

and therefore

$$\lim_{k\to\infty} \langle w_x^{\gamma_k} - x, u\rangle = 0, \qquad \forall u \in S(x). \quad (13)$$

Since $x, w_x^{\gamma_k} \in S(x)$, the sequence $w_x^{\gamma_k} - x$ is confined to the subspace $S(x)$. Since $S(x)$ is finite-dimensional, (13) implies that $w_x^{\gamma_k}$ converges strongly to x. By passing to the limit inferior in (11) and using the lower semicontinuity of Ω, inequality (7) follows. □

3.1. Loss Functions Which Lead to Orthomonotonicity

Observe that part 1 of Theorem 3.1 (sufficiency of orthomonotonicity) only requires existence of minimizers of J, without any specific additional assumptions on the error term. On the other side, part 2 (necessity of orthomonotonicity) holds under additional assumptions on f. In the following, we provide examples of functions f that satisfy such assumptions, showing that most of the error functions considered in practice do so. The vast majority of error functions used are *additively separable*, namely of the form

$$f(z) = \sum_{i=1}^{m} V(z_i, y_i), \quad (14)$$

where $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ and $y_i \in \mathbb{R}$ are prescribed output data.

3.1.1. Regression loss functions

In this section we show that, for a broad class of regression loss functions, it is possible to find output data such that, if the family of regularization functionals (6) admits a representer theorem, then Ω is orthomonotone.

Definition 3.4. *We call the function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ a* regression loss function *if*

$$V(z, y) = \phi(z - y),$$

where $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous with bounded sublevel sets and minimized at zero.

The class of functions defined above includes any loss of the form $\phi(t) = |t|^p$ with $p > 0$ (in particular, square and absolute loss), the interpolation loss

$$\phi(t) = \begin{cases} 0 & \text{if } t = 0 \\ +\infty & \text{otherwise} \end{cases},$$

as well as the ε-insensitive loss $\phi(t) = \max\{0, |t| - \varepsilon\}$, which is not uniquely minimized at zero.

Lemma 3.1. *Assume that V is a regression loss function. Then, for every $p \in \mathbb{N}$, there exist output data $\{y_i : i \in \mathbb{N}_{2p}\} \subset \mathbb{R}$ and a function $f_u : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ satisfying the hypothesis of Theorem 3.1, Part 2, such that the error functional*

$$w \mapsto f(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle, \langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) ,$$

with f defined by (14) for $m = 2p$, equals the error functional

$$w \mapsto f_u(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) .$$

3.1.2. Binary classification loss functions

Definition 3.5. *We call the function $V : \mathbb{R} \times \{-1, +1\} \to \mathbb{R}$ a* regular binary classification loss function *if*

$$V(z, y) = \phi(yz),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is lower semicontinuous, nonincreasing, there exists $\alpha > 0$ such that the function

$$\psi_\alpha(t) = \phi(t) + \phi(-\alpha t),$$

admits a unique minimizer $\hat{t} \neq 0$ and there exists $\varepsilon > 0$ such that the sublevel set $\{t \in \mathbb{R} : \psi_\alpha(t) \leq \min \psi_\alpha + \varepsilon\}$ is bounded.

It can be seen easily that this definition is satisfied by most commonly used binary classification loss functions, including the logistic loss $\phi(t) = \log(1 + e^{-t})$, the exponential loss $\phi(t) = e^{-t}$ and the hinge loss $\phi(t) = \max\{0, 1 - t\}$. To verify the uniqueness of the minimizer for these three losses, choose for instance $\alpha = 1/2$.

Lemma 3.2. *Assume that V is a regular binary classification loss function. Then, for every $p \in \mathbb{N}$, there exist output data $\{y_i : i \in \mathbb{N}_{2p}\} \subseteq \{-1, +1\}$ and a function $f_u : \mathbb{R}^p \to \mathbb{R}$ satisfying the hypothesis of Theorem 3.1, Part 2, such that the error functional*

$$w \mapsto f(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle, \langle w, \alpha w_1\rangle, \ldots, \langle w, \alpha w_p\rangle) ,$$

with f defined by (14) for $m = 2p$, equals the error functional

$$w \mapsto f_u(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) .$$

3.2. Properties of Orthomonotone Functions

An obvious first fact about orthomonotone functions is that nesting of maps preserves orthomonotonicity.

