Journal of Machine Learning Research 11 (2010) 935-953 Submitted 4/09; Published 2/10

On Spectral Learning

Andreas Argyriou A.ARGYRIOU@CS.UCL.AC.UK
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT, UK

Charles A. Micchelli
Department of Mathematics and Statistics
State University of New York
The University at Albany
Albany, New York 12222, USA

Massimiliano Pontil M.PONTIL@CS.UCL.AC.UK
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT, UK

Editor: Alexander Smola

Abstract

In this paper, we study the problem of learning a matrix W from a set of linear measurements. Our formulation consists in solving an optimization problem which involves regularization with a spectral penalty term. That is, the penalty term is a function of the spectrum of the covariance of W. Instances of this problem in machine learning include multi-task learning, collaborative filtering and multi-view learning, among others. Our goal is to elucidate the form of the optimal solution of spectral learning. The theory of spectral learning relies on the von Neumann characterization of orthogonally invariant norms and their association with symmetric gauge functions. Using this tool we formulate a representer theorem for spectral regularization and specify it to several useful example, such as Schatten $p-$norms, trace norm and spectral norm, which should proved useful in applications.

Keywords: kernel methods, matrix learning, minimal norm interpolation, multi-task learning, orthogonally invariant norms, regularization

1. Introduction

In this paper, we study the problem of learning a matrix from a set of linear measurements. Our formulation consists in solving for the matrix

$$\hat{W} = \operatorname{argmin}\{E(I(W),y) + \gamma\Omega(W) : W \in M_{d,n}\}, \tag{1}$$

where $M_{d,n}$ is the set of $d \times n$ real matrices, y an $m-$dimensional real vector of observations and $I : M_{d,n} \to \mathbb{R}^m$ a linear operator, whose components are given by the Frobenius inner product between the matrix W and prescribed data matrices. The objective function in (1) combines a data term, $E(I(W),y)$, which measures the fit of W to available training data and a penalty term or regularizer, $\Omega(W)$. The positive constant γ controls the trade-off between the two terms and may be chosen

by prior information on the noise underlying the data. A typical example for the data term is $E(I(W),y) = \|I(W)-y\|_2^2$, where the subscript indicates the Euclidean norm.

In the design of learning algorithms from the point of view of regularization the choice of the penalty term is essential. To obtain insights into this issue, we shall investigate in this paper the form of matrices which solve the variational problem (1) when the penalty term is an *orthogonally invariant norm* (OI-norm). This means, for any pair of orthogonal matrices U and V, that

$$\Omega(UWV) = \Omega(W).$$

There are many important examples of OI-norms. Among them, the family of Schatten $p-$norms, $1 \leq p \leq \infty$, namely the ℓ_p-norm of the singular values of a matrix, are especially useful.

Our main motivation for studying the optimization problem (1) arises from its application to multi-task learning, see Argyriou et al. (2007a) and references therein. In this context, the matrix columns are interpreted as the parameters of different regression or classification tasks and the regularizer Ω is chosen in order to favor certain kinds of dependencies across the tasks. The operator I consists of inner products formed from the inputs of each task and the error term $E(I(W),y)$ is the sum of losses on the individual tasks. Collaborative filtering (Srebro et al., 2005) provides another interesting instance of problem (1), in which the operator I is formed from a subset of the matrix elements. Further examples in which OI-norm have been used include multi-class classification (Amit et al., 2007), multi-view learning (Cavallanti et al., 2008) and similarity learning (Maurer, 2008a).

A recent trend in regularization methods in machine learning is to use matrix regularizers which are orthogonally invariant (Argyriou et al., 2007a,b; Abernethy et al., 2009; Srebro et al., 2005). In particular, an important case is the Schatten one norm of W, which is often referred to as the *trace norm*. The general idea behind this methodology is that a small trace norm favors low-rank solution matrices to (1). This means that the tasks (the columns of W) are related in that they all lie in a low-dimensional subspace of $\mathbb{R}^d$. Indeed, if we choose the regularizer to be the rank of a matrix, we obtained a non-convex NP-hard problem. However, the trace norm provides a convex relaxation of this problem, which has been justified in various ways (see, for example, Fazel et al., 2001; Candès and Recht, 2008).

The main purpose of this paper is to characterize the form of the solutions to problem (1). Specifically, we provide what in machine learning is known as a *representer theorem*. Namely we show, for a wide variety of OI-norm regularizers, that it is possible to compute the inner product $\langle \hat{W}, X\rangle$ only in terms of the $m \times m$ Gram matrix I^*I and $I(X)$. A representer theorem is appealing from a practical point of view, because it ensures that the cost of solving the optimization problem (1) depends on the size m of the training sample, which can be much smaller than the number of elements of the matrix W. For example, in multi-task learning, the number of rows in the matrix W may be much larger than the number of data per task. More fundamentally, the task vectors (the columns of matrix W) may be elements of a reproducing kernel Hilbert Space.

Our point of view in developing these theorems is through the study of the *minimal norm interpolation* problem

$$\min\{\|W\| : I(W) = \hat{y},\ W \in M_{d,n}\}.$$

The reason for this is that the solution $\hat{W}$ of problem (1) also solves the above problem for an appropriately chosen $\hat{y} \in \mathbb{R}^m$. Specifically, this is the case if we choose $\hat{y} = I(\hat{W})$. In the development of these results, tools from convex analysis are needed. In particular, a key tool that we use in this

paper is a classical result of von Neumann (1962), which characterizes OI-norms in terms of the notion of *symmetric gauge function*; see also Lewis (1995) for a discussion of the von Neumann theorem in the context of convex analysis. We record some of these facts which we need in Section 4.

The paper is organized in the following manner. In Section 2 we introduce our notation and describe the connection between minimal norm interpolation and regularization. In Section 3 we describe the relationship between any solution of (1) and any solution of a dual problem, which involves a number of variables equal to the training set size. In Section 4 we specify this result to the class of OI-norms. In particular, we describe a special case of such norms, which contains the Schatten $p-$norms, and derive a linear representer theorem for this case. As we shall see, this computation in general involves a nonlinear function and a singular value decomposition of an appropriate matrix.

2. Background

Before proceeding, we introduce some of the notation used in the paper and review some basic facts.

2.1 Notation

We use $\mathbb{N}_d$ as a shorthand for the set of integers $\{1,\ldots,d\}$, $\mathbb{R}^d$ for the linear space of vectors with d real components and $M_{d,n}$ for the linear space of $d\times n$ real matrices. For any vector $a\in\mathbb{R}^d$ we use a_i to denote its i-th component and for any matrix $W\in M_{d,n}$ we use w_t to denote the t-th column of W, for $t\in\mathbb{N}_n$. For a vector $\lambda\in\mathbb{R}^d$, we let $\mathrm{Diag}(\lambda)$ or $\mathrm{Diag}(\lambda_i)_{i\in\mathbb{N}_d}$ to denote the $d\times d$ diagonal matrix having the elements of λ on the diagonal. We denote the trace of matrix W by $\mathrm{tr}(W)$. We use $\mathbf{S}^d$ to denote the set of $d\times d$ real symmetric matrices and $\mathbf{S}^d_+$ and $\mathbf{S}^d_{++}$ to denote the subsets of positive semidefinite and positive definite ones, respectively. We use $\succ$ and $\succeq$ for the positive definite and positive semidefinite partial orderings on $\mathbf{S}^d$, respectively. We also let O_d be the set of $d\times d$ orthogonal matrices and $\mathcal{P}_d$ the set of $d\times d$ permutation matrices. Finally, in this paper, the notation $\langle\cdot,\cdot\rangle$ denotes the standard inner products on $\mathbb{R}^d$ and $M_{d,n}$, that is, $\langle a,b\rangle=\sum_{i\in\mathbb{N}_d}a_ib_i$ for any vectors $a,b\in\mathbb{R}^d$ and $\langle W,V\rangle=\mathrm{tr}(W^\top V)$ for any matrices $W,V\in M_{d,n}$.

