

When is there a representer theorem? Vector versus matrix regularizers

Andreas Argyriou⁽¹⁾, Charles A. Micchelli⁽²⁾, Massimiliano Pontil⁽¹⁾

(1) Department of Computer Science
University College London
Gower Street, London WC1E, England, UK
E-mail: *m.pontil@cs.ucl.ac.uk*

(2) Department of Mathematics and Statistics
State University of New York
The University at Albany
1400 Washington Avenue, Albany, NY, 12222, USA
E-mail: *cam@math.albany.edu*

September 9, 2008

Abstract

We consider a general class of regularization methods which learn a vector of parameters on the basis of linear measurements. It is well known that if the regularizer is a nondecreasing function of the inner product then the learned vector is a linear combination of the input data. This result, known as the *representer theorem*, is at the basis of kernel-based methods in machine learning. In this paper, we prove the necessity of the above condition, thereby completing the characterization of kernel methods based on regularization. We further extend our analysis to regularization methods which learn a matrix, a problem which is motivated by the application to multi-task learning. In this context, we study a more general representer theorem, which holds for a larger class of regularizers. We provide a necessary and sufficient condition for these class of matrix regularizers and highlight them with some concrete examples of practical importance. Our analysis uses basic principles from matrix theory, especially the useful notion of matrix nondecreasing function.

1 Introduction

Regularization in Hilbert spaces is an important methodology for learning from examples and has a long history in a variety of fields. It has been studied, from different perspectives, in statistics [Wahba, 1990], in optimal estimation [Micchelli and Rivlin, 1985] and recently has been a focus of attention in machine learning theory – see, for example, [Cucker and Smale, 2001, De Vito et al., 2004, Micchelli and Pontil, 2005a, Shawe-Taylor and Cristianini, 2004, Vapnik, 2000] and references therein. Regularization is formulated as an *optimization problem* involving an *error term* and a *regularizer*. The regularizer plays an important role, in that it favors solutions with certain desirable properties. It has long been observed that certain regularizers exhibit an appealing property, called the *representer theorem*, which states that there exists a solution of the regularization problem that is a linear combination of the data [Wahba, 1990]. This property has important computational implications in the context of regularization with positive semidefinite *kernels*, because it makes high or infinite-dimensional problems of this type into finite dimensional problems of the size of the number of available data [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004].

The topic of interest in this paper will be to determine the conditions under which representer theorems hold. In the first half of the paper, we describe a property which a regularizer should satisfy in order to give rise to a representer theorem. It turns out that this property has a simple geometric interpretation and that the regularizer can be equivalently expressed as a *nondecreasing* function of the Hilbert space norm. Thus, we show that this condition, which has already been known to be sufficient for representer theorems, is also *necessary*. In the second half of the paper, we depart from the context of Hilbert spaces and focus on a class of problems in which a *matrix structure* plays an important role. For such problems, which have recently appeared in several machine learning applications, we show a modified version of the representer theorem that holds for a class of regularizers significantly larger than in the former context. As we shall see, these matrix regularizers are important in the context of multi-task learning: the matrix columns are the parameters of different regression tasks and the regularizer encourages certain dependences across the tasks.

In general, we consider problems in the framework of *Tikhonov regularization* [Tikhonov and Arsenin, 1977]. This regularization approach receives a set of input/output data $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{H} \times \mathcal{Y}$ and selects a vector in \mathcal{H} as the solution of an optimization problem. Here, \mathcal{H} is a prescribed Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle$ and $\mathcal{Y} \subseteq \mathbb{R}$ a set of possible output values. The optimization problems encountered in regularization are of the type

$$\min \{ \mathcal{E}(\langle w, x_1 \rangle, \dots, \langle w, x_m \rangle), (y_1, \dots, y_m) \} + \gamma \Omega(w) : w \in \mathcal{H} \}, \quad (1.1)$$

where $\gamma > 0$ is a regularization parameter. The function $\mathcal{E} : \mathbb{R}^m \times \mathcal{Y}^m \rightarrow \mathbb{R}$ is called an *error function* and $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ is called a *regularizer*. The error function measures the error on the data. Typically, it decomposes as a sum of univariate functions. For example, in regression, a common choice would be the sum of square errors, $\sum_{i=1}^m (\langle w, x_i \rangle - y_i)^2$. The function Ω , called the regularizer, favors certain regularity properties of the vector w (such as a small norm) and can be chosen based on available prior information about the target vector. In some Hilbert spaces such as Sobolev spaces the regularizer is measure of smoothness: the smaller the norm the smoother the function.

This framework includes several well-studied learning algorithms, such as ridge regression [Hoerl and Kennard, 1970], support vector machines [Boser et al., 1992], and many more – see [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004] and references therein.

An important aspect of the practical success of this approach is the observation that, for certain choices of the regularizer, solving (1.1) reduces to identifying m parameters and not $\dim(\mathcal{H})$. Specifically, when the regularizer is the square of the Hilbert space norm, the representer theorem holds: there exists a solution \hat{w} of (1.1) which is a linear combination of the input vectors,

$$\hat{w} = \sum_{i=1}^m c_i x_i, \quad (1.2)$$

where c_i are some real coefficients. This result is simple to prove and dates at least from the 1970’s, see, for example, [Kimeldorf and Wahba, 1970]. It is also known that it extends to any regularizer that is a *nondecreasing* function of the norm [Schölkopf et al., 2001]. Several other variants and results about the representation form (1.2) have also appeared in recent years [De Vito et al., 2004, Dinuzzo et al., 2007, Evgeniou et al., 2000, Girosi et al., 1995, Micchelli and Pontil, 2005b, Steinwart, 2003]. Moreover, the representer theorem has been important in machine learning, particularly within the context of learning in reproducing kernel Hilbert spaces [Aronszajn, 1950] – see [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004] and references therein.

Our first objective in this paper is to derive necessary and sufficient conditions for representer theorems to hold. Even though one is mainly interested in regularization problems, it is more convenient to study *interpolation* problems, that is, problems of the form

$$\min \{ \Omega(w) : w \in \mathcal{H}, \langle w, x_i \rangle = y_i, \forall i = 1, \dots, m \} . \quad (1.3)$$

Thus, we begin this paper (Section 2) by showing how representer theorems for interpolation and regularization relate. On one side, a representer theorem for interpolation easily implies such a theorem for regularization with the same regularizer and any error function. Therefore, *all representer theorems obtained in this paper apply equally to interpolation and regularization*. On the other side, though, the converse implication is true under certain weak qualifications on the error function.

Having addressed this issue, we concentrate in Section 3 on proving that an interpolation problem (1.3) admits solutions representable in the form (1.2) *if and only if* the regularizer is a *nondecreasing function of the Hilbert space norm*. That is, we provide a complete characterization of regularizers that give rise to representer theorems, which had been an open question. Furthermore, we discuss how our proof is motivated by a geometric understanding of the representer theorem, which is equivalently expressed as a monotonicity property of the regularizer.

Our second objective is to formulate and study the novel question of representer theorems for *matrix problems*. To make our discussion concrete, let us consider the problem of learning n linear regression vectors, represented by the parameters $w_1, \dots, w_n \in \mathbb{R}^d$, respectively. Each vector can be thought of as a “task” and the goal is to *jointly* learn these n tasks. In such problems, there is usually prior knowledge that *relates* these tasks and it is often the case that learning can improve if this knowledge is appropriately taken into account. Consequently, a good regularizer should favor such task relations and involve *all tasks jointly*.

In the case of interpolation, this learning framework can be formulated concisely as

$$\min \{ \Omega(W) : W \in \mathbf{M}_{d,n}, w_t^\top x_{ti} = y_{ti} \quad \forall i = 1, \dots, m_t, t = 1, \dots, n \}, \quad (1.4)$$

where $\mathbf{M}_{d,n}$ denotes the set of $d \times n$ real matrices and the column vectors $w_1, \dots, w_n \in \mathbb{R}^d$ form the matrix W . Each task t has its own input data $x_{t1}, \dots, x_{tm_t} \in \mathbb{R}^d$ and corresponding output values $y_{t1}, \dots, y_{tm_t} \in \mathcal{Y}$.

An important feature of such problems that distinguishes them from the type (1.3) is the appearance of *matrix products* in the constraints, unlike the inner products in (1.3). In fact, as we will discuss in Section 4.1, problems of the type (1.4) can be written in the form (1.3). Consequently, the representer theorem applies if the matrix regularizer is a nondecreasing function of the Frobenius norm¹. However, the optimal vector \hat{w}_t for each task can be represented as a linear combination of *only those input vectors corresponding to this particular task*. Moreover, with such regularizers it is easy to see that each task in (1.4) can be optimized independently. Hence, these regularizers are of no practical interest if the tasks are expected to be related.