Proposition 3.1. *If $S, S' : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ are such that $S(x) \subseteq S'(x)$ for all $x \in \mathcal{H}$, then any $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ orthomonotone with respect to S is also orthomonotone with respect to S'.*

Thus, enlarging the map S enlarges the class of orthomonotone functions as well. In the extreme case when S maps every point to $\mathcal{H}$, the orthomonotone class includes all functions. At the other extreme, S maps every point to $\{0\}$ and the orthomonotone class equals the set of constant functions.

A convenient way to obtain new orthomonotone functions (and hence new representer theorems) is by applying simple operations to known orthomonotone functions. For example, shifting the argument inside an orthomonotone function yields an orthomonotone function with respect to a larger map. This fact implies that Theorem 3.1 can be modified to apply to functions Ω that are minimized at points other than 0.

Proposition 3.2. *Let $a \in \mathcal{H}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ orthomonotone with respect to the map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. If S is quasilinear then the function $x \mapsto \Omega(x + a)$ is orthomonotone with respect to the map $x \mapsto S(x) + S(a)$.*

Another useful rule combines functions which are orthomonotone with respect to different maps.

Proposition 3.3. *Let $\Omega_1 : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be orthomonotone with respect to a map $S_1 : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ and $\Omega_2 : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be orthomonotone with respect to a map $S_2 : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. Also let $h : (\mathbb{R} \cup \{+\infty\})^2 \to \mathbb{R} \cup \{+\infty\}$ be elementwise nondecreasing, that is, $h(a', b') \geq h(a, b)$ whenever $a' \geq a$ and $b' \geq b$. Then the function $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$,*

$$\Omega(w) = h\left(\Omega_1(w), \Omega_2(w)\right), \qquad \forall w \in \mathcal{H},$$

is orthomonotone with respect to the map $S_1 + S_2$.

This rule holds more generally for any finite number of orthomonotone functions. In particular, any nonnegative linear combination of orthomonotone functions is also orthomonotone with respect to the sum of the corresponding maps. The same applies to the maximum and to the minimum of orthomonotone functions.

Finally, there is a composition rule for orthomonotone functions, similar to the chain rule for differentiation.

Proposition 3.4. *Let $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be orthomonotone with respect to a map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ and let $T \in \mathscr{L}(\mathcal{H})$ be a continuous operator. Then the function $\Omega \circ T$ is orthomonotone with respect to $T^* \circ S \circ T$.*

4. Examples of Representer Theorems

We now proceed to describe the set of orthomonotone functions for specific regularization problems of interest. For each problem, we describe the map S, provide a class of orthomonotone functions and state the resulting representer theorem.

Example 4.1. *Assume that the dimension of $\mathcal{H}$ is at least two and let S be defined as in Example 2.1. Then, the definition of representer theorem 3.1 reduces to the classical linear combination of the representers*

$$\hat{w} = \sum_{i=1}^{m} c_i w_i,$$

where $c_i \in \mathbb{R}$ and the definition of orthomonotonicity (5) reduces to

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

If Ω is lower-semicontinuous, this last condition is satisfied if and only if

$$\Omega(w) = h(\|w\|)$$

with $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ nondecreasing.

See Theorem 1 of (Dinuzzo & Schölkopf, 2012) and (Argyriou et al., 2009; Yu et al., 2013) for related results. This is a generalized version of the well known "classical" representer theorem (Girosi, 1998; Kimeldorf & Wahba, 1970; Schölkopf et al., 2001) which has found wide application to regularization methods in Hilbert spaces.

The case of regularization with a *bias term* can be recovered easily by choosing the error function f as a minimum with respect to the bias variable. This technique also yields *semiparametric theorems* (Schölkopf & Smola, 2002).

Example 4.2. *Let $\mathcal{H}$ and S be defined as in Example 2.3. Then the representer theorem 3.1 reduces to*

$$\hat{W} = \sum_{i=1}^{m} W_i C_i \,,$$

where C_i are matrices in $\mathbf{M}_n$ and the definition of orthomonotonicity (5) reduces to

$$\Omega(X + Y) \geq \Omega(X), \qquad \forall X, Y \in \mathbf{M}_{d,n} : \quad X^T Y = 0.$$

Moreover, in this case the orthomonotonicity property is equivalent to

$$\Omega(W) = h(W^\top W)$$

with $h : \mathbf{S}_+^n \to \mathbb{R} \cup \{+\infty\}$ being a matrix nondecreasing *function (with respect to the partial order of positive semidefinite matrices).*

