Journal of Machine Learning Research 13 (2012) 671-690 Submitted 9/11; Revised 2/12; Published 3/12

Structured Sparsity and Generalization

Andreas Maurer AM@ANDREAS-MAURER.EU
Adalbertstr. 55
D-80799, München
GERMANY

Massimiliano Pontil M.PONTIL@CS.UCL.AC.UK
Department of Computer Science
University College London
Gower St.
London, UK

Editor: Gabor Lugosi

Abstract

We present a data dependent generalization bound for a large class of regularized algorithms which implement structured sparsity constraints. The bound can be applied to standard squared-norm regularization, the Lasso, the group Lasso, some versions of the group Lasso with overlapping groups, multiple kernel learning and other regularization schemes. In all these cases competitive results are obtained. A novel feature of our bound is that it can be applied in an infinite dimensional setting such as the Lasso in a separable Hilbert space or multiple kernel learning with a countable number of kernels.

Keywords: empirical processes, Rademacher average, sparse estimation.

1. Introduction

We study a class of regularization methods used to learn a linear function from a finite set of examples. The regularizer is expressed as an infimum convolution which involves a set $\mathcal{M}$ of linear transformations (see Equation (1) below). As we shall see, this regularizer generalizes, depending on the choice of the set $\mathcal{M}$, the regularizers used by several learning algorithms, such as ridge regression, the Lasso, the group Lasso (Yuan and Lin, 2006), multiple kernel learning (Lanckriet et al., 2004; Bach et al., 2004), the group Lasso with overlap (Obozinski et al., 2009), and the regularizers in Micchelli et al. (2010).

We give a bound on the Rademacher average of the linear function class associated with this regularizer. The result matches existing bounds in the above mentioned cases but also admits a novel, dimension free interpretation. In particular, the bound applies to the Lasso in a separable Hilbert space or to multiple kernel learning with a countable number of kernels, under certain finite second-moment conditions.

We now introduce some necessary notation and state our main results. Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $\mathcal{M}$ be an at most countable set of symmetric bounded linear operators on H such that for every $x \in H$, $x \neq 0$, there is some linear operator $M \in \mathcal{M}$ with $Mx \neq 0$ and that $\sup_{M \in \mathcal{M}} |||M||| < \infty$, where $||| \cdot |||$ is the operator norm. Define the function $\|\cdot\|_{\mathcal{M}} : H \to \mathbb{R}_+ \cup \{\infty\}$ by

$$\|\beta\|_{\mathcal{M}} = \inf\left\{ \sum_{M\in\mathcal{M}} \|v_M\| : v_M \in H,\ \sum_{M\in\mathcal{M}} M v_M = \beta \right\}. \tag{1}$$

It is shown in Section 3.2 that the chosen notation is justified, because $\|\cdot\|_{\mathcal{M}}$ is indeed a norm on the subspace of H where it is finite, and the dual norm is, for every $z \in H$, given by

$$\|z\|_{\mathcal{M}^*} = \sup_{M\in\mathcal{M}} \|Mz\|.$$

The somewhat complicated definition of $\|\cdot\|_{\mathcal{M}}$ is contrasted by the simple form of the dual norm.

As an example, if $H = \mathbb{R}^d$ and $M = \{P_1, \dots, P_d\}$, where P_i is the orthogonal projection on the i-th coordinate, then the function (1) reduces to the ℓ_1 norm.

Using well known techniques, as described in Koltchinskii and Panchenko (2002) and Bartlett and Mendelson (2002), our study of generalization reduces to the search for a good bound on the empirical Rademacher complexity of a set of linear functionals with $\|\cdot\|_{\mathcal{M}}$-bounded weight vectors

$$\mathcal{R}_{\mathcal{M}}(\mathbf{x}) = \frac{2}{n}\mathbb{E} \sup_{\beta:\ \|\beta\|_{\mathcal{M}}\leq 1} \sum_{i=1}^{n} \varepsilon_i \langle \beta, x_i \rangle, \tag{2}$$

where $\mathbf{x} = (x_1, \dots, x_n) \in H^n$ is a sample vector representing observations, and $\varepsilon_1, \dots, \varepsilon_n$ are Rademacher variables, mutually independent and each uniformly distributed on $\{-1, 1\}$.[1] Given a bound on $\mathcal{R}_{\mathcal{M}}(\mathbf{x})$ we obtain uniform bounds on the estimation error, for example using the following standard result (adapted from Bartlett and Mendelson 2002), where the Lipschitz function ϕ is to be interpreted as a loss function.