This observation leads us to formulate a *modified representer theorem*, which is appropriate for matrix problems, namely,

$$\hat{w}_t = \sum_{s=1}^n \sum_{i=1}^{m_s} c_{si}^{(t)} x_{si} \quad \forall t = 1, \dots, n, \quad (1.5)$$

where $c_{si}^{(t)}$ are scalar coefficients, for $t, s = 1, \dots, n, i = 1, \dots, m_s$. In other words, we now allow for *all input vectors* to be present in the linear combination representing each column of the optimal matrix. As a result, this definition greatly expands the class of regularizers that give rise to representer theorems.

Moreover, this framework can be applied to many applications where matrix optimization problems are involved. Our immediate motivation, however, has been more specific than that, namely *multi-task learning*. Learning multiple tasks jointly has been a growing area of interest in machine learning, especially during the past few years [Abernethy et al., 2006, Argyriou et al., 2006, 2007a,b, Candès and Recht, 2008, Cavallanti et al., 2008, Izenman, 1975, Maurer, 2006a,b, Srebro et al., 2005, Wolf et al., 2007, Xiang and Bennett, 2005, Yuan et al., 2007]. For instance, some of these works use regularizers which involve the *trace norm*² of matrix W . The general idea behind this methodology is that a small trace norm favors low-rank matrices. This means that the tasks (the columns of W) are related in that they all lie in a low-dimensional subspace of \mathbb{R}^d . In the case of the trace norm, the representer theorem (1.5) is known to hold – see [Abernethy et al., 2006, Argyriou et al., 2007a, Amit et al., 2007], also discussed in Section 4.1.

It is natural, therefore, to ask a question similar to that in the standard Hilbert space (or single-task) setting. That is, under which conditions on the regularizer a representer theorem holds. In Section 4.2, we provide an answer by *proving a necessary and sufficient condition for representer theorems to hold, expressed as a simple monotonicity property*. This property is analogous to the one in the Hilbert space setting, but its geometric interpretation is now algebraic in nature. We also give a functional description equivalent to this property, that is, *we show that the regularizers of interest are the matrix nondecreasing functions of the quantity $W^\top W$* .

¹Defined as $\|W\|_2 = \sqrt{\text{tr}(W^\top W)}$.

²Equal to the sum of the singular values of W .

Our results cover matrix problems of the type (1.4) which have already been studied in the literature. But they also point towards some new learning methods that may perform well in practice and can now be made computationally efficient. Thus, we close the paper with a discussion of possible regularizers that satisfy our conditions and have been used or can be used in the future in machine learning problems.

1.1 Notation

Before proceeding, we introduce the notation used in this paper. We use \mathbb{N}_d as a shorthand for the set of integers $\{1, \dots, d\}$. We use \mathbb{R}^d to denote the linear space of vectors with d real components. The standard inner product in this space is denoted by $\langle \cdot, \cdot \rangle$, that is, $\langle w, v \rangle = \sum_{i \in \mathbb{N}_d} w_i v_i$, $\forall w, v \in \mathbb{R}^d$, where w_i, v_i are the i -th components of w, v respectively. More generally, we will consider Hilbert spaces which we will denote by \mathcal{H} , equipped with an inner product $\langle \cdot, \cdot \rangle$.

We also let $\mathbf{M}_{d,n}$ be the linear space of $d \times n$ real matrices. If $W, Z \in \mathbf{M}_{d,n}$ we define their Frobenius inner product as $\langle W, Z \rangle = \text{tr}(W^\top Z)$, where tr denotes the trace of a matrix. With \mathbf{S}^d we denote the set of $d \times d$ real symmetric matrices and with \mathbf{S}_+^d (\mathbf{S}_{++}^d) its subset of positive semidefinite (definite) ones. We use \succ and \succeq for the positive definite and positive semidefinite partial orderings, respectively. Finally, we let \mathbf{O}^d be the set of $d \times d$ orthogonal matrices.

2 Regularization versus Interpolation

The line of attack we shall follow in this paper will go through *interpolation*. That is, our main concern will be to obtain necessary and sufficient conditions for representer theorems that hold for interpolation problems. However, in practical applications one encounters *regularization* problems more frequently than interpolation problems.

First of all, the family of the former problems is more general than that of the latter ones. Indeed, an interpolation problem can be simply obtained in the limit as the *regularization parameter* goes to zero [Micchelli and Pinkus, 1994]. More importantly, regularization enables one to trade off interpolation of the data against smoothness or simplicity of the model, whereas interpolation frequently suffers from *overfitting*.

Thus, frequently one considers problems of the form

$$\min \left\{ \mathcal{E}(\langle w, x_1 \rangle, \dots, \langle w, x_m \rangle), (y_1, \dots, y_m) \right\} + \gamma \Omega(w) : w \in \mathcal{H} \quad , \quad (2.1)$$

where $\gamma > 0$ is called the regularization parameter. This parameter is not known in advance but can be tuned with techniques like *cross validation* [Wahba, 1990]. Here, $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ is a *regularizer*, $\mathcal{E} : \mathbb{R}^m \times \mathcal{Y}^m \rightarrow \mathbb{R}$ is an error function and $x_i \in \mathcal{H}, y_i \in \mathcal{Y}, \forall i \in \mathbb{N}_m$, are given input and output data. The set \mathcal{Y} is a subset of \mathbb{R} and varies depending on the context, so that it is typically assumed equal to \mathbb{R} in the case of regression or equal to $\{-1, 1\}$ in binary classification problems. One may also consider the associated interpolation problem, which is

$$\min \left\{ \Omega(w) : w \in \mathcal{H}, \langle w, x_i \rangle = y_i, \forall i \in \mathbb{N}_m \right\} . \quad (2.2)$$

Under certain assumptions, the minima in problems (2.1) and (2.2) are attained (whenever the constraints in (2.2) are satisfiable). Such assumptions could involve, for example, lower semi-continuity and boundedness of sublevel sets for Ω and boundedness from below for \mathcal{E} . These

issues will not concern us here, as we shall assume the following about the error function \mathcal{E} and the regularizer Ω , from now on.

Assumption 2.1. *The minimum (2.1) is attained for any $\gamma > 0$, any input and output data $\{x_i, y_i : i \in \mathbb{N}_m\}$ and any $m \in \mathbb{N}$. The minimum (2.2) is attained for any input and output data $\{x_i, y_i : i \in \mathbb{N}_m\}$ and any $m \in \mathbb{N}$, whenever the constraints in (2.2) are satisfiable..*

The main objective of this paper is to obtain *necessary and sufficient* conditions on Ω so that the solution of problem (2.1) satisfies a *linear representer theorem*.

Definition 2.1. *We say that a class of optimization problems such as (2.1) or (2.2) satisfies the linear representer theorem if, for any choice of data $\{x_i, y_i : i \in \mathbb{N}_m\}$ such that the problem has a solution, there exists a solution that belongs to $\text{span}\{x_i : i \in \mathbb{N}_m\}$.*

In this section, we show that the existence of representer theorems for regularization problems is equivalent to the existence of representer theorems for interpolation problems, under a quite general condition that has a rather simple geometric interpretation.

We first recall a lemma from [Micchelli and Pontil, 2004, Sec. 2] which states that (linear or not) representer theorems for interpolation lead to representer theorems for regularization, under no conditions on the error function.

Lemma 2.1. *Let $\mathcal{E} : \mathbb{R}^m \times \mathcal{Y}^m \rightarrow \mathbb{R}$, $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ satisfying Assumption 2.1. Then if the class of interpolation problems (2.2) satisfies the linear representer theorem, so does the class of regularization problems (2.1).*

Proof. Consider a problem of the form (2.1) and let \hat{w} be a solution. We construct an associated interpolation problem

$$\min \{ \Omega(w) : w \in \mathcal{H}, \langle w, x_1 \rangle = \langle \hat{w}, x_1 \rangle, \dots, \langle w, x_m \rangle = \langle \hat{w}, x_m \rangle \} . \quad (2.3)$$

By hypothesis, there exists a solution \tilde{w} of (2.3) that lies in $\text{span}\{x_i : i \in \mathbb{N}_m\}$. But then $\Omega(\tilde{w}) \leq \Omega(\hat{w})$ and hence \tilde{w} is a solution of (2.1) and the result follows. \blacksquare

This lemma requires no special properties of the functions involved. Its converse, in contrast, requires assumptions about the analytical properties of the error function. We provide one such natural condition in the theorem below, but other conditions could conceivably work too. The main idea in the proof is, based on a single input, to construct a sequence of appropriate regularization problems for different values of the regularization parameter γ . Then, it suffices to show that letting $\gamma \rightarrow 0^+$ yields a limit of the minimizers that satisfies an interpolation constraint.