2.2 Regularization and Interpolation with Matrices

Let us first describe the type of optimization problems of interest in this paper. Our motivation comes from recent work in machine learning which deals with the problem of multi-task learning. Beyond these practical concerns, the matrix optimization problems we consider here have the property that the *matrix structure* is important.

We shall consider *regularization* problems of the type

$$\min\left\{E(I(W),y)+\gamma\Omega(W):W\in M_{d,n}\right\}, \tag{2}$$

where $E:\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}$ is called a *loss function*, $\Omega:M_{d,n}\to\mathbb{R}$ a *regularizer*, $\gamma>0$ the *regularization parameter*, $I:M_{d,n}\to\mathbb{R}^m$ is a *linear operator* and $y\in\mathbb{R}^m$. Associated to the above regularization problem is the *interpolation problem*

$$\min\left\{\Omega(W):W\in M_{d,n},\ I(W)=y\right\}. \tag{3}$$

Unless otherwise stated, we always assume that the minima in problems (2) and (3) are attained.

Regularization enables one to trade off interpolation of the data against smoothness or simplicity of the model, whereas interpolation frequently suffers from *overfitting*. Note that the family of the former problems encompasses the latter ones. Indeed, an interpolation problem can be simply obtained in the limit as the regularization parameter γ goes to zero (see, for example, Micchelli and Pinkus, 1994).

For example, a special case of matrix regularization problems of the type (2) is obtained with the choice

$$I(W) = (\langle w_t, x_{ti} \rangle : t \in \mathbb{N}_n, i \in \mathbb{N}_{m_t}) \,,$$

where the x_{ti} are given input vectors in $\mathbb{R}^d$. This occurs, for example, in multi-task learning and problems closely related to it (Abernethy et al., 2009; Argyriou et al., 2007a,b; Candès and Recht, 2008; Cavallanti et al., 2008; Izenman, 1975; Maurer, 2006a,a; 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 which 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.

In multi-task learning, the error term E in (2) 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. For example, one such regularizer, considered in Evgeniou et al. (2005), is a specific quadratic form in W, namely

$$\Omega(W) = \sum_{s,t \in \mathbb{N}_n} \langle w_s, E_{st} w_t \rangle,$$

where the matrices $E_{st} \in \mathbf{S}^d$ are chosen to model cross-tasks interactions.

Another common choice for the regularizer is the *trace norm*, which is defined to be the sum of the singular values of a matrix,

$$\Omega(W) = \sum_{j \in \mathbb{N}_r} \sigma_j(W),$$

where $r = \min(d, m)$. Equivalently this regularizer can be expressed as $\Omega(W) = \mathrm{tr}(W^\top W)^{\frac{1}{2}}$. Regularization with the trace norm learns the tasks as one joint optimization problem, by favoring matrices with low rank (Argyriou et al., 2007a). 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 a *few* data points available for each task.

In general, the linear operator I can be written in the form

$$I(W) = (\langle W, X_i \rangle : i \in \mathbb{N}_m) \,, \tag{4}$$

where the inputs matrices X_i are in $M_{d,n}$. Recall that the *adjoint* operator, $I^* : \mathbb{R}^m \to M_{d,n}$, is defined by the property that

$$\langle I^*(c), W \rangle = \langle c, I(W) \rangle \,,$$

for all $c \in \mathbb{R}^m, W \in M_{d,n}$. Therefore, it follows that I^* is given at $c \in \mathbb{R}^m$, by

$$I^*(c) = \sum_{i \in \mathbb{N}_m} c_i X_i.$$

We denote by $\mathcal{R}(I)$ and $\mathcal{N}(I)$ the range and the null space of the operator I, respectively.

In this paper, we are interested in studying the form of the solution to matrix problems (2) or (3). For certain families of regularizers, the solutions can be expressed in terms of the given inputs X_i in (4). Such facts are known in machine learning as *representer theorems*, see Argyriou et al. (2009) and reference therein.

The line of attack we shall follow in this paper will go through *interpolation*. That is, our main concern will be to obtain representer theorems which hold for problems like (3). This in turn will imply representer theorems for the associated regularization problems. This is justified by the next lemma.

Lemma 1 *Let $E : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, a linear operator $I : M_{d,n} \to \mathbb{R}^m$, $\Omega : M_{d,n} \to \mathbb{R}$, $\gamma > 0$ such that the problems* (2) *and* (3) *admit a minimizer for every $y \in \mathbb{R}^m$. Then for every $y \in \mathbb{R}^m$ there exists $\hat{y} \in \mathbb{R}^m$ such that any solution of the interpolation problem* (3) *with $y = \hat{y}$ is a solution of the regularization problem* (2)*.*

Proof If $\hat{W}$ solves (2), we may define $\hat{y} := I(\hat{W})$. It then readily follows that any solution of (3) with $\hat{y}$ in place of y is a solution of (2). ■

For some other results relating optimality conditions for regularization and interpolation problems, see Argyriou et al. (2009). We shall return to this issue in Section 4, where we study representer theorems of a particular type for regularizers which are OI-norms.

3. Duality and Minimal Norm Interpolation

In this section, we turn our attention to the study of the interpolation problem (3) when the function Ω is a *norm* on $M_{d,n}$. That is we prescribe a linear operator $I : M_{d,n} \to \mathbb{R}^m$, a vector $y \in \mathcal{R}(I) \setminus \{0\}$ and study the minimal norm interpolation problem

$$\phi := \min\{\|W\| : I(W) = y, W \in M_{d,n}\}. \tag{5}$$

The approach we take to analyze problem (5) makes use of a dual problem. To identify it, we recall the definition of the dual norm, given by

$$\|X\|_D = \max\{\langle X, W\rangle : W \in M_{d,n}, \|W\| \leq 1\}.$$

Consequently, it follows, for every $X, W \in M_{d,n}$, that

$$|\langle X, W\rangle| \leq \|X\|_D \|W\|. \tag{6}$$

Associated with this inequality is the notion of the *peak set* of the norm $\|\cdot\|$ at X, namely

$$\mathscr{D}\|X\| = \{W : \langle X, W\rangle = \|X\|_D, : W \in M_{d,n}, \|W\| = 1\}.$$

Note that, for each $X \in M_{d,n}\backslash\{0\}$ the peak set $\mathscr{D}\|X\|$ is a nonempty compact convex set which contains all $W \in M_{d,n}\backslash\{0\}$ that make the bound in (6) tight.

As we shall see in the theorem below, the dual norm leads to the following *dual problem*

$$\theta := \min\left\{\|I^*(c)\|_D : c \in \mathcal{R}(I),\ \langle c, y\rangle = 1\right\}. \tag{7}$$

Let us first observe that both the primal and dual problem have solutions. In the primal problem we minimize a norm which is a function which grows at infinity and, so, the existence of a solution is

assured. Similarly, the quantity $\|I^*(c)\|_D$ which is minimized in the dual problem is also norm on $c \in \mathcal{R}(I)$.

The main result of this section establishes the relationship between the solutions of the primal problem (5) and those of the dual problem (7).