Theorem 1 *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a vector of iid random variables with values in H, let X be iid to X_1, let $\phi : \mathbb{R} \to [0, 1]$ have Lipschitz constant L and $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ in the draw of $\mathbf{X}$ it holds, for every $\beta \in \mathbb{R}^d$ with $\|\beta\|_{\mathcal{M}} \leq 1$, that*

$$\mathbb{E}\phi(\langle \beta, X \rangle) \leq \frac{1}{n}\sum_{i=1}^{n} \phi(\langle \beta, X_i \rangle) + L\ \mathcal{R}_{\mathcal{M}}(\mathbf{X}) + \sqrt{\frac{9 \ln 2/\delta}{2n}}.$$

A similar (slightly better) bound is obtained if $\mathcal{R}_{\mathcal{M}}(\mathbf{X})$ is replaced by its expectation $\mathcal{R}_{\mathcal{M}} = \mathbb{E}\mathcal{R}_{\mathcal{M}}(\mathbf{X})$ (see Bartlett and Mendelson 2002).

The following is the main result of this paper and leads to consistency proofs and finite sample generalization guarantees for all algorithms which use a regularizer of the form (1). A proof is given in Section 3.3.

Theorem 2 *Let $\mathbf{x} = (x_1, \dots, x_n) \in H^n$ and $\mathcal{R}_{\mathcal{M}}(\mathbf{x})$ be defined as in (2). Then*

$$\begin{aligned}
\mathcal{R}_{\mathcal{M}}(\mathbf{x}) &\leq \frac{2^{3/2}}{n}\sqrt{\sup_{M\in\mathcal{M}} \sum_{i=1}^{n} \|Mx_i\|^2}\left(2 + \sqrt{\ln\left(\sum_{M\in\mathcal{M}} \frac{\sum_i \|Mx_i\|^2}{\sup_{N\in\mathcal{M}} \sum_j \|Nx_j\|^2}\right)}\right) \\
&\leq \frac{2^{3/2}}{n}\sqrt{\sum_{i=1}^{n} \|x_i\|_{\mathcal{M}^*}^2}\left(2 + \sqrt{\ln|\mathcal{M}|}\right).
\end{aligned}$$

[1] Our definition coincides with the one in Bartlett and Mendelson (2002), while other authors omit the factor of 2. This is relevant when comparing the constants in different bounds.

The second inequality follows from the first one, the inequality

$$\sup_{M\in\mathcal{M}}\sum_{i=1}^{n}\|Mx_i\|^2\leq\sum_{i=1}^{n}\|x_i\|_{\mathcal{M}^*}^2,$$

a fact which will be tacitly used in the sequel, and the observation that every summand in the logarithm appearing in the first inequality is bounded by 1. Of course the second inequality is relevant only if $\mathcal{M}$ is finite. In this case we can draw the following conclusion: If we have an a priori bound on $\|X\|_{\mathcal{M}^*}$ for some data distribution, say $\|X\|_{\mathcal{M}^*}\leq C$, and $\mathbf{X}=(X_1,\ldots,X_n)$, with X_i iid to X, then

$$\mathcal{R}_{\mathcal{M}}(\mathbf{X})\leq\frac{2^{3/2}C}{\sqrt{n}}\left(2+\sqrt{\ln|\mathcal{M}|}\right),$$

thus passing from a data-dependent to a distribution dependent bound. In Section 2 we show that this recovers existing results (Cortes et al., 2010; Kakade et al., 2010; Kloft et al., 2011; Meir and Zhang, 2003; Ying and Campbell, 2009) for many regularization schemes.[2]

But the first bound in Theorem 2 can be considerably smaller than the second and may be finite even if $\mathcal{M}$ is infinite. This gives rise to some novel features, even in the well studied case of the Lasso, when there is a (finite but potentially large) ℓ_2-bound on the data.