Theorem 2.1. *Let $\mathcal{E} : \mathbb{R}^m \times \mathcal{Y}^m \rightarrow \mathbb{R}$ and $\Omega : \mathcal{H} \rightarrow \mathbb{R}$. Assume that \mathcal{E}, Ω are lower semi-continuous, that Ω has bounded sublevel sets and that \mathcal{E} is bounded from below. Assume also that, for some $v \in \mathbb{R}^m \setminus \{0\}, y \in \mathcal{Y}^m$, there exists a unique minimizer of $\min\{\mathcal{E}(av, y) : a \in \mathbb{R}\}$ and that this minimizer does not equal zero. Then if the class of regularization problems (2.1) satisfies the linear representer theorem, so does the class of interpolation problems (2.2).*

Proof. Fix an arbitrary $x \neq 0$ and let a_0 be the minimizer of $\min\{\mathcal{E}(av, y) : a \in \mathbb{R}\}$. Consider the problems

$$\min \left\{ \mathcal{E} \left(\frac{a_0}{\|x\|^2} \langle w, x \rangle v, y \right) + \gamma \Omega(w) : w \in \mathcal{H} \right\},$$

for every $\gamma > 0$, and let w_γ be a solution in the span of x (known to exist by hypothesis). We then obtain that

$$\mathcal{E}(a_0 v, y) + \gamma \Omega(w_\gamma) \leq \mathcal{E} \left(\frac{a_0}{\|x\|^2} \langle w_\gamma, x \rangle v, y \right) + \gamma \Omega(w_\gamma) \leq \mathcal{E}(a_0 v, y) + \gamma \Omega(x). \quad (2.4)$$

Thus, $\Omega(w_\gamma) \leq \Omega(x)$ and so, by the hypothesis on Ω , the set $\{w_\gamma : \gamma > 0\}$ is bounded. Therefore, there exists a convergent subsequence $\{w_{\gamma_\ell} : \ell \in \mathbb{N}\}$, with $\gamma_\ell \rightarrow 0^+$, whose limit we call \bar{w} . By taking the limits as $\ell \rightarrow \infty$ on the inequality on the right in (2.4), we obtain

$$\mathcal{E} \left(\frac{a_0}{\|x\|^2} \langle \bar{w}, x \rangle v, y \right) \leq \mathcal{E}(a_0 v, y)$$

and consequently

$$\frac{a_0}{\|x\|^2} \langle \bar{w}, x \rangle = a_0$$

or

$$\langle \bar{w}, x \rangle = \|x\|^2.$$

In addition, since w_γ belongs to the span of x for every $\gamma > 0$, so does \bar{w} . Thus, we obtain that $\bar{w} = x$. Moreover, from the definition of w_γ we have that

$$\mathcal{E} \left(\frac{a_0}{\|x\|^2} \langle w_\gamma, x \rangle v, y \right) + \gamma \Omega(w_\gamma) \leq \mathcal{E}(a_0 v, y) + \gamma \Omega(w) \quad \forall w \in \mathcal{H} \text{ such that } \langle w, x \rangle = \|x\|^2$$

and, combining with the definition of a_0 , that

$$\Omega(w_\gamma) \leq \Omega(w) \quad \forall w \in \mathcal{H} \text{ such that } \langle w, x \rangle = \|x\|^2$$

Taking the limits as $\ell \rightarrow \infty$, we conclude that $\bar{w} = x$ is a solution of the problem

$$\min\{\Omega(w) : w \in \mathcal{H}, \langle w, x \rangle = \|x\|^2\}.$$

Moreover, this assertion holds even when $x = 0$, since the hypothesis implies that 0 is a global minimizer of Ω . Indeed, any regularization problem of the type (2.1) with zero inputs, $x_i = 0, \forall i \in \mathbb{N}_m$, admits a solution in their span. Thus, we have shown that Ω satisfies property (3.3) and the result follows immediately from Lemma 3.1. \blacksquare

We now comment on some commonly used error functions. The first is the *square loss*,

$$\mathcal{E}(z, y) = \sum_{i \in \mathbb{N}_m} (z_i - y_i)^2,$$

for $z, y \in \mathbb{R}^m$. It is immediately apparent that Theorem 2.1 applies in this case.

The second case is the *hinge loss*,

$$\mathcal{E}(z, y) = \sum_{i \in \mathbb{N}_m} \max(1 - z_i y_i, 0),$$

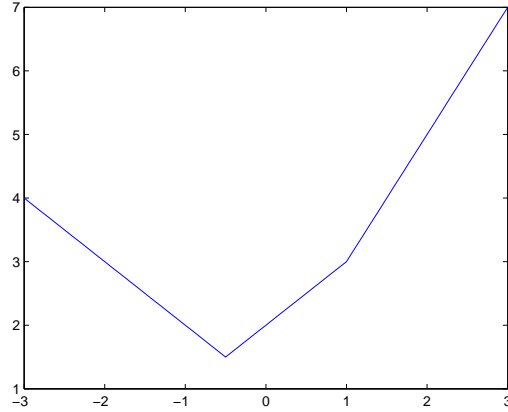


Figure 1: Hinge loss along the direction $(1, -2, 0, \dots, 0)$.

where the outputs y_i are assumed to belong to $\{-1, 1\}$ for the purpose of classification. In this case, we may select $y_i = 1, \forall i \in \mathbb{N}_m$, and $v = (-1, -2, 0, \dots, 0)^\top$ for $m \geq 2$. Then the function $\mathcal{E}(\cdot, v, y)$ is the one shown in Figure 1.

Finally, the *logistic loss*,

$$\mathcal{E}(z, y) = \sum_{i \in \mathbb{N}_m} \log(1 + e^{-z_i y_i}),$$

is also used in classification problems. In this case, we may select $y_i = 1, \forall i \in \mathbb{N}_m$, and $v = (2, -1)^\top$ for $m = 2$ or $v = (m - 2, -1, \dots, -1)^\top$ for $m > 2$. In the latter case, for example, setting to zero the derivative of $\mathcal{E}(\cdot, v, y)$ yields the equation $(m - 1)e^{a(m-1)} + e^a - m + 2 = 0$, which can easily be seen to have a unique solution.

Summarizing, we obtain the following corollary.

Corollary 2.1. *If $\mathcal{E} : \mathbb{R}^m \times \mathcal{Y}^m \rightarrow \mathbb{R}$ is the square loss, the hinge loss (for $m \geq 2$) or the logistic loss (for $m \geq 2$) and $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ is lower semi-continuous with bounded sublevel sets, then the class of problems (2.1) satisfies the linear representer theorem if and only if the class of problems (2.2) does.*

Note also that the condition on \mathcal{E} in Theorem 2.1 is rather weak in that an error function \mathcal{E} may satisfy it without being convex. At the same time, an error function that is “too flat”, such as a constant loss, will not do.

We conclude with a remark about the situation in which the inputs x_i are *linearly independent*.³ It has a brief and straightforward proof, which we do not present here.

Remark 2.1. *Let \mathcal{E} be the hinge loss or the logistic loss and $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ be of the form $\Omega(w) = h(\|w\|)$, where $h : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a lower semi-continuous function with bounded sublevel sets. Then the class of regularization problems (2.1) in which the inputs $x_i, i \in \mathbb{N}_m$, are linearly independent, satisfies the linear representer theorem.*

³This occurs frequently in practice, especially when the dimensionality d is high.

3 Representer Theorems for Interpolation Problems

The results of the previous section allow us to focus on linear representer theorems for interpolation problems of the type (2.2). We are going to consider the case of a Hilbert space \mathcal{H} as the domain of an interpolation problem. Interpolation constraints will be formed as inner products of the variable with the input data. For all purposes in this context, it makes no difference to think of \mathcal{H} as being equal to \mathbb{R}^d .

In this section, we consider the interpolation problem

$$\min\{\Omega(w) : w \in \mathcal{H}, \langle w, x_i \rangle = y_i, i \in \mathbb{N}_m\}, \quad (3.1)$$

We coin the term *admissible* to denote the class of regularizers we are interested in.

Definition 3.1. *We say that the function $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ is admissible if, for every $m \in \mathbb{N}$ and any data set $\{(x_i, y_i) : i \in \mathbb{N}_m\} \subseteq \mathcal{H} \times \mathcal{Y}$ such that the interpolation constraints are satisfiable, problem (3.1) admits a solution \hat{w} of the form*

$$\hat{w} = \sum_{i \in \mathbb{N}_m} c_i x_i,$$

where c_i are some real parameters.