See (Argyriou et al., 2009; Yu et al., 2013) as well as (Amit et al., 2007; Argyriou et al., 2008; 2010; Evgeniou et al., 2005) for special cases. The above representation extends the classical representer theorem to *matrix learning* problems, such as regularization with penalties involving the Frobenius norm, the trace norm and general spectral penalties. These methods have been used for multitask learning, collaborative filtering, kernel learning, domain adaptation and other problems. Problems like multitask learning benefit substantially from the representer theorem since in those cases the data matrices are rank-one (and in collaborative filtering they are also sparse). Indeed, whenever the data matrices are rank-one, that is, $W_i = a_i b_i^T$, we can let $v_i = b_i^T C_i$ and write $\hat{W} = \sum_{i=1}^{m} a_i v_i^T$, so that an equivalent optimization problem with substantially fewer degrees of freedom can be obtained. This last representation ensures that (4) is equivalent to an optimization problem whose number of variables is mn, which can be much smaller than dn, the size of matrix W.

It can also be seen that for other penalties of the type $\Omega(W) = g(R^\top W^\top G W R)$, with $R \in \mathbf{M}_{n,k}$, $G \in \mathbf{S}^d_{++}$ and g matrix nondecreasing, other representer theorems can be derived from the above result, by applying the change of variable $W' = G^{\frac{1}{2}} W$. These apply, for example, to spectral functions of QWR, with $Q \in \mathbf{M}_{\ell,d}$, which have been proposed for multi-task learning (Dinuzzo, 2013; Dinuzzo & Fukumizu, 2011).

In addition, Example 4.2 relates to certain optimization problems with *positive semidefinite* matrix variables. Indeed, problem (4) with $\mathcal{H} = \mathbf{M}_n$, $\Omega(W) = g(R^\top W^\top W R)$, g matrix nondecreasing and rank-one data, yields a problem of the type

$$\min\{f(y_1^\top Z x_1, \ldots, y_m^\top Z x_m) + \gamma g(R^\top Z R) : Z \in \mathbf{S}^n_+\} \tag{15}$$

by the change of variable $Z = W^\top W$. Thus a representer theorem for this family of problems follows directly from Example 4.2. Some results for special cases of (15), applied to metric and semisupervised learning, have already appeared in (Jain et al., 2010; 2012).

Example 4.3. *Assume $\mathcal{H} = \mathbf{M}_{d,n}$ equipped with the standard inner product, and suppose that S maps*

$$X \mapsto \{XC + DX : C \in \mathbf{M}_n, D \in \mathbf{M}_d\} .$$

Then S is nd-regular quasilinear. For this map, definition 3.1 reads

$$\hat{W} = \sum_{i=1}^{m} (W_i C_i + D_i W_i) ,$$

and the definition of orthomonotonicity (5) reduces to

$$\Omega(X+Y) \geq \Omega(X), \qquad \forall X, Y \in \mathbf{M}_{d,n} : \begin{cases} X^T Y = 0 \\ X Y^T = 0 \end{cases} ,$$

which is satisfied by all functions such that

$$\Omega(W) = h(W^T W, W W^T),$$

where $h : \mathbf{S}^n_+ \times \mathbf{S}^d_+ \to \mathbb{R} \cup \{+\infty\}$ is matrix nondecreasing *in each matrix argument.*

The family of regularizers described in the last example includes, for instance, functions of the form

$$\Omega(W) = \|QW\| + \|WR\| ,$$

where $\|\cdot\|$ is any orthogonally invariant norm. Such penalties are of considerable interest in many matrix learning problems, since they allow for incorporating information about both row and column dependencies, by designing the matrices Q and R. This can be applied, for instance, to collaborative filtering problems when side information about both users and items is available. When the data matrices are rank-one (such as in multi-task learning and collaborative filtering problems), the representer theorem above makes it again possible to obtain a significant reduction in the number of degrees of freedom, since the solution $\hat{W}$ can be rewritten in the form $\hat{W} = \sum_{i=1}^{m} (a_i v_i^T + u_i b_i^T)$, where the number of variables, $m(n+d)$, can be much smaller than nd, the size of W.