Theorem 2 *A vector $\hat{c} \in \mathbb{R}^m$ solves the dual problem* (7) *if and only if there exists $\hat{W} \in \theta^{-1}\mathscr{D}\|I^*(\hat{c})\|$ such that $I(\hat{W}) = y$. Moreover, in this case $\hat{W}$ solves the primal problem* (5) *and $\phi\theta = 1$. Conversely, for every $\hat{W}$ solving the primal problem* (5) *and* any *solution $\hat{c}$ of the dual problem* (7)*, it holds that $\hat{W} \in \theta^{-1}\mathscr{D}\|I^*(\hat{c})\|$.*

Before we proceed with a proof, let us explain the rationale behind this result. The number of free parameters in the dual problem is at most $m-1$, while the primal problem involves $dn-m$ parameters. Typically, in applications, dn is much larger than m. Recalling the connection to multi-task learning in Section 2, this means that d is much larger than the number of data per task, $\frac{m}{n}$. Therefore, from the perspective of this parameter count, solving the dual problem may be advantageous. More importantly, any solution of the dual problem will provide us with a solution of the primal problem and conditions on the latter are obtained from a study of the peak set $\mathscr{D}\|I^*(\hat{c})\|$. For example, as we shall see in Section 4, in the case of OI-norms, this fact will be facilitated by fundamental matrix inequalities.

Proof of Theorem 2 First let us establish that

$$\frac{1}{\theta} \leq \phi. \tag{8}$$

To this end, consider any $c \in \mathbb{R}^m$ with $\langle c, y\rangle = 1$ and $W \in M_{d,n}$ with $I(W) = y$. Then

$$1 = \langle c, y\rangle = \langle c, I(W)\rangle = \langle I^*(c), W\rangle \leq \|I^*(c)\|_D \|W\|. \tag{9}$$

From this inequality we get the desired claim. To prove the reverse inequality in (8), we let $\hat{c} \in \mathcal{R}(I)$ be a solution of the dual problem (7) and conclude, for any $b \in \mathcal{R}(I)$ such that $\langle b, y\rangle = 0$, that

$$\lim_{\varepsilon \to 0^+} \frac{\|I^*(\hat{c} + \varepsilon b)\|_D - \|I^*(\hat{c})\|_D}{\varepsilon} \geq 0.$$

Since the dual norm is a maximum of linear functions over a compact set, we may apply Theorem 22 in the case that $\mathcal{X} = \{X : X \in M_{d,n}, \|X\| \leq 1\}$, $\mathcal{W} = M_{d,n}$, $f(W,X) = \langle W, X\rangle$, and evaluate Equation (16) for $W = I^*(\hat{c})$ and $\Delta = I^*(b)$ to obtain the inequality

$$\max\{\langle I^*(b), T\rangle \ : \ T \in \mathscr{D}\|I^*(\hat{c})\|\} \geq 0\,.$$

Using the fact that $b \in \mathcal{R}(I)$ and $\langle b, y\rangle = 0$, we can rephrase this inequality in the following fashion. For every $Z \in M_{d,n}$ such that $\langle Z, I^*(y)\rangle = 0$ we have that

$$\max\{\langle Z, I^*I(T)\rangle \ : \ T \in \mathscr{D}\|I^*(\hat{c})\|\} \geq 0.$$

To resolve this set of inequalities we use Lemma 21 in the appendix with $k = 1$, $J : M_{d,n} \to \mathbb{R}$ defined at $W \in M_{d,n}$ as $J(W) = \langle I^*(y), W\rangle$ and $\mathcal{W} := I^*I(\mathscr{D}\|I^*(\hat{c})\|)$. Since $J^* : \mathbb{R} \to M_{d,n}$ is given for $a \in \mathbb{R}$ as $J^*(a) = aI^*(y)$, we conclude that there exist $\lambda \in \mathbb{R}$ and $\tilde{W} \in \mathscr{D}\|I^*(\hat{c})\|$ such that

$$\lambda I^*(y) = I^*I(\tilde{W}).$$

This equation implies that $\lambda y - I(\tilde{W}) \in \mathcal{N}(I^*)$. However, recalling the fact that $y \in \mathcal{R}(I)$, we also have that $\lambda y - I(\tilde{W}) \in \mathcal{R}(I)$. Therefore, we have established that

$$\lambda y = I(\tilde{W}).$$

To identify the value of λ we use the fact that $\langle y, \hat{c} \rangle = 1$ and obtain that

$$\lambda = \langle \hat{c}, I(\tilde{W}) \rangle = \langle I^*(\hat{c}), \tilde{W} \rangle = \|I^*(\hat{c})\|_D = \theta.$$

Now, we define $\hat{W} = \frac{1}{\theta}\tilde{W}$ and note that $I(\hat{W}) = y$ and since $\|\tilde{W}\| = 1$ we obtain that

$$\phi \leq \|\hat{W}\| = \frac{1}{\theta}.$$

This inequality, combined with inequality (8) demonstrates that $\phi\theta = 1$ and that $\hat{W}$ is a solution to the primal problem (5).

To complete the proof, consider any $\hat{W}$ solving (5) and $\hat{c}$ solving (7) and it easily follows from inequality (9) and $\phi\theta = 1$ that $\hat{W} \in \theta^{-1}\mathscr{D}\|I^*(\hat{c})\|$. ■

Theorem 2 describes the relation between the set of solutions of the primal problem (5) and the dual problem (7). It also relates the set of solutions of the primal problem to the range of the adjoint operator I^*. This latter property, combined with Lemma 1, may be viewed as a general *representer theorem*, that is, the theorem implies that the solutions of the regularization problem (2) are matrices in the set $\mathscr{D}\|I^*(\tilde{c})\|$, for some $\tilde{c} \in \mathbb{R}^m$. However, additional effort is required to obtain a concrete representation of such solution. For example, for the Frobenius norm, $\mathscr{D}\|X\| = \{X/\|X\|\}$ and, so, the optimality condition becomes $\hat{W} = I^*(\tilde{c})$. We refer to this condition throughout the paper as the *standard representer theorem*, see Argyriou et al. (2009) and references therein. In other words, the standard representer theorem for $\hat{W}$ means that $\hat{W} \in \mathcal{R}(I^*)$.

We make no claim of originality for Theorem 2 as its proof uses well established tools of convex analysis. On the contrary, we emphasize the utility of this result for machine learning. Alternatively, we can approach the minimal norm interpolation problem by use of the Lagrangian, defined, for $W \in M_{d,n}$ and $\lambda \in \mathbb{R}^m$, as

$$\mathcal{L}(W, \lambda) = \|W\| + \langle W, I^*(\lambda) \rangle - \langle y, \lambda \rangle.$$

4. Representer Theorems for Orthogonally Invariant Norms

In this section, we focus our attention on matrix norms which are invariant under left and right multiplication by orthogonal matrices. As we shall see, for such norms, the representer theorem can be written in terms of the singular value decomposition. In addition, in Section 4.3, we shall describe a subclass of OI-norms for which representer theorems can be phrased in terms of matrix multiples of the adjoint operator value $I^*(\hat{c})$. This type of representer theorem arises in *multi-task learning* as described in Argyriou et al. (2009). That is, each of the columns of the optimal matrix lies in the span of the corresponding columns of the input matrices X_i.