We say that $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ is differentiable if, for every $w \in \mathcal{H}$, there is a unique vector denoted by $\nabla\Omega(w)$, such that for all $p \in \mathcal{H}$,

$$\lim_{t \rightarrow 0} \frac{\Omega(w + tp) - \Omega(w)}{t} = \langle \nabla\Omega(w), p \rangle.$$

This notion corresponds to the usual notion of directional derivative on \mathbb{R}^d and in that case $\nabla\Omega(w)$ is the gradient of Ω at w .

In the remainder of the section, we always assume that Assumption 2.1 holds for Ω . The following theorem provides a necessary and sufficient condition for a regularizer to be admissible.

Theorem 3.1. *Let $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ be a differentiable function and $\dim(\mathcal{H}) \geq 2$. Then Ω is admissible if and only if*

$$\Omega(w) = h(\langle w, w \rangle) \quad \forall w \in \mathcal{H}, \quad (3.2)$$

for some nondecreasing function $h : \mathbb{R}_+ \rightarrow \mathbb{R}$.

It is well known that the above functional form is sufficient for a representer theorem to hold (see for example [Schölkopf et al., 2001]). Here we show that it is also necessary.

The route we follow to proving the above theorem is based on a geometric interpretation of representer theorems. This intuition can be formally expressed as condition (3.3) in the lemma below. Both condition (3.3) and functional form (3.2) express the property that the contours of Ω are *spheres* (or regions between spheres), which is apparent from Figure 2.

Lemma 3.1. *A function $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ is admissible if and only if it satisfies the property that*

$$\Omega(w + p) \geq \Omega(w) \quad \forall w, p \in \mathcal{H} \text{ such that } \langle w, p \rangle = 0. \quad (3.3)$$

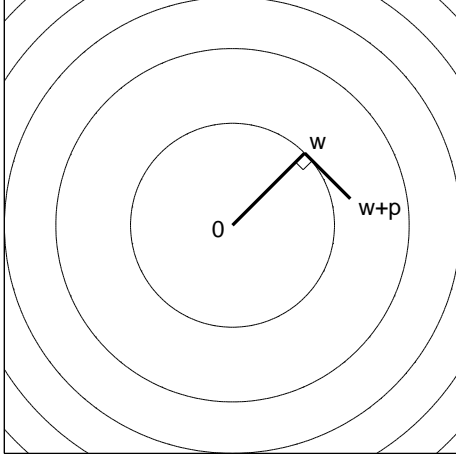


Figure 2: Geometric interpretation of Theorem 3.1. The function Ω should not decrease when moving to orthogonal directions. The contours of such a function should be spherical.

Proof. Suppose that Ω satisfies property (3.3), consider arbitrary data $x_i, y_i, i \in \mathbb{N}_m$, and let \hat{w} be a solution to problem (3.1). We can uniquely decompose \hat{w} as $\hat{w} = \bar{w} + p$ where $\bar{w} \in \mathcal{L} := \text{span}\{x_i : i \in \mathbb{N}_m\}$ and $p \in \mathcal{L}^\perp$. From (3.3) we obtain that $\Omega(\hat{w}) \geq \Omega(\bar{w})$. Also \bar{w} satisfies the interpolation constraints and hence we conclude that \bar{w} is a solution to problem (3.1).

Conversely, if Ω is admissible choose any $w \in \mathcal{H}$ and consider the problem $\min\{\Omega(z) : z \in \mathcal{H}, \langle z, w \rangle = \langle w, w \rangle\}$. By hypothesis, there exists a solution belonging in $\text{span}\{w\}$ and hence w is a solution to this problem. Thus, we have that $\Omega(w+p) \geq \Omega(w)$ for every p such that $\langle w, p \rangle = 0$. ■

It remains to establish the equivalence of the geometric property (3.3) to condition (3.2) that Ω is a nondecreasing function of the L_2 norm.

Proof of Theorem 3.1. Assume first that (3.3) holds and $\dim(\mathcal{H}) < \infty$. In this case, we only need to consider the case that $\mathcal{H} = \mathbb{R}^d$ since (3.3) can always be rewritten as an equivalent condition on \mathbb{R}^d , using an orthonormal basis of \mathcal{H} .

First we observe that, since Ω is differentiable, this property implies the condition that

$$\langle \nabla \Omega(w), p \rangle = 0, \quad (3.4)$$

for all $w, p \in \mathbb{R}^d$ such that $\langle w, p \rangle = 0$.

Now, fix any $w_0 \in \mathbb{R}^d$ such that $\|w_0\| = 1$. Consider an arbitrary $w \in \mathbb{R}^d$. Then there exists an orthogonal matrix $U \in \mathbf{O}^d$ such that $w = \|w\|Uw_0$ and $\det(U) = 1$ (see Lemma 5.1 in the appendix). Moreover, we can write $U = e^D$ for some skew-symmetric matrix $D \in \mathbf{M}_{d,d}$ — see [Horn and Johnson, 1991, Example 6.2.15]. Consider now the path $z : [0, 1] \rightarrow \mathbb{R}^d$ with

$$z(\lambda) = \|w\|e^{\lambda D}w_0 \quad \forall \lambda \in [0, 1].$$

We have that $z(0) = \|w\|w_0$ and $z(1) = w$. Moreover, since $\langle z(\lambda), z(\lambda) \rangle = \langle w, w \rangle$, we obtain that

$$\langle z'(\lambda), z(\lambda) \rangle = 0 \quad \forall \lambda \in [0, 1].$$

Applying (3.4) with $w = z(\lambda)$, $p = z'(\lambda)$, it follows that

$$\frac{d\Omega(z(\lambda))}{d\lambda} = \langle \nabla\Omega(z(\lambda)), z'(\lambda) \rangle = 0.$$

Consequently, $\Omega(z(\lambda))$ is constant and hence $\Omega(\|w\|w_0) = \Omega(w)$. Setting $h(\xi) = \Omega(\sqrt{\xi}w_0)$, $\forall \xi \in \mathbb{R}_+$, yields (3.2). In addition, h must be nondecreasing in order for Ω to satisfy property (3.3).

For the case $\dim(\mathcal{H}) = \infty$ we can argue similarly using instead the path

$$z(\lambda) = \frac{(1-\lambda)w_0 + \lambda w}{\|(1-\lambda)w_0 + \lambda w\|} \|w\|$$

which is differentiable on $[0, 1]$ when $w \notin \text{span}\{w_0\}$. We confirm equation (3.2) for vectors in $\text{span}\{w_0\}$ by a limiting argument on vectors not in $\text{span}\{w_0\}$ since Ω is surely continuous.

Conversely, if $\Omega(w) = h(\langle w, w \rangle)$ and h is nondecreasing, property (3.3) follows immediately. \blacksquare

We note that we could modify Definition 3.1 by requiring that *any* solution of problem (3.1) be in the linear span of the input data. We call such regularizers *strictly admissible*. Then with minor modifications to Lemma 3.1 (namely, requiring that equality in (3.3) holds only if $p = 0$) and to the proof of Theorem 3.1 (namely, requiring h to be strictly increasing) we have the following corollary.

Corollary 3.1. *Let $\Omega : \mathcal{H} \rightarrow \mathbb{R}$ be a differentiable function. Then Ω is strictly admissible if and only if $\Omega(w) = h(\langle w, w \rangle)$, $\forall w \in \mathcal{H}$, where $h : \mathbb{R}_+ \rightarrow \mathbb{R}$ is strictly increasing.*

Theorem 3.1 can be used to verify whether the linear representer theorem can be obtained when using a regularizer Ω . For example, the function $\|w\|_p = (\sum_{i \in \mathbb{N}_d} |w_i|^p)^{\frac{1}{p}}$ is not admissible for any $p \geq 0$, $p \neq 2$, because it cannot be expressed as a function of the Hilbert space norm. Indeed, if we choose any $a \in \mathbb{R}$ and let $w = (a\delta_{i1} : i \in \mathbb{N}_d)$, the requirement that $\|w\|_p = h(\langle w, w \rangle)$ would imply that $h(a^2) = |a|$, $\forall a \in \mathbb{R}$, and hence that $\|w\|_p = \|w\|$.

4 Matrix Learning Problems

In this section, we investigate how representer theorems and results like Theorem 3.1 can be extended in the context of optimization problems that involve matrices.