5. Conclusion

We have presented a framework which unifies existing results about representer theorems for regularization problems and allows for a more formal study of these results. We introduced a new definition of representer theorem to include a broader family of representation results. We showed that each theorem in this family corresponds to a regular quasilinear subspace-valued map. Moreover, we characterized the class of regularization penalties corresponding to each representer theorem via the orthomonotonicity property. Orthomonotone functions exhibit simple calculus rules, which can be used to obtain new representer theorems by combining existing ones.

Our new framework opens a number of possibilities for further investigation. First of all, it calls for more detailed characterizations of regular quasilinear subspace-valued maps and orthomonotone functions, given their importance in the mathematical construction that leads to the representer theorems. Secondly, it can lead to the derivation of new families of regularization penalties and corresponding methodologies, for example, for matrix and tensor regularization. Finally, it lays the foundation for a new and more general class of kernel methods, obtained by plugging the expression of the generalized representer theorem into the objective functional J and considering the resulting optimization problem.

References

Abernethy, J., Bach, F., Evgeniou, T., and Vert, J-P. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.

Amit, Y., Fink, M., Srebro, N., and Ullman, S. Uncovering shared structures in multiclass classification. In *Proceedings of the Twenty-Fourth International Conference on Machine learning*, 2007.

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

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

Argyriou, A., Micchelli, C.A., and Pontil, M. On spectral learning. *The Journal of Machine Learning Research*, 11:935–953, 2010.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

De Vito, E., Rosasco, L., Caponnetto, A., Piana, M., and Verri, A. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

Dinuzzo, F. Learning output kernels for multi-task problems. *Neurocomputing*, 118:119–126, 2013.

Dinuzzo, F. and Fukumizu, K. Learning low-rank output kernels. *Journal of Machine Learning Research*, 20:181–196, 2011.

Dinuzzo, F. and Schölkopf, B. The representer theorem for Hilbert spaces: a necessary and sufficient condition. In *Advances in Neural Information Processing Systems 25*, pp. 189–196, 2012.

Dinuzzo, F., Neve, M., De Nicolao, G., and Gianazza, U. P. On the representer theorem and equivalent degrees of freedom of SVR. *Journal of Machine Learning Research*, 8:2467–2495, 2007.

Evgeniou, T., Micchelli, C. A., and Pontil, M. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Girosi, F. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

Gnecco, G. and Sanguineti, M. Regularization techniques and suboptimal solutions to optimization problems in learning from data. *Neural Computation*, 22(3):793–829, 2010.

Jain, P., Kulis, B., and Dhillon, I. S. Inductive regularized learning of kernel functions. In *Advances in Neural Information Processing Systems*, pp. 946–954, 2010.

Jain, P., Kulis, B., Davis, J. V., and Dhillon, I. S. Metric and kernel learning using a linear transformation. *The Journal of Machine Learning Research*, 13:519–547, 2012.

Kimeldorf, G. S. and Wahba, G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

Kulis, B., Saenko, K., and Darrell, T. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1785–1792, 2011.

Lafferty, J., Zhu, X., and Liu, Y. Kernel conditional random fields: representation and clique selection. In *Proceedings of the Twenty-First International Conference on Machine learning*, pp. 64, 2004.

Micchelli, C. A. and Pontil, M. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25*, pp. 10–18. MIT Press, 2012.

Mukherjee, S. and Wu, Q. Estimation of gradients and coordinate covariation in classification. *The Journal of Machine Learning Research*, 7:2481–2514, 2006.

Pillai, N. S., Wu, Q., Liang, F., Mukherjee, S., and Wolpert, R. L. Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research*, 8:1769–1797, 2007.

Schölkopf, B. and Smola, A. J. *Learning with Kernels*. MIT Press, 2002.

Schölkopf, B., Herbrich, R., and Smola., A.J. A generalized representer theorem. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, 2001.

Signoretto, M., De Lathauwer, L., and Suykens, J. A. K. Learning tensors in reproducing kernel Hilbert spaces with multilinear spectral penalties. arXiv preprint arXiv:1310.4977, 2013.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G.R.G. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Steinwart, I. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

Warmuth, M., Kotłowski, W., and Zhou, S. Kernelization of matrix updates, when and how? In *Algorithmic Learning Theory*, pp. 350–364. Springer, 2012.

Yu, Y., Cheng, H., Schuurmans, D., and Szepesvari, C. Characterizing the representer theorem. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 570–578, 2013.

Zhang, H. and Zhang, J. Regularized learning in Banach spaces as an optimization problem: representer theorems. *Journal of Global Optimization*, 54(2):235–250, 2012.