4.1 Notation

Let $W \in M_{d,n}$ be a matrix and set $r = \min\{d,n\}$. We express the singular value decomposition of the matrix W in the form

$$W = U\Sigma V^\top,$$

where $U \in O_d, V \in O_n$ and $\Sigma \in M_{d,n}$ is a diagonal matrix with nonnegative elements, that is $\Sigma = \mathrm{diag}(\sigma(W))$, where $\sigma(W) = (\sigma_i(W) : i \in \mathbb{N}_r) \in \mathbb{R}^r_+$. We assume that the singular values are *ordered* in a non-increasing sense, that is,

$$\sigma_1(W) \geq \cdots \geq \sigma_r(W) \geq 0\,.$$

Note that $\sigma(W)$ is uniquely defined in this way. Sometimes we also use $\Sigma(W)$ to denote the diagonal matrix Σ. The components of $\sigma(W)$ are the *singular values* of W. They are equal to the square root of the largest r eigenvalues of $W^\top W$, which are the same as those of $WW^\top$. We shall call functions of the singular values of a matrix *spectral functions*.

In the case of a symmetric matrix $A \in \mathbf{S}^n$, we similarly write

$$A = U\Lambda U^\top$$

for a spectral decomposition of A, where $U \in O_n$, $\Lambda = \mathrm{Diag}(\lambda(A))$ and $\lambda(A) = (\lambda_j : j \in \mathbb{N}_n)$ has components ordered in non-decreasing sense

$$\lambda_1(A) \geq \cdots \geq \lambda_n(A)\,.$$

In addition, for $x \in \mathbb{R}^r$, we shall use $|x|$ to denote the vector of absolute values $(|x_i| : i \in \mathbb{N}_r)$. Finally, for two vectors $x, y \in \mathbb{R}^r$ we write $x \leq y$ whenever, for all $i \in \mathbb{N}_r$, $x_i \leq y_i$.

4.2 Orthogonally Invariant Norms

A norm $\|\cdot\|$ on $M_{d,n}$ is called *orthogonally invariant* whenever, for every $U \in O_d$, $V \in O_n$ and $W \in M_{d,n}$, we have that

$$\|UWV^\top\| = \|W\|\,.$$

It is clear from the definition that an OI-norm that $\|\cdot\|$ is a spectral function. That is, for some function f, we have that $\|W\| = f(\sigma(W))$.

The remaining conditions on f which characterize OI-norms were given by von Neumann (1962) (see also Horn and Johnson, 1991, Section 3.5). He established that OI-norms are exactly *symmetric gauge functions* (SG-functions) of the singular values. To this end, we let $\mathcal{P}_r$ be the subset of $r \times r$ permutation matrices.

Definition 3 *A function $f : \mathbb{R}^r \to \mathbb{R}_+$ is called an SG-function whenever the following properties hold:*

1. *f is a norm on $\mathbb{R}^r$;*
2. *$f(x) = f(|x|)$ for all $x \in \mathbb{R}^r$;*
3. *$f(Px) = f(x)$ for all $x \in \mathbb{R}^r$ and all permutation matrices $P \in \mathcal{P}_r$.*

Property 2 states that f is *absolutely* or *gauge invariant*. Property 3 states that f is *symmetric* or *permutation invariant*. Hence, an SG-function is an absolutely symmetric norm.

Von Neumann's result is stated in the following theorem.

Theorem 4 *If $\|\cdot\|$ is an OI-norm on $M_{d,n}$ then there exists an SG-function $f:\mathbb{R}^r\to\mathbb{R}_+$ such that $\|W\|=f(\sigma(W))$, for all $W\in M_{d,n}$. Conversely, if $f:\mathbb{R}^r\to\mathbb{R}$ is an SG-function then the norm defined at $W\in M_{d,n}$, as $\|W\|=f(\sigma(W))$ is orthogonally invariant.*

The best known example of OI-norms are the *Schatten p-norms*, where $p\geq 1$, They are defined, for every $W\in M_{d,n}$, as

$$\|W\|_p=\left(\sum_{i\in\mathbb{N}_r}(\sigma_i(W))^p\right)^{\frac{1}{p}}$$

and, for $p=\infty$, as

$$\|W\|_\infty=\sigma_1(W).$$

The Schatten $1-$norm is sometimes called the *trace norm* or *nuclear norm*. Other common values of p give rise to the *Frobenius norm* ($p=2$) and the *spectral norm* ($p=\infty$). The Frobenius norm can also be written as $\sqrt{\mathrm{tr}W^\top W}$ and the spectral norm is alternatively expressed as $\max\{\|Wx\|_2:\|x\|_2=1\}$, where the subscript on the vector norm indicates the Euclidean norm of that vector. Another well-known family of OI-norms are the *Ky Fan norms* defined, for every $W\in M_{d,n}$ as

$$\|W\|_{(k)}=\sum_{i\in\mathbb{N}_k}\sigma_i(W)$$

where $1\leq k\leq r$ (the cases $k=1$ and $k=r$ are the spectral and trace norms, respectively). For more examples and for many interesting results involving OI-norms, we refer the reader to (Bhatia, 1997, Sec. IV.2) and (Horn and Johnson, 1991, Sec. 3.5).

We also mention, in passing, a formula from Argyriou et al. (2007b) which is useful for algorithmic developments. Specifically, we recall, for $p\in(0,2]$, that

$$\|W\|_p=\inf\left\{\langle WW^\top,D^{-\frac{2-p}{p}}\rangle:D\in\mathbf{S}^d_{++},\mathrm{tr}D\leq 1\right\}. \tag{10}$$

When $p\in[1,2]$, the function

$$(W,D)\mapsto\langle WW^\top,D^{-\frac{2-p}{p}}\rangle$$

is jointly convex in W and D and, so, the infimum in (10) is convex in W, in agreement with the convexity of the norm of $\|W\|_p$. Furthermore, if $WW^\top$ is invertible and $p\in(0,2]$, then the infimum is uniquely attained by the matrix

$$D=\frac{(WW^\top)^{\frac{p}{2}}}{\mathrm{tr}(WW^\top)^{\frac{p}{2}}}.$$

In machine learning practice, regularization with the trace norm has been proposed for collaborative filtering and multi-task learning (Abernethy et al., 2009; Argyriou et al., 2007a,b; Maurer, 2006a; Srebro et al., 2005, and references therein) and related problems (Yuan et al., 2007). If $\Omega(W)=\mathrm{rank}(W)$ the regularization problem (1) is non-convex. However, a common technique that overcomes this issue is to replace the rank by the trace norm (Fazel et al., 2001). The trace norm is the ℓ_1 norm on the singular values and hence there is an analogy to regularization of a vector

variable with the ℓ_1 norm, which is often used to obtain sparse solutions, see Candès and Recht (2008) and reference therein. In analogy to ℓ_1 regularization, it has recently been shown that for certain configurations of the input data the low rank solution can be recovered using the trace norm approach (Candès and Recht, 2008; Recht et al., 2008). More generally, regardless of the rank of the solution, it has been demonstrated that this approach allows for accurate joint estimation of multiple related tasks even when there are only *few* data points available for each task (Srebro et al., 2005; Argyriou et al., 2007a). One motivation is to approximate a matrix with a (possibly low-rank) factorization (Srebro et al., 2005). Another is that fitting multiple learning tasks simultaneously, so that they share a small set of orthogonal features, leads to a trace norm problem (Argyriou et al., 2007a).

The spectral norm, $\|\cdot\|_\infty$, is also of interest in the context of filter design (Zames, 1981) in control theory. Moreover, Schatten p-norms in the range $p \in [1,2]$ can be used for trading off sparsity of the model against task independence in multi-task learning (Argyriou et al., 2007b). In general, OI-norms are a natural class of regularizers to consider, since many matrix optimization problems can be posed in terms of the spectrum of the matrix.

We now proceed by reviewing some facts on duals of OI-norms. To this end, we first state a useful inequality, which can be found, for example, in (Horn and Johnson, 1991, ex. 3.3.10). This inequality also originates from von Neumann (1962) and is sometimes called *von Neumann's trace theorem* or *Ky Fan inequality*.