4.1 Exploiting Matrix Structure

As we have already seen, our discussion in Section 3 applies to any Hilbert space. Thus, we may consider the finite Hilbert space of $d \times n$ matrices $\mathbf{M}_{d,n}$ equipped with the Frobenius inner product $\langle \cdot, \cdot \rangle$. As in Section 3, we could consider interpolation problems of the form

$$\min \{ \Omega(W) : W \in M_{d,n}, \langle W, X_i \rangle = y_i, i \in \mathbb{N}_m \} \quad (4.1)$$

where $X_i \in M_{d,n}$ are prescribed input matrices and $y_i \in \mathcal{Y}$ scalar outputs, for $i \in \mathbb{N}_m$. Then Theorem 3.1 states that such a problem admits a solution of the form

$$\hat{W} = \sum_{i \in \mathbb{N}_m} c_i X_i, \quad (4.2)$$

where c_i are some real parameters, if and only if Ω can be written in the form

$$\Omega(W) = h(\langle W, W \rangle) \quad \forall W \in \mathbf{M}_{d,n}, \quad (4.3)$$

where $h : \mathbb{R}_+ \rightarrow \mathbb{R}$ is nondecreasing.

However, optimization problems of the form (4.1) do not occur frequently in machine learning practice. The constraints of (4.1) do not utilize the structure inherent in matrices – that is, it makes no difference whether the variable is regarded as a matrix or as a vector – and hence have limited applicability. In contrast, in many recent applications, some of which we shall briefly discuss below, it is natural to consider problems like

$$\min \{ \Omega(W) : W \in \mathbf{M}_{d,n}, w_t^\top x_{ti} = y_{ti} \quad \forall i \in \mathbb{N}_{m_t}, t \in \mathbb{N}_n \}. \quad (4.4)$$

Here, $w_t \in \mathbb{R}^d$ denote the columns of matrix W , for $t \in \mathbb{N}_n$, and $x_{ti} \in \mathbb{R}^d, y_{ti} \in \mathcal{Y}$ are prescribed inputs and outputs, for $i \in \mathbb{N}_{m_t}, t \in \mathbb{N}_n$. In addition, the desired representation form for solutions of such matrix problems is different from (4.2). In this case, one may encounter representer theorems of the form

$$\hat{w}_t = \sum_{s \in \mathbb{N}_n} \sum_{i \in \mathbb{N}_{m_s}} c_{si}^{(t)} x_{si} \quad \forall t \in \mathbb{N}_n, \quad (4.5)$$

where $c_{si}^{(t)}$ are scalar coefficients for $s, t \in \mathbb{N}_n, i \in \mathbb{N}_{m_s}$.

To illustrate the above, consider the problem of multi-task learning and problems closely related to it [Abernethy et al., 2006, Argyriou et al., 2006, 2007a,b, Candès and Recht, 2008, Cavallanti et al., 2008, Izenman, 1975, Maurer, 2006a,b, Srebro et al., 2005, Yuan et al., 2007, etc.]. In learning multiple tasks jointly, each task may be represented by a vector of regression parameters that corresponds to the column w_t in our notation. There are n tasks and m_t data examples $\{(x_{ti}, y_{ti}) : i \in \mathbb{N}_{m_t}\}$ for the t -th task. The learning algorithm used is

$$\min \{ \mathcal{E}(w_t^\top x_{ti}, y_{ti} : i \in \mathbb{N}_{m_t}, t \in \mathbb{N}_n) + \gamma \Omega(W) : W \in \mathbf{M}_{d,n} \}, \quad (4.6)$$

where $\mathcal{E} : \mathbb{R}^M \times \mathcal{Y}^M \rightarrow \mathbb{R}, M = \sum_{t \in \mathbb{N}_n} m_t$. The error term expresses the objective that the regression vector for each task should fit well the data for this particular task. The choice of the regularizer Ω is important in that it captures certain relationships between the tasks. One common choice is the *trace norm*, which is defined to be the sum of the singular values of a matrix or, equivalently,

$$\Omega(W) = \|W\|_1 := \text{tr}(W^\top W)^{\frac{1}{2}}.$$

Regularization with the trace norm learns the tasks as one joint optimization problem, by favoring matrices with low rank. In other words, the vectors w_t are related in that they are *all* linear combinations of a *small* set of basis vectors. It has been demonstrated that this approach allows for accurate estimation of related tasks even when there are only *few* data points available for each task.

Thus, it is natural to consider optimization problems of the form (4.4). In fact, these problems can be seen as instances of problems of the form (4.1), because the quantity $w_t^\top x_{ti}$ can be written as the inner product between W and a matrix having all its columns equal to zero except for the t -th column being equal to x_{ti} . It is also easy to see that (4.1) is a richer class since the corresponding constraints are less restrictive.

Despite this fact, by focusing on the class (4.4) we concentrate on problems of more practical interest and we can obtain representer theorems for a richer class of regularizers, which includes the trace norm and other useful functions. In contrast, regularization with the functional form (4.3) is not a satisfactory approach since it ignores matrix structure. In particular, regularization with the Frobenius norm (and a separable error function) corresponds to learning each task *independently*, ignoring relationships among the tasks.

A representer theorem of the form (4.5) for regularization with the trace norm has been shown in [Argyriou et al., 2007a]. Related results have also appeared in [Abernethy et al., 2006, Amit et al., 2007]. We repeat here the statement and the proof of this theorem, in order to better motivate our proof technique of Section 4.2.

Theorem 4.1. *If Ω is the trace norm then problem (4.4) (or problem (4.6)) admits a solution \hat{W} of the form (4.5), for some $c_{si}^{(t)} \in \mathbb{R}$, $i \in \mathbb{N}_{m_s}$, $s, t \in \mathbb{N}_n$.*

Proof. Let \hat{W} be a solution of (4.4) and let $\mathcal{L} := \text{span}\{x_{si} : s \in \mathbb{N}_n, i \in \mathbb{N}_{m_s}\}$. We can decompose the columns of \hat{W} as $\hat{w}_t = \bar{w}_t + p_t$, $\forall t \in \mathbb{N}_n$, where $\bar{w}_t \in \mathcal{L}$ and $p_t \in \mathcal{L}^\perp$. Hence $\hat{W} = \bar{W} + P$, where \bar{W} is the matrix with columns \bar{w}_t and P is the matrix with columns p_t . Moreover we have that $P^\top \bar{W} = 0$. From Lemma 5.2 in the appendix, we obtain that $\|\hat{W}\|_1 \geq \|\bar{W}\|_1$. We also have that $\langle w_t, x_{ti} \rangle = \langle \bar{w}_t, x_{ti} \rangle$, for every $i \in \mathbb{N}_{m_t}$, $t \in \mathbb{N}_n$. Thus, \bar{W} preserves the interpolation constraints (or the value of the error term) while not increasing the value of the regularizer. Hence, it is a solution of the optimization problem and the assertion follows. \blacksquare

A simple but important observation about this and related results is that each task vector w_t is a linear combination of the data for *all* the tasks. This contrasts to the representation form (4.2) obtained by using Frobenius inner product constraints. Interpreting (4.2) in a multi-task context, by appropriately choosing the X_i as described above, would imply that each w_t is a linear combination of only the data for task t .

Finally, in some applications the following variant, similar to the type (4.4), has appeared,

$$\min \{ \Omega(W) : W \in \mathbf{M}_{d,n}, w_t^\top x_i = y_{ti} \quad \forall i \in \mathbb{N}_m, t \in \mathbb{N}_n \} . \quad (4.7)$$

Problems of this type corresponds to a special case in multi-task learning applications in which the input points are the same for all the tasks. For instance, this is the case with collaborative filtering or applications in marketing where the same products/entities are rated by all users/consumers (see, for example, [Aaker et al., 2004, Evgeniou et al., 2005, Lenk et al., 1996, Srebro et al., 2005] for various approaches to this problem).

4.2 Characterization of Matrix Regularizers

Our objective in this section will be to state and prove a general representer theorem for problems of the form (4.4) or (4.7) using a functional form analogous to (3.2). The key insight used in the proof of [Argyriou et al., 2007a] has been that the trace norm is defined in terms of a matrix function that preserves the partial ordering of matrices. That is, it satisfies Lemma 5.2, which is a matrix analogue of the geometric property (3.3). To prove our main result (Theorem 4.2), we shall build on this observation in a way similar to the approach followed in Section 3.