Corollary 3 *Under the conditions of Theorem 2 we have*

$$\mathcal{R}_{\mathcal{M}}(\mathbf{x})\leq\frac{2^{3/2}}{n}\sqrt{\sup_{M\in\mathcal{M}}\sum_i\|Mx_i\|^2}\left(2+\sqrt{\ln\frac{1}{n}\sum_i\sum_{M\in\mathcal{M}}\|Mx_i\|^2}\right)+\frac{2}{\sqrt{n}}.$$

A proof is given in Section 3.3. To obtain a distribution dependent bound we retain the condition $\|X\|_{\mathcal{M}^*}\leq C$ and replace finiteness of $\mathcal{M}$ by the condition that

$$R^2:=\mathbb{E}\sum_{M\in\mathcal{M}}\|MX\|^2<\infty. \tag{3}$$

Taking the expectation in Corollary 3 and using Jensen's inequality then gives a bound on the expected Rademacher complexity

$$\mathcal{R}_{\mathcal{M}}\leq\frac{2^{3/2}C}{\sqrt{n}}\left(2+\sqrt{\ln R^2}\right)+\frac{2}{\sqrt{n}}. \tag{4}$$

The key features of this result are the dimension-independence and the only logarithmic dependence on R^2, which in many applications turns out to be simply $R^2=\mathbb{E}\|X\|^2$.

The rest of the paper is organized as follows. In the next section, we specialize our results to different regularizers. In Section 3, we present the proof of Theorem 2 as well as the proof of other results mentioned above. In Section 4, we discuss the extension of these results to the ℓ_q case. Finally, in Section 5, we draw our conclusions and comment on future work.

[2] We note that the numerical implementation and practical application of specific cases of the regularizer described here have been addressed in detail in a number of papers. We recommend Baldassarre et al. (2012), Obozinski et al. (2009) and Jenatton et al. (2011) and references therein for detailed information on such matters. We also refer to Baraniuk et al. (2010) and Huang et al. (2009) for related work using greedy methods.

2. Examples

Before giving the examples we mention a great simplification in the definition of the norm $\|\cdot\|_{\mathcal{M}}$ which occurs when the members of $\mathcal{M}$ have mutually orthogonal ranges. A simple argument, given in Proposition 8 below shows that in this case

$$\|\beta\|_{\mathcal{M}} = \sum_{M \in \mathcal{M}} \left\| M^{+}\beta \right\|,$$

where M^{+} is the pseudoinverse of M. If, *in addition*, every member of $\mathcal{M}$ is an orthogonal projection P, the norm further simplifies to

$$\|\beta\|_{\mathcal{M}} = \sum_{P \in \mathcal{M}} \|P\beta\|,$$

and the quantity R^2 occurring in the second moment condition (3) simplifies to

$$R^2 = \mathbb{E} \sum_{P \in \mathcal{M}} \|PX\|^2 = \mathbb{E}\,\|X\|^2.$$

For the remainder of this section $\mathbf{X} = (X_1, \dots, X_n)$ will be a generic iid random vector of data points, $X_i \in H$, and X will be a generic data variable, iid to X_i. If $H = \mathbb{R}^d$ we write $(X)_k$ for the k-th coordinate of X, not to be confused with X_k, which would be the k-th member of the vector $\mathbf{X}$.

2.1 The Euclidean Regularizer

In this simplest case we set $\mathcal{M} = \{I\}$, where I is the identity operator on the Hilbert space H. Then $\|\beta\|_{\mathcal{M}} = \|\beta\|$, $\|z\|_{\mathcal{M}^*} = \|z\|$, and the bound on the empirical Rademacher complexity becomes

$$\mathcal{R}_{\mathcal{M}}(\mathbf{x}) \leq \frac{2^{5/2}}{n} \sqrt{\sum_i \|x_i\|^2},$$

worse by a constant factor of $2^{3/2}$ than the corresponding result in Bartlett and Mendelson (2002), a tribute paid to the generality of our result.