Lemma 5 *For any $X, Y \in M_{d,n}$, we have that*

$$\langle X, Y \rangle \leq \langle \sigma(X), \sigma(Y) \rangle \tag{11}$$

and equality holds if and only if there are $U \in O_d$ and $V \in O_n$ such that $X = U\Sigma(X)V^\top$ and $Y = U\Sigma(Y)V^\top$.

We emphasize that equality in (11) implies that the matrices X and Y admit the same ordered system of singular vectors, where the ordering is given by *ordering of the singular values*. It is also important to note that this inequality is stronger than the Cauchy-Schwarz inequality for the Frobenius norm, $\langle X, Y \rangle \leq \|X\|_2 \|Y\|_2$. Moreover, in the case of diagonal matrices one obtains a vector inequality due to Hardy et al. (1988)

$$\langle x, y \rangle \leq \langle [x], [y] \rangle ,$$

where $x, y \in \mathbb{R}^d$ and $[x]$ denotes the vector consisting of the components of x in non-increasing order. Let us also mention that apart from norm duality, Lemma 5 underlies many analytical properties of spectral functions, such as convexity, Fenchel conjugacy, subgradients and differentiability (see, for example, Lewis, 1995 for a review).

For our purposes, inequality (11) can be used to compute the dual of an OI-norm in terms of the dual of the corresponding SG-function. This is expressed in the following lemmas, which follow easily from (11) (see also Bhatia, 1997, Secs. IV.1, IV.2).

Lemma 6 *If the norm $\|\cdot\|$ on $M_{d,n}$ is orthogonally invariant and f is the corresponding SG-function, then the dual norm is given, for $W \in M_{d,n}$, by*

$$\|W\|_D = f_D(\sigma(W))$$

where $f_D : \mathbb{R}^r \to \mathbb{R}_+$ is the dual norm of f.

Lemma 7 *$\|\cdot\|$ is an OI-norm on $M_{d,n}$ if and only if $\|\cdot\|_D$ is orthogonally invariant. Also, $f:\mathbb{R}^r\to\mathbb{R}$ is an SG-function if and only if f_D is an SG-function.*

The next useful formula describes the peak set for any OI-norm.

Lemma 8 *Let $W\in M_{d,n}\setminus\{0\}$ and $W=U\Sigma(W)V^\top$ its singular value decomposition. If the norm $\|\cdot\|$ is orthogonally invariant and f is the corresponding SG-function, then*

$$\mathscr{D}\|W\| = \{Z : Z = U\Sigma(Z)V^\top, \sigma(Z)\in\mathscr{D}f(\sigma(W))\}.$$

Lemma 9 *Let $W\in M_{d,n}\setminus\{0\}$ and $W=U\Sigma(X)V^\top$ its singular value decomposition. If the norm $\|\cdot\|$ is orthogonally invariant and the corresponding SG-function f is* differentiable *at $\sigma(W)$, then $\|\cdot\|$ is differentiable at W and*

$$\nabla\|W\| = U\,\nabla f(\sigma(W))\,V^\top.$$

Lemma 10 *If f is an SG-function, $x\in\mathbb{R}^r\setminus\{0\}$ and $w\in\mathscr{D}f(x)$, then $\overline{w}\in\mathscr{D}f(\overline{x})$, where $\overline{w},\overline{x}\in\mathbb{R}^r$ are the vectors with elements the absolute values of w,x, respectively, in decreasing order. Moreover, $|w|,|x|$ can yield $\overline{w},\overline{x}$ with a simultaneous permutation of their elements.*

In other words, duality of OI-norms translates to duality of SG-functions. Norm duality preserves orthogonal invariance as well as the symmetric gauge properties. And dual pairs of matrices with respect to OI-norms directly relate to dual pairs of vectors with respect to SG-functions. Similarly, (sub)gradients of OI-norms correspond to (sub)gradients of SG-functions. In fact, Lemmas 8 and 9 hold, more generally, for all symmetric functions of the singular values (Lewis, 1995).

As an example, the dual of a Schatten p-norm $\|\cdot\|_p$ is the norm $\|\cdot\|_q$, where $\frac{1}{p}+\frac{1}{q}=1$. For $p>1$ and every $W=U\Sigma(W)V^\top\in M_{d,n}\setminus\{0\}$, one can readily obtain the set of duals from the equality conditions in Hölder's inequality. These give that

$$\mathscr{D}\|W\|_p = \left\{Z : Z = U\Sigma(Z)V^\top, \sigma_i(Z) = \frac{(\sigma_i(W))^{q-1}}{\|\sigma(W)\|_q^{q-1}}, i\in\mathbb{N}_r\right\}.$$

Moreover, this norm is differentiable for $p>1$ and the gradient is given by

$$\nabla\|W\|_p = U\,\mathrm{Diag}(\lambda)V^\top\frac{1}{\|\sigma(X)\|_p^{p-1}},$$

where $\lambda_i=(\sigma_i(W))^{p-1}, i\in\mathbb{N}_r$.

Before continuing to the main result about OI-norms, we briefly review the relation between *regularization* and *interpolation* problems, mentioned at the end of Section 2. We are interested in obtaining representer theorems and optimality conditions, in general, for regularization problems of the form (2). We shall focus, however, on representer theorems for interpolation problems of the form (3).

Let $\Omega:M_{d,n}\to\mathbb{R}$ be a given regularizer and assume that, for every $y\in\mathbb{R}^m$ and linear operator $I:M_{d,n}\to\mathbb{R}^m$, there exists some solution of (3) satisfying a prescribed representer theorem. Then, by Lemma 1, for every $y\in\mathbb{R}^m$, $I:M_{d,n}\to\mathbb{R}^m$ and $E:\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}$, the same representer theorem holds for some solution of problem (2). In the remainder of the paper we shall prove optimality conditions for interpolation problems, which thus equally apply to regularization problems.

Conversely, the representer theorem for the regularization problem (2) associated with certain choices of the function Ω and E , will also hold for the corresponding interpolation problems (3). To illustrate this idea, we adopt a result from Argyriou et al. (2009), which concerns the standard representer theorem,

$$\hat{W} \in \mathcal{R}(I^*)\,.$$

Theorem 11 *Let $E : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ and $\Omega : M_{d,n} \to \mathbb{R}$ be a function with the following properties:*

(i) *E is lower semicontinuous and bounded from below;*

(ii) *Ω is lower semicontinuous and has bounded sublevel sets, that is, for every $\lambda \in \mathbb{R}$, the set $\{W : W \in M_{d,n}, \Omega(W) \leq \lambda\}$ is bounded;*

(iii) *for some $v \in \mathbb{R}^m \setminus \{0\}, y \in \mathbb{R}^m$, there exists a* unique *minimizer of $\min\{E(av, y) : a \in \mathbb{R}\}$ and this minimizer does not equal zero.*

If, for all choices of I and y, there exists a solution $\hat{W} \in \mathcal{R}(I^*)$ of (2)*, then, for all choices of I and y such that $y \in \mathcal{R}(I)$, there exists a solution $\hat{W} \in \mathcal{R}(I^*)$ of* (3)*.*

As noted in Argyriou et al. (2009), the square loss, hinge loss or logistic loss are all valid error functions in this theorem. The above results allow us to focus on the interpolation problems, as a devise to study the regularization problem.

We are now ready to describe the main result of this section, which concerns the form of the solution of interpolation problems (5) for the class of OI-norms.