Before proceeding to a study of matrix interpolation problems, it should be remarked that our results will apply equally to matrix regularization problems. That is, a variant of Theorem 2.1 can

be shown for matrix regularization and interpolation problems, following along the lines of the proof of that theorem. The hypothesis now becomes that for some $V, Y \in \mathbf{M}_{n,n}$, V nonsingular, the minimizer of $\min\{\mathcal{E}(AV, Y) : A \in \mathbf{M}_{n,n}\}$ is unique and nonsingular. As a result, matrix regularization with the square loss, the hinge loss or the logistic loss does not differ from matrix interpolation with respect to representer theorems.

Thus, we may focus on the interpolation problems (4.4) and (4.7). First of all, observe that, by definition, problems of the type (4.4) include those of type (4.7). Conversely, consider a set of constraints of the type (4.4) with one input per task ($m_t = 1, \forall t \in \mathbb{N}_n$) and not all input vectors collinear. Then any matrix W such that each w_t lies on a fixed hyperplane perpendicular to x_{t1} satisfies these constraints. At least two of these hyperplanes do not coincide, whereas each constraint in (4.7) implies that all vectors w_t lie on the same hyperplane. Therefore, the class of problems (4.4) is strictly larger than the class (4.7).

However, it turns out that with regard to representer theorems of the form (4.5) there is no distinction between the two types of problems. In other words, the representer theorem holds for the same regularizers Ω , independent of whether each task has its own sample or not. More importantly, we can connect the existence of representer theorems to a geometric property of the regularizer, in a way analogous to property (3.3) in Section 3. These facts are stated in the following lemma.

Lemma 4.1. *The following statements are equivalent:*

- (a): *Problem (4.7) admits a solution of the form (4.5), for every data set $\{(x_i, y_{ti}) : i \in \mathbb{N}_m, t \in \mathbb{N}_n\} \subseteq \mathbf{M}_{d,n} \times \mathbf{M}_{n,n}$ and every $m \in \mathbb{N}$, such that the interpolation constraints are satisfiable.*
- (b): *Problem (4.4) admits a solution of the form (4.5), for every data set $\{(x_{ti}, y_{ti}) : i \in \mathbb{N}_{m_t}, t \in \mathbb{N}_n\} \subseteq \mathbb{R}^d \times \mathbb{R}$ and every $m_t \in \mathbb{N}$, such that the interpolation constraints are satisfiable.*
- (c): *The function Ω satisfies the property*

$$\Omega(W + P) \geq \Omega(W) \quad \forall W, P \in \mathbf{M}_{d,n} \text{ such that } W^\top P = 0. \quad (4.8)$$

Proof. We will show that (a) \implies (c), (c) \implies (b) and (b) \implies (a).

[(a) \implies (c)] Consider any $W \in \mathbf{M}_{d,n}$. Choose $m = n$ and the input data to be the columns of W . In other words, consider the problem

$$\min\{\Omega(Z) : Z \in \mathbf{M}_{d,n}, Z^\top W = W^\top W\}.$$

By hypothesis, there exists a solution $\hat{Z} = WC$ for some $C \in \mathbf{M}_{n,n}$. Since $(\hat{Z} - W)^\top W = 0$, all columns of $\hat{Z} - W$ have to belong to the null space of W . But, at the same time, they have to lie in the range of W and hence we obtain that $\hat{Z} = W$. Therefore, we obtain property (4.8) after the variable change $P = Z - W$.

[(c) \implies (b)] Consider arbitrary $x_{ti} \in \mathbb{R}^d, y_{ti} \in \mathcal{Y}, i \in \mathbb{N}_{m_t}, t \in \mathbb{N}_n$, and let \hat{W} be a solution to problem (4.4). We can decompose the columns of \hat{W} as $\hat{w}_t = \bar{w}_t + p_t$ where $\bar{w}_t \in \mathcal{L} := \text{span}\{x_{si}, i \in \mathbb{N}_{m_s}, s \in \mathbb{N}_n\}$, and $p_t \in \mathcal{L}^\perp, \forall t \in \mathbb{N}_n$. By hypothesis $\Omega(\hat{W}) \geq \Omega(\bar{W})$. Since \hat{W} interpolates the data, so does \bar{W} and therefore \bar{W} is a solution to (4.4).

[(b) \implies (a)] Trivial, since any problem of type (4.7) is also of type (4.4). ■

The above lemma provides us with a criterion for characterizing all regularizers satisfying representer theorems of the form (4.5), in the context of problems (4.4) or (4.7). Our objective will be to obtain a functional form analogous to (3.2) that describes functions satisfying property (4.8). This property does not have a simple geometric interpretation, unlike (3.3) which describes functions with spherical contours. The reason is that the matrix product in the constraint is more difficult to tackle than an inner product.

Similar to the Hilbert space setting (3.2), where we required h to be a nondecreasing real function, the functional description of the regularizer now involves the notion of a *matrix nondecreasing* function.

Definition 4.1. We say that the function $h : \mathbf{S}_+^n \rightarrow \mathbb{R}$ is nondecreasing in the order of matrices if $h(A) \leq h(B)$ for all $A, B \in \mathbf{S}_+^n$ such that $A \preceq B$.

Theorem 4.2. Let $d, n \in \mathbb{N}$ with $d \geq 2n$. The differentiable function $\Omega : \mathbf{M}_{d,n} \rightarrow \mathbb{R}$ satisfies property (4.8) if and only if there exists a matrix nondecreasing function $h : \mathbf{S}_+^n \rightarrow \mathbb{R}$ such that

$$\Omega(W) = h(W^\top W), \quad \forall W \in \mathbf{M}_{d,n}. \quad (4.9)$$

Proof. We first assume that Ω satisfies property (4.8). From this property it follows that, for all $W, P \in \mathbf{M}_{d,n}$ with $W^\top P = 0$,

$$\langle \nabla \Omega(W), P \rangle = 0. \quad (4.10)$$

To see this, observe that if the matrix $W^\top P$ is zero then, for all $\varepsilon > 0$, we have that

$$\frac{\Omega(W + \varepsilon P) - \Omega(W)}{\varepsilon} \geq 0.$$

Taking the limit as $\varepsilon \rightarrow 0^+$ we obtain that $\langle \nabla \Omega(W), P \rangle \geq 0$. Similarly, choosing $\varepsilon < 0$ we obtain that $\langle \nabla \Omega(W), P \rangle \leq 0$ and equation (4.10) follows.

Now, consider any matrix $W \in \mathbf{M}_{d,n}$. Let $r = \text{rank}(W)$ and let us write W in a singular value decomposition as follows

$$W = \sum_{i \in \mathbb{N}_r} \sigma_i u_i v_i^\top,$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ are the singular values and $u_i \in \mathbb{R}^d$, $v_i \in \mathbb{R}^n$, $i \in \mathbb{N}_r$, sets of singular vectors, so that $u_i^\top u_j = v_i^\top v_j = \delta_{ij}$, $\forall i, j \in \mathbb{N}_r$. Also, let $u_{r+1}, \dots, u_d \in \mathbb{R}^d$ be vectors that together with u_1, \dots, u_r form an orthonormal basis of \mathbb{R}^d . Without loss of generality, let us pick u_1 and consider any *unit* vector z orthogonal to the vectors u_2, \dots, u_r . Let $k = d - r + 1$ and $q \in \mathbb{R}^k$ be the unit vector such that

$$z = Rq,$$

where $R = (u_1, u_{r+1}, \dots, u_d)$. We can complete q by adding $d - r$ columns to its right in order to form an orthogonal matrix $Q \in \mathbf{O}^k$ and, since $d > n$, we may select these columns so that $\det(Q) = 1$. Furthermore, we can write this matrix as $Q = e^D$ with $D \in \mathbf{M}_{k,k}$ a skew-symmetric matrix (see [Horn and Johnson, 1991, Example 6.2.15]).

We also define the path $Z : [0, 1] \rightarrow \mathbf{M}_{d,n}$ as

$$Z(\lambda) = \sigma_1 R e^{\lambda D} e_1 v_1^\top + \sum_{i=2}^r \sigma_i u_i v_i^\top \quad \forall \lambda \in [0, 1],$$

where e_1 denotes the vector $(1, 0, \dots, 0)^\top$. In other words, we fix the singular values, the right singular vectors and the $r - 1$ left singular vectors u_2, \dots, u_r and only allow the first left singular vector to vary. This path has the properties that $Z(0) = W$ and $Z(1) = \sigma_1 z v_1^\top + \sum_{i=2}^r \sigma_i u_i v_i^\top$.

By construction of the path, it holds that

$$Z'(\lambda) = \sigma_1 R e^{\lambda D} D e_1 v_1^\top$$

and hence

$$Z(\lambda)^\top Z'(\lambda) = (\sigma_1 R e^{\lambda D} e_1 v_1^\top)^\top \sigma_1 R e^{\lambda D} D e_1 v_1^\top = \sigma_1^2 v_1 e_1^\top D e_1 v_1^\top = 0,$$

for every $\lambda \in [0, 1]$, because $D_{11} = 0$. Hence, using equation (4.10), we have that

$$\langle \nabla \Omega(Z(\lambda)), Z'(\lambda) \rangle = 0$$

and, since $\frac{d\Omega(Z(\lambda))}{d\lambda} = \langle \nabla \Omega(Z(\lambda)), Z'(\lambda) \rangle$, we conclude that $\Omega(Z(\lambda))$ equals a constant independent of λ . In particular, $\Omega(Z(0)) = \Omega(Z(1))$, that is,

$$\Omega(W) = \Omega \left(\sigma_1 z v_1^\top + \sum_{i=2}^r \sigma_i u_i v_i^\top \right).$$

In other words, if we fix the singular values of W , the right singular vectors and all the left singular vectors but one, Ω does not depend on the remaining left singular vector (because the choice of z is independent of u_1).

In fact, this readily implies that Ω does not depend on the left singular vectors at all. Indeed, fix an arbitrary $Y \in \mathbf{M}_{d,n}$ such that $Y^\top Y = I$. Consider the matrix $Y(W^\top W)^{\frac{1}{2}}$, which can be written using the same singular values and right singular vectors as W . That is,

$$Y(W^\top W)^{\frac{1}{2}} = \sum_{i \in \mathbb{N}_r} \sigma_i \tau_i v_i^\top,$$

where $\tau_i = Y v_i$, $\forall i \in \mathbb{N}_r$. Now, we select unit vectors z_1, \dots, z_r as follows:

$$\begin{aligned} z_1 &= u_1 \\ z_2 &\perp z_1, u_3, \dots, u_r, \tau_1 \\ &\vdots \\ z_r &\perp z_1, \dots, z_{r-1}, \tau_1, \dots, \tau_{r-1}. \end{aligned}$$

This construction is possible since $d \geq 2n$. Replacing successively u_i with z_i and then z_i with τ_i , $\forall i \in \mathbb{N}_r$, and applying the invariance property, we obtain that

$$\begin{aligned}
\Omega(W) &= \Omega\left(\sum_{i \in \mathbb{N}_r} \sigma_i u_i v_i^\top\right) \\
&= \Omega\left(\sigma_1 z_1 v_1^\top + \sigma_2 z_2 v_2^\top + \sum_{i=3}^r \sigma_i u_i v_i^\top\right) \\
&\quad \vdots \quad \vdots \\
&= \Omega\left(\sum_{i \in \mathbb{N}_r} \sigma_i z_i v_i^\top\right) \\
&= \Omega\left(\sigma_1 \tau_1 v_1^\top + \sum_{i=2}^r \sigma_i z_i v_i^\top\right) \\
&\quad \vdots \quad \vdots \\
&= \Omega\left(\sum_{i \in \mathbb{N}_r} \sigma_i \tau_i v_i^\top\right) = \Omega\left(Y(W^\top W)^{\frac{1}{2}}\right).
\end{aligned}$$

Therefore, defining the function $h : \mathbf{S}_+^n \rightarrow \mathbb{R}$ as $h(A) = \Omega(YA^{\frac{1}{2}})$, we deduce that $\Omega(W) = h(W^\top W)$.

Finally, we show that h is matrix nondecreasing, that is, $h(A) \leq h(B)$ if $0 \preceq A \preceq B$. For any such A, B and since $d \geq 2n$, we may define $W = [A^{\frac{1}{2}}, 0, 0]^\top$, $P = [0, (B - A)^{\frac{1}{2}}, 0]^\top \in \mathbf{M}_{d,n}$. Then $W^\top P = 0$, $A = W^\top W$, $B = (W + P)^\top (W + P)$ and thus, by hypothesis,

$$h(B) = \Omega(W + P) \geq \Omega(W) = h(A).$$

This completes the proof in one direction of the theorem.

To show the converse, assume that $\Omega(W) = h(W^\top W)$, where the function h is matrix nondecreasing. Then for any $W, P \in \mathbf{M}_{d,n}$ with $W^\top P = 0$, we have that $(W + P)^\top (W + P) = W^\top W + P^\top P \succeq W^\top W$ and, so, $\Omega(W + P) \geq \Omega(W)$, as required. \blacksquare

We conclude this section by providing a necessary and sufficient condition on the matrix nondecreasing property of the function h .

Proposition 4.1. *Let $h : \mathbf{S}_+^n \rightarrow \mathbb{R}$ be differentiable function. The following properties are equivalent:*

- (a) h is matrix nondecreasing
- (b) the matrix $\nabla h(A) := \left(\frac{\partial h}{\partial a_{ij}} : i, j \in \mathbb{N}_n\right)$ is positive semidefinite, for every $A \in \mathbf{S}_+^n$.

Proof. If (a) holds, we choose $x \in \mathbb{R}^n$, $t \in \mathbb{R}$ and note that

$$\frac{h(A + txx^\top) - h(A)}{t} \geq 0.$$

Letting t go to zero gives that $x^\top \nabla h(A)x \geq 0$.

Conversely, if (b) is true we have, for every $x \in \mathbb{R}^n$, that $x^\top \nabla h(A)x = \langle \nabla h(A), xx^\top \rangle \geq 0$ and, so, $\langle \nabla h(A), C \rangle \geq 0$ for all $C \in \mathbf{S}_+^n$. For any $A, B \in \mathbf{S}_+^n$ such that $A \preceq B$, consider the univariate function $g : [0, 1] \rightarrow \mathbb{R}$, $g(t) = h(A + t(B - A))$. By the chain rule it is easy to verify that g is nondecreasing. Therefore we conclude that $h(A) = g(0) \leq g(1) = h(B)$. ■

4.3 Examples

We have briefly mentioned already that functional description (4.9) subsumes the special case of *monotone spectral functions*. By spectral functions we simply mean those real-valued functions of matrices that depend only on the singular values of their argument. Monotonicity in this case simply means that one-by-one orderings of the singular values are preserved. In addition, the monotonicity of h in (4.9) is a direct consequence of Weyl's monotonicity theorem [Horn and Johnson, 1985, Cor. 4.3.3], which states that if $A \preceq B$ then the spectra of A and B are ordered.

Interesting examples of such functions are the *Schatten L_p norms* and *pre norms*,

$$\Omega(W) = \|W\|_p := \|\sigma(W)\|_p,$$

where $p \in [0, +\infty)$ and $\sigma(W)$ denotes the n -dimensional vector of the singular values of W . For instance, we have already mentioned in Section 4.1 that the representer theorem holds when the regularizer is the trace norm (the L_1 norm of the spectrum). But it also holds for the *rank* of a matrix, which is the L_0 pre norm of the spectrum. Regularization with the rank is an NP-hard optimization problem but the representer theorem implies that it can be solved in time dependent on the total sample size.

If we exclude spectral functions, the functions that remain are invariant under *left* multiplication with an orthogonal matrix. Examples of such functions are Schatten norms and pre norms composed with *right* matrix scaling,

$$\Omega(W) = \|WM\|_p, \tag{4.18}$$

where $M \in \mathbf{S}^n$. In this case, the corresponding h is the function $S \mapsto \|\sqrt{\sigma(MSM)}\|_p$. To see that this function is matrix nondecreasing, observe that if $A, B \in \mathbf{S}_+^n$ and $A \preceq B$ then $0 \preceq MAM \preceq MBM$ and hence $\sigma(MAM) \preceq \sigma(MBM)$ by Weyl's monotonicity theorem. Therefore, $\|\sqrt{\sigma(MAM)}\|_p \leq \|\sqrt{\sigma(MBM)}\|_p$.

Also, the matrix M above can be used to select a subset of the columns of W . In addition, more complicated structures can be obtained by summation of matrix nondecreasing functions and by taking minima or maxima over sets. For example, we can obtain a regularizer such as

$$\Omega(W) = \min_{\{I_1, \dots, I_K\} \in \mathcal{P}} \sum_{k \in \mathbb{N}_K} \|W(I_k)\|_1,$$

where \mathcal{P} is the set of partitions of \mathbb{N}_n in K subsets and $W(I_k)$ denotes the submatrix of W formed by just the columns indexed by I_k . This regularizer is an extension of the trace norm and can be used for learning multiple tasks via dimensionality reduction on more than one subspaces [Argyriou et al., 2008].