Theorem 12 *Assume that $\|\cdot\|$ is an OI-norm and let f be the corresponding SG-function. If the matrix $\hat{W} \in M_{d,n} \setminus \{0\}$ is a solution of* (5) *and the vector $\hat{c} \in \mathbb{R}^m$ is a solution of* (7)*, then*

$$\hat{W} = U\,\Sigma(\hat{W})\,V^\top\,,\; I^*(\hat{c}) = U\,\Sigma(I^*(\hat{c}))\,V^\top$$

for some $U \in O^d, V \in O^n$, and

$$\sigma(\hat{W}) \in \frac{1}{\|I^*(\hat{c})\|_D}\,\partial f(\sigma(I^*(\hat{c})))\,.$$

Proof By Theorem 2 we obtain that $\|I^*(\hat{c})\|_D \hat{W} \in \partial\|I^*(\hat{c})\|$. We can write $I^*(\hat{c})$ in terms of its singular value decomposition, $I^*(\hat{c}) = U\,\Sigma(I^*(\hat{c}))\,V^\top$ with $U \in O^d, V \in O^n$. Using Lemma 8 we conclude that

$$\|I^*(\hat{c})\|_D \hat{W} = U\,(\|I^*(\hat{c})\|_D \Sigma(\hat{W}))\,V^\top,$$

where $\|I^*(\hat{c})\|_D\,\sigma(\hat{W}) \in \partial f(\sigma(I^*(\hat{c})))$. ■

This theorem implies that, in order to solve the minimal norm interpolation problem (5), we may first solve the dual problem (7) and then find a matrix in the peak set of $I^*(\hat{c})$ scaled by $1/\|I^*(\hat{c})\|_D$, which interpolates the data. The latter step in turn requires computing a singular value decomposition of $I^*(\hat{c})$ and then solving a non-linear system of equations. However, when the SG-function is smooth, there is a unique elements in the peak set and, so, there is no need to solve the non-linear equations. For example, if $\|\cdot\|$ is the Frobenius norm, Theorem 12 readily yields the standard representer theorem.

4.3 Admissible Orthogonally Invariant Norms

In this section, we define a subclass of OI-norms, which obey an improved version of the representer theorem presented above.

We begin with a definition.

Definition 13 *A norm $\|\cdot\|$ on $\mathbb{R}^r$ is said to be* admissible *if for any $x \in \mathbb{R}^r$, any $k \in \mathbb{N}_r$ such that $x_k \neq 0$ we have that*

$$\|x^k\| < \|x\|$$

where x^k is the vector all of whose components agree with x, except the k-th component which is zero.

The simplest example of admissible norms are the ℓ_p norm on $\mathbb{R}^d$, $\|\cdot\|_p$, for $p \in [1,\infty)$. From this norm we can form other admissible norms in various ways. Specifically, for any $p_1, p_2 \in [1,\infty)$, we see that the norm $\|\cdot\|_{p_1} + \|\cdot\|_{p_2}$ or the norm $\max\{\|\cdot\|_{p_1}, \|\cdot\|_{p_2}\}$ are both admissible. Note that some of these norms are not strictly convex. Also compare this definition to that of weakly monotone norms (Horn and Johnson, 1985, Def. 5.5.13).

Lemma 14 *If $\|\cdot\|$ is an admissible norm on $\mathbb{R}^r$, $x \in \mathbb{R}^r \backslash \{0\}$ and $w \in \mathscr{D}\|x\|$, then for any $k \in \mathbb{N}_r$ with $x_k = 0$ it holds that $w_k = 0$.*

Conversely, assume that, for every $x \in \mathbb{R}^r \backslash \{0\}$, $w \in \mathscr{D}\|x\|$ and $k \in \mathbb{N}_r$, if $x_k = 0$ it holds that $w_k = 0$. Then $\|\cdot\|$ is admissible.

Proof Let $w \in \mathscr{D}\|x\|$, where $x \in \mathbb{R}^r \backslash \{0\}$, with $x_k = 0$. Suppose to the contrary that $w_k \neq 0$. Since $\|\cdot\|$ is admissible it follows that $\|w^k\| < \|w\|$, and so, we get that $\|w^k\| < 1$, because $\|w\| = 1$. However, we also have that

$$\|x\|_D = \langle w, x\rangle = \langle w^k, x\rangle \leq \|w^k\|\|x\|_D$$

from which it follows that $\|w^k\| \geq 1$. This proves the first part of the claim.

For the converse, we consider a $w \in \mathbb{R}^r \backslash \{0\}$ with $w_k \neq 0$. We shall show that $\|w^k\| < \|w\|$. To this end, we choose $x \in \mathscr{D}\|w^k\|_D$ and then we choose $y \in \mathscr{D}\|x^k\|$. By our hypothesis, we conclude that $y_k = 0$ and by our choice we have, in particular that $1 = \|y\| = \|x\|_D$. Consequently, it follows that

$$\|x^k\|_D = \langle y, x^k\rangle = \langle y, x\rangle \leq \|y\|\|x\|_D = 1$$

from which conclude that

$$\|w^k\| = \langle w^k, x\rangle = \langle w, x^k\rangle \leq \|w\|. \tag{12}$$

Moreover, if equality holds in this inequality it would follow that $\frac{w}{\|w\|} \in \mathscr{D}\|x^k\|$. But then, we can invoke our hypothesis once again and obtain a contradiction. That is, inequality (12) is strict and therefore $\|\cdot\|$ is an admissible norm, as asserted. ■

The above observation leads us to consider the following subclass of OI-norms.

Definition 15 *A norm $\|\cdot\|$ on $M_{d,n}$ is said to be* admissible orthogonally invariant *if there exists an admissible vector norm $\|\cdot\|$ on $\mathbb{R}^r$ such that, for every $W \in M_{d,n}$, we have that $\|W\| = \|\sigma(W)\|$.*

Examples of non-admissible OI-norms are the spectral norm, the Ky Fan norms $\|\cdot\|_{(k)}$ for $1 \leq k < r$ and the norm $\max\{\|\cdot\|_1, \alpha\|\cdot\|_\infty\}$ for $\alpha \in (1,\infty)$.

We have now accumulated sufficient information on admissible OI-norms to present an improved representer theorem for problem (5). We shall prove below, for any admissible OI-norm, $\hat{W}$ can be expressed as

$$\hat{W} = \sum_{i \in \mathbb{N}_m} \hat{c}_i X_i R.$$

In other words, $\hat{W}$ is obtained by first applying the standard representer theorem and then multiplying it from the right by the matrix R. In the case of the Frobenius norm $R = I_n$.

Theorem 16 *If $\|\cdot\|$ is admissible orthogonally invariant, the matrix $\hat{W} \in M_{d,n} \setminus \{0\}$ is a solution of* (5) *and the vector $\hat{c} \in \mathbb{R}^m$ is a solution of* (7)*, then there exists a matrix $R \in \mathbf{S}^n_+$ such that*

$$\hat{W} = I^*(\hat{c})R \tag{13}$$

and the eigenvectors of R are right singular vectors of $I^*(\hat{c})$.

Proof By Theorem 12, there exists $I^*(\hat{c}) = U\Sigma(I^*(\hat{c}))V^\top$, obtained from a dual solution $\hat{c}$ of (7), such that $\|I^*(\hat{c})\|_D \hat{W} = U\,\mathrm{Diag}(\lambda)V^\top$, where $\lambda \in \partial f(\sigma(I^*(\hat{c})))$ and f is the SG-function associated with $\|\cdot\|$. Since f is admissible, Lemma 14 implies that $\lambda_i = 0$ whenever $\sigma_i(I^*(\hat{c})) = 0$. Hence there exists $\mu \in \mathbb{R}^r_+$ such that $\lambda_i = \sigma_i(I^*(\hat{c}))\mu_i$, $i \in \mathbb{N}_r$, and $\mu_i = 0$ whenever $\sigma_i(I^*(\hat{c})) = 0$. Thus, $\|I^*(\hat{c})\|_D \hat{W} = U\Sigma(I^*(\hat{c}))V^\top V\,\mathrm{Diag}(\mu)V^\top$ and the corollary follows by selecting $R = \dfrac{1}{\|I^*(\hat{c})\|_D} V\,\mathrm{Diag}(\mu)V^\top$. ■

Note that, in the above theorem, the eigenvectors of R need not correspond to right singular vectors of $I^*(\hat{c})$ according to a simultaneous ordering of the eigenvalues / singular values.

We may also state a converse of Theorem 16, that is, the only OI-norms which satisfy property (13) are admissible.

Theorem 17 *If $\|\cdot\|$ is orthogonally invariant and condition* (13) *holds (without any conditions on $R \in M_{n,n}$), for every linear operator $I : M_{d,n} \to \mathbb{R}^m$, $y \in \mathcal{R}(I) \setminus \{0\}$, every solution $\hat{W}$ of* (5) *and every solution $\hat{c}$ of* (7)*, then the norm $\|\cdot\|$ is admissible orthogonally invariant.*

Proof Let f be the SG-function corresponding to $\|\cdot\|$. Let arbitrary $x \in \mathbb{R}^r \setminus \{0\}$ and $w \in \partial f(x)$. Define $\overline{x}, \overline{w} \in \mathbb{R}^r_+$ to be the vectors with elements the absolute values of x, w, respectively, in descending order. By Lemma 10, we obtain that $\overline{w} \in \partial f(\overline{x})$. Define also $X = \mathrm{Diag}(\overline{x}) \in M_{d,n}$ and $W = \mathrm{Diag}(\overline{w}) \in M_{d,n}$. By Lemma 8, we obtain that $W \in \partial\|X\|$. Now, consider the problem $\min\{\|Z\| : Z \in M_{d,n}, \langle Z, X\rangle = \|X\|_D\}$, whose set of solutions is $\partial\|X\|$. By hypothesis, $W = cXR$ for some $R \in M_{n,n}$ and for $c = \frac{1}{\|X\|_D}$ (the only solution of the dual problem). Therefore, $\mathrm{Diag}(\overline{w}) = c\,\mathrm{Diag}(\overline{x})R$ and hence $\overline{x}_k = 0$ implies $\overline{w}_k = 0$, for all $k \in \mathbb{N}_r$. By Lemma 10, this implies in turn that $w_k = 0$ if $x_k = 0$, for all $k \in \mathbb{N}_r$. Combining with Lemma 14, we deduce that f is admissible, as required. ■

We remark that there exist norms on $M_{d,n}$ which are not orthogonally invariant and satisfy condition (13). In fact, given two non-singular matrices $Q \in M_{d,d}$ and $M \in M_{n,n}$, the norm $W \mapsto \|QWM\|_2$ is not orthogonally invariant, but the representer theorem is easily seen to be

$$\hat{W} = (Q^\top Q)^{-1} I^*(\hat{c})(MM^\top)^{-1}.$$

Furthermore, concerning the converse of Theorem 16, it can be shown that if we restrict the eigenvectors of R to be right singular vectors of $I^*(\hat{c})$, then $\|\cdot\|$ has to be orthogonally invariant.

Moreover, if the norm $\|\cdot\|$ is not admissible, then it can be shown that for every solution $\hat{c}$ of the dual problem there exists a solution of the primal satisfying (13). As an example, see Corollary 20 below (spectral norm). For a characterization of functions yielding such representer theorems, see Argyriou et al. (2009).

Returning to Theorem 12, for the Schatten p-norms we have the following corollary. To state it, we use the notation A^{q-1} as a shorthand for the matrix $U \operatorname{Diag}(\sigma_i(A)^{q-1})_{i \in \mathbb{N}_r} V^\top$ when $A = U\Sigma(A)V^\top$.

Corollary 18 *If the matrix $\hat{W} \in M_{d,n} \setminus \{0\}$ is a solution of* (5) *for the Schatten $p-$norm, with $p \in (1,\infty)$, then there exists a vector $\hat{c} \in \mathbb{R}^m$ such that*

$$\hat{W} = \frac{I^*(\hat{c})^{q-1}}{\|I^*(\hat{c})\|_q^q},$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof The corollary follows directly from Theorem 12 and the description of $\partial \|\cdot\|_p$ in Section 4.2. ■

The above corollary does not cover the cases that $p = 1$ or $p = \infty$. We state them separately.

Corollary 19 *If $\hat{W} \in M_{d,n} \setminus \{0\}$ is a solution of* (5) *for the trace norm, $\hat{c} \in \mathbb{R}^m$ a solution of* (7) *and $I^*(\hat{c}) = \sum_{i \in \mathbb{N}_r} \sigma_i(I^*(\hat{c})) u_i v_i^\top$ is a singular value decomposition, then*

$$\hat{W} = \frac{1}{\sigma_1(I^*(\hat{c}))} \sum_{i \in \mathbb{N}_{r_{\max}}} \lambda_i u_i v_i^\top,$$

for some $\lambda_i \geq 0, i \in \mathbb{N}_{r_{\max}}$ such that $\sum_{i \in \mathbb{N}_{r_{\max}}} \lambda_i = 1$, where $r_{\max}$ is the multiplicity of the largest singular value $\sigma_1(I^*(\hat{c}))$. Moreover, $\hat{W} = I^*(\hat{c})R$, where

$$R = \frac{1}{\sigma_1^2(I^*(\hat{c}))} \sum_{i \in \mathbb{N}_{r_{\max}}} v_i v_i^\top.$$

Proof The corollary follows from Theorem 12 and the description of $\partial \|\cdot\|_1$. From the definition, it is easy to obtain that, for every $x \in \mathbb{R}^r_+$, $\partial \|x\|_1 = \{y \in \mathbb{R}^r_+ : y_i = 0, \text{ if } x_i < \|x\|_\infty, \sum_{i \in \mathbb{N}_r} y_i = 1\}$. Thus, $\sigma_1(I^*(\hat{c}))\hat{W} = \|I^*(\hat{c})\|_\infty \hat{W} = U\Lambda V^\top$, for $\Lambda = \operatorname{Diag}(\lambda)$ and $\lambda_i = 0$ for $i > r_{\max}$, $\sum_{i \in \mathbb{N}_{r_{\max}}} \lambda_i = 1$. Since $\Lambda = \frac{1}{\sigma_1(I^*(\hat{c}))}\Sigma(I^*(\hat{c}))\Lambda$, R can be selected as $\frac{1}{\sigma_1^2(I^*(\hat{c}))} V\Lambda V^\top$. ■

Corollary 20 *If the matrix $\hat{W} \in M_{d,n} \setminus \{0\}$ is a solution of* (5) *for the spectral norm, $\hat{c} \in \mathbb{R}^m$ a solution of* (7) *and $I^*(\hat{c}) = \sum_{i \in \mathbb{N}_r} \sigma_i(I^*(\hat{c}))u_i v_i^\top$ is a singular value decomposition, then*

$$\hat{W} = \frac{1}{\|I^*(\hat{c})\|_1} \sum_{i=1}^{\operatorname{rank}(I^*(\hat{c}))} u_i v_i^\top + \sum_{i=\operatorname{rank}(I^*(\hat{c}))+1}^{r} \alpha_i u_i v_i^\top , \tag{14}$$

for some $\alpha_i \in [0,1]$, $i = \operatorname{rank}(I^*(\hat{c})) + 1, \ldots, r$.