Yet another example of valid regularizer is that considered in [Evgeniou et al., 2005, Sec. 3.1], which encourages the tasks to be close to each others, namely

$$\Omega(W) = \sum_{t=1}^n \left\| w_t - \frac{1}{n} \sum_{s=1}^n w_s \right\|^2.$$

This regularizer immediately verifies property (4.8), and so by Theorem 4.2 it is a matrix non-decreasing function of $W^\top W$. One can also verify that this regularizer is the square of the form (4.18) with $p = 2$.

Finally, it is worth noting that the representer theorem does *not* apply to a family of “mixed” matrix norms that have been used in both statistics and machine learning, in formulations such as the “group Lasso” [Antoniadis and Fan, 2001, Argyriou et al., 2006, Bakin, 1999, Grandvalet and Canu, 1999, Lin and Zhang, 2003, Obozinski et al., 2006, Yuan and Lin, 2006]. These norms are of the form

$$\Omega(W) = \|W\|_{p,q} := \left(\sum_{i \in \mathbb{N}_d} \|w^i\|_p^q \right)^{\frac{1}{q}},$$

where w^i denotes the i -th row of W and $(p, q) \neq (2, 2)$. Typically in the literature, q is chosen equal to one in order to favor sparsity of the coefficient vectors *at the same covariates*.

5 Conclusion

We have characterized the classes of vector and matrix regularizers which lead to certain forms of the solution of the associated regularization problems. In the vector case, we have proved the necessity of a well-known sufficient condition for the “standard representer theorem”, which is encountered in many learning and statistical estimation problems. In the matrix case, we have described a novel class of regularizers which lead to a modified representer theorem. This class, which relies upon the notion of matrix nondecreasing function, includes and extends significantly the vector class. To motivate the need for our study, we have discussed some examples of regularizers, which have been recently used in the context of multi-task learning and collaborative filtering.

In the future, it would be valuable to study more in detail special cases of the matrix regularizers which we have encountered, such as those based on orthogonally invariant functions. It would also be interesting to investigate how the presence of additional constraints affects the representer theorem. In particular, we have in mind the possibility that the matrix may be constrained to be in a convex cone, such as the set of positive semidefinite matrices. Finally, we leave to future studies the extension of the ideas presented here to the case in which matrices are replaced by operators between two Hilbert spaces.

Acknowledgments

The work of the first and third authors was supported by EPSRC Grant EP/D052807/1 and by the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778. The second author is supported by NSF grant DMS 0712827.

Appendix

Here we collect some auxiliary results which are used in the above analysis.

The first result states a basic property of connectedness through rotations.

Lemma 5.1. *Let $w, v \in \mathbb{R}^d$ and $d \geq 2$. Then there exists $U \in \mathbf{O}^d$ with determinant 1 such that $v = Uw$ if and only if $\|w\| = \|v\|$.*

Proof. If $v = Uw$ we have that $v^\top v = w^\top w$. Conversely, if $\|w\| = \|v\|$, we may choose orthonormal vectors $\{x_\ell : \ell \in \mathbb{N}_{d-1}\} \perp w$ and $\{z_\ell : \ell \in \mathbb{N}_{d-1}\} \perp v$ and form the matrices $R = (w, x_1, \dots, x_{d-1})$ and $S = (v, z_1, \dots, z_{d-1})$. We have that $R^\top R = S^\top S$. We wish to solve the equation $UR = S$. For this purpose we choose $U = SR^{-1}$ and note that $U \in \mathbf{O}^d$ because $U^\top U = (R^{-1})^\top S^\top S R^{-1} = (R^{-1})^\top R^\top R R^{-1} = I$. Since $d \geq 2$, in the case that $\det(U) = -1$ we can simply change the sign of one of the x_ℓ or z_ℓ to get $\det(U) = 1$ as required. ■

The second result concerns the monotonicity of the trace norm.

Lemma 5.2. *Let $W, P \in \mathbf{M}_{d,n}$ such that $W^\top P = 0$. Then $\|W + P\|_1 \geq \|W\|_1$.*

Proof. It is known that the square root function, $t \mapsto t^{\frac{1}{2}}$, is *matrix monotone* – see, for example, [Bhatia, 1997, Sec. V.1]. This means that for any matrices $A, B \in \mathbf{S}_+^n$, $A \succeq B$ implies $A^{\frac{1}{2}} \succeq B^{\frac{1}{2}}$. Hence, for any matrices $A, B \in \mathbf{S}_+^n$, $A \succeq B$ implies $\text{tr} A^{\frac{1}{2}} \geq \text{tr} B^{\frac{1}{2}}$. We apply this fact to the matrices $W^\top W + P^\top P$ and $W^\top W$ to obtain that

$$\|W + P\|_1 = \text{tr}((W + P)^\top(W + P))^{\frac{1}{2}} = \text{tr}(W^\top W + P^\top P)^{\frac{1}{2}} \geq \text{tr}(W^\top W)^{\frac{1}{2}} = \|W\|_1. \quad \blacksquare$$

References

- D.A. Aaker, V. Kumar, and G.S. Day. *Marketing Research*. John Wiley & Sons, 2004. 8th edition.
- J. Abernethy, F. Bach, T. Evgeniou, and J-P. Vert. Low-rank matrix factorization with attributes. Technical Report 2006/68/TOM/DS, INSEAD, 2006. Working paper.
- Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the Twenty-Fourth International Conference on Machine learning*, 2007.
- A. Antoniadis and J. Fan. Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association*, 96:939–967, 2001.
- A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, 2006.
- A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 2007a. <http://www.springerlink.com/content/161105027v344n03>.

- A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, 2007b.
- A. Argyriou, A. Maurer, and M. Pontil. An algorithm for transfer learning in a heterogeneous environment. In *European Conference on Machine Learning*, 2008. To appear.
- N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 686:337–404, 1950.
- S. Bakin. *Adaptive regression and model selection in data mining problems*. PhD thesis, Australian National University, Canberra, 1999.
- R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer, 1997.
- B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA, 1992. ACM.
- E. J. Candès and B. Recht. Exact matrix completion via convex optimization. Submitted for publication, 2008.
- G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.
- F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.
- E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.
- F. Dinuzzo, M. Neve, G. De Nicolao, and U. P. Gianazza. On the representer theorem and equivalent degrees of freedom of SVR. *Journal of Machine Learning Research*, 8:2467–2495, 2007.
- T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.
- T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.
- F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.
- Yves Grandvalet and Stéphane Canu. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 445–451, Cambridge, MA, USA, 1999. MIT Press.
- A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42:80–86, 1970.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

- R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.
- A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264, 1975.
- G.S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.
- P. J. Lenk, W. S. DeSarbo, P. E. Green, and M. R. Young. Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.
- Y. Lin and H. H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models – COSSO. Technical Report 2556, Institute of Statistics Mimeo Series, NCSU, 2003.
- A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7: 117–139, 2006a.
- A. Maurer. The Rademacher complexity of linear transformation classes. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, volume 4005 of *LNAI*, pages 65–78. Springer, 2006b.
- C. A. Micchelli and A. Pinkus. Variational problems arising from balancing several error criteria. *Rendiconti di Matematica, Serie VII*, 14:37–86, 1994.
- C. A. Micchelli and M. Pontil. A function representation for learning in Banach spaces. In *Proceedings of the Nineteenth Annual Conference on Learning Theory*, volume 3120 of *Lecture Notes in Computer Science*, pages 255–269. Springer, 2004.
- C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005a.
- C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17: 177–204, 2005b.
- C. A. Micchelli and T. J. Rivlin. Lectures on optimal recovery. In P. R. Turner, editor, *Lecture Notes in Mathematics*, volume 1129. Springer Verlag, 1985.
- G. Obozinski, B. Taskar, and M.I. Jordan. Multi-task feature selection. Technical report, Dept. of Statistics, UC Berkeley, June 2006.
- B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, 2002.
- B. Schölkopf, R. Herbrich, and A.J. Smola. A generalized representer theorem. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, 2001.
- J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

- N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pages 1329–1336. MIT Press, 2005.
- I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071–1105, 2003.
- A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill posed problems*. V. H. Winston and Sons (distributed by Wiley), 1977. F. John, Translation Editor.
- V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.
- G. Wahba. *Spline Models for Observational Data*, volume 59 of *Series in Applied Mathematics*. SIAM, Philadelphia, 1990.
- L. Wolf, H. Jhuang, and T. Hazan. Modeling appearances with low-rank SVM. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2007.
- Z. Xiang and K. P. Bennett. Inductive transfer using kernel multitask latent analysis. In *Inductive Transfer : 10 Years Later, NIPS 2005 Workshop*, 2005.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 68(1):49–67, 2006.
- M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346, 2007.