Proof The corollary follows from Theorem 12 and the fact that, for every $x \in \mathbb{R}^r_+$, $\mathscr{D}\|x\|_\infty = \{y \in [-1,1]^r : y_i = 1, \text{ if } x_i > 0\}$. ■

The above corollary also confirms that representation (13) does not apply to the spectral norm (which is not admissible orthogonally invariant). Indeed, from (14) it is clear that the range of $\hat{W}$ can be a superset of the range of $I^*(\hat{c})$.

To recapitulate the results presented in this section, Theorem 12 allows one to obtain the solutions of the primal minimum norm interpolation problem (5) from those of its dual problem (7), which involves m variables. This is true for *all OI-norms*, even though the representer theorem in the form (13) applies only to admissible OI-norms. Part of the appeal of OI-norms is that computing primal solutions from dual ones reduces to a vector norm optimization problem. Indeed, given a solution of the dual problem, one just needs to compute the *singular value decomposition* of the matrix $I^*(\hat{c})$ and *the peak set of the SG-function f at the singular values*. The associated primal solutions are then easily obtained by keeping the same row and column spaces and using elements of the peak set in place of the singular values. In fact, in many cases, the latter problem of computing the peak set of f may be straightforward. For example, if f_D is differentiable (except at zero), each dual solution is associated with a single primal one, which equals a multiple of the gradient of f_D at the dual solution.

4.4 Related Work

The results of Section 4 are related to other prior work, besides the already mentioned literature on representer theorems for the case of the vector L_2 norm (that is, for $n = 1$). In particular, the representer theorem for the trace norm (Corollary 19) has been stated in Srebro et al. (2005). Also, the representation (13) in Theorem 16 relates to the representer theorems proven in Argyriou et al. (2009); Abernethy et al. (2009). The results in Abernethy et al. (2009) apply to the case of the trace norm and when the X_i are rank one matrices. The results in Argyriou et al. (2009) give representer theorems for a broad class of functions, of which differentiable OI-norms are members. However, as mentioned before, Theorem 16 requires additional conditions on matrix R. In particular, the requirement on the eigenvectors of this matrix holds only for admissible OI-norms.

5. Conclusion and Future Work

We have characterized the form of the solution of regularization with an orthogonally invariant penalty term. Our result depends upon a detailed analysis of the corresponding minimal norm interpolation problem. In particular, we have derived a dual problem of the minimal norm interpolation problem and established the relationship between these two problems. The dual problem involves optimization over a vector of parameters whose size equals the number of data points. In practical circumstances, this number may be smaller than the dimension of the matrix we seek, thus our result

should prove useful in the development of optimization algorithms for orthogonally invariant norm regularization. For example, one could combine our result with Lemma 9 in order to implement gradient methods for solving the dual problem. Note however that the dual problem involves a singular value decomposition, and more effort is needed in elucidating the algorithmic implications of the results presented here.

Acknowledgments

The work of the first author and third author was partially supported by EPSRC Grant EP/D071542/1. The work of the second author was supported by NSF Grant ITR-0312113 and Air Force Grant AFOSR-FA9550.

Appendix A.

Here, we describe two results which we have used in the paper. Recall that for every linear operator $J : M_{d,n} \to \mathbb{R}^k$, the linear spaces $\mathcal{R}(J)$ and $\mathcal{N}(J)$ denote the range and the kernel of J, respectively.

Lemma 21 *Let $\mathcal{W}$ be a nonempty, convex and compact subset of $M_{d,n}$ and let $J : M_{d,n} \to \mathbb{R}^k$ be a linear operator. The set $\mathcal{R}(J^*)$ intersects $\mathcal{W}$ if and only if, for every $X \in \mathcal{N}(J)$ the following inequality holds*

$$\max\{\langle X, W\rangle : W \in \mathcal{W}\} \geq 0. \tag{15}$$

Proof Suppose that there exist $z \in \mathbb{R}^k$ and $T \in \mathcal{W}$ such that $J^*(z) = T$. Then for any $X \in M_{d,n}$ with $J(X) = 0$ we have that $\langle X, T\rangle = 0$ and, so, inequality (15) holds true.

Now, suppose that $\mathcal{R}(J^*) \cap \mathcal{W} = \emptyset$. Then, there is a hyperplane which strictly separates $\mathcal{R}(J^*)$ from $\mathcal{W}$ (see, for example, Rockafellar, 1970, Cor. 11.4.2). That is, there exist $W_0 \in M_{d,n}$ and $\mu \in \mathbb{R}$ such that, for all $z \in \mathbb{R}^k$,

$$\langle W_0, J^*(z)\rangle + \mu \geq 0,$$

while, for all $W \in \mathcal{W}$,

$$\langle W_0, W\rangle + \mu < 0.$$

The first inequality implies that $J(W_0) = 0$. To see this, we choose any $z_0 \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}$ and let $z = \lambda z_0$ in the first inequality. Now, we allow $\lambda \to \pm\infty$, to obtain that $\langle W_0, J^*(z_0)\rangle = 0$. Therefore, the first inequality simplifies to the statement that $\mu \geq 0$.

The second inequality implies that

$$\max\{\langle W_0, W\rangle : W \in \mathcal{W}\} < -\mu \leq 0,$$

which contradicts (15) and proves the result. ■

Next, we state an important rule for taking directional derivatives of a convex function expressed as a maximum of a family of convex functions. For this purpose, recall that the right directional derivative of a function $g : \mathcal{W} \to \mathbb{R}$ in the direction Δ at $W \in \mathcal{W}$ is defined as

$$g'_+(W;\Delta) = \lim_{\lambda \to 0^+} \frac{g(W + \lambda\Delta) - g(W)}{\lambda}.$$

Theorem 22 *Let $\mathcal{W}$ be a convex subset and $\mathcal{X}$ a compact subset of $M_{d,n}$ and $f : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$. If, for every $W \in \mathcal{W}$, the function $X \mapsto f(W,X)$ is continuous on $\mathcal{X}$ and, for every $X \in \mathcal{X}$ the function $W \mapsto f(W,X)$ is convex on $\mathcal{W}$, then the convex function $g : \mathcal{W} \to \mathbb{R}$ defined at $W \in \mathcal{W}$ as*

$$g(W) := \max\{ f(W,X) : X \in \mathcal{X}\}$$

has a right directional derivative at W in the direction $\Delta \in M_{d,n}$, given as

$$g'_+(W;\Delta) = \max\{ f'_+(W;\Delta,X) : X \in M(W)\}, \tag{16}$$

where $M(W) = \{X : X \in \mathcal{X}, f(W,X) = g(W)\}$.

References

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

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

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

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

R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer, 1997.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer, 2005.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717-772, 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.

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

M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings, American Control Conference*, volume 6, pages 4734–4739, 2001.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1988.

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.

A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1/2):173–183, 1995.

A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006a.

A. Maurer. Learning similarity with operator-valued large-margin classifier. *Journal of Machine Learning Research*, 9:1049–1082, 2008a.

C. A. Micchelli and A. Pinkus. Variational problems arising from balancing several error criteria. *Rendiconti di Matematica, Serie VII*, 14:37–86, 1994.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. Preprint, 2008.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

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

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.

J. von Neumann. Some matrix inequalities and metrization of matric-space. In *Tomsk University Review, 1:286-300, 1937*, volume IV of *Collected Works*, pages 205–218. Pergamon, Oxford, 1962.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

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.

G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26(2):301–320, 1981.