2.2 The Lasso

Let us first assume that $H = \mathbb{R}^d$ is finite dimensional and set $\mathcal{M} = \{P_1, \dots, P_d\}$ where P_k is the orthogonal projection onto the 1-dimensional subspace generated by the basis vector e_k. All the above mentioned simplifications apply and we have $\|\beta\|_{\mathcal{M}} = \|\beta\|_1$ and $\|z\|_{\mathcal{M}^*} = \|z\|_\infty$. The bound on $\mathcal{R}_{\mathcal{M}}(\mathbf{x})$ now reads

$$\mathcal{R}_{\mathcal{M}}(\mathbf{x}) \leq \frac{2^{3/2}}{n} \sqrt{\sum_i \|x_i\|_\infty^2} \left(2 + \sqrt{\ln d}\right).$$

If $\|X\|_\infty \leq 1$ almost surely we obtain

$$\mathcal{R}_{\mathcal{M}}(\mathbf{X}) \leq \frac{2^{3/2}}{\sqrt{n}} \left(2 + \sqrt{\ln d}\right),$$

which agrees with the bound in Kakade et al. (2010) on the dominant term (see also Bartlett and Mendelson 2002 and Meir and Zhang 2003).

Our last bound is useless if $d \geq e^n$ or if d is infinite. But whenever the norm of the data has finite second moments we can use Corollary 3 and inequality (4) to obtain

$$\mathcal{R}_{\mathcal{M}}(\mathbf{X}) \leq \frac{2^{3/2}}{\sqrt{n}}\left(2+\sqrt{\ln \mathbb{E}\|X\|_2^2}\right)+\frac{2}{\sqrt{n}}.$$

For nontrivial results $\mathbb{E}\|X\|^2$ only needs to be subexponential in n.

We remark that a similar condition to Equation (3) for the Lasso, replacing the expectation with the supremum over X, has been considered within the context of elastic net regularization (De Mol et al., 2009).

2.3 The Weighted Lasso

The Lasso assigns an equal penalty to all regression coefficients, while there may be a priori information on the respective significance of the different coordinates. For this reason different weightings have been proposed (see, for example, Shimamura et al. 2007). In our framework an appropriate set of operators is $\mathcal{M} = \{\alpha_1 P_1, \ldots, \alpha_k P_k, \ldots\}$, with $\alpha_k > 0$ where α_k^{-1} is the penalty weight associated with the k-th coordinate. Then

$$\|\beta\|_{\mathcal{M}} = \sum_k \alpha_k^{-1} |\beta_k|$$

and

$$\|z\|_{\mathcal{M}^*} = \sup_k \alpha_k |z_k|.$$

To further illustrate the use of Corollary 3 let us assume that the underlying space H is infinite dimensional (that is, $H = \ell_2(\mathbb{N})$), and make the compensating assumption that $\alpha \in H$, that is $\sum_k \alpha_k^2 = R^2 < \infty$. For simplicity we also assume that $\sup_k \alpha_k \leq 1$. Then, if $\|X\|_\infty \leq 1$ almost surely, we have both $\|X\|_{\mathcal{M}^*} \leq 1$ and $\sum_k \alpha_k^2 (X)_k^2 \leq R^2$. Again we obtain

$$\mathcal{R}_{\mathcal{M}}(\mathbf{X}) \leq \frac{2^{3/2}}{\sqrt{n}}\left(2+\sqrt{\ln R^2}\right)+\frac{2}{\sqrt{n}}.$$

So in this case the second moment bound is enforced by the weighting sequence.

2.4 The Group Lasso

Let $H = \mathbb{R}^d$ and let $\{J_1, \ldots, J_r\}$ be a partition of the index set $\{1, \ldots, d\}$. We take $\mathcal{M} = \{P_{J_1}, \ldots, P_{J_r}\}$ where $P_{J_\ell} = \sum_{i \in J_\ell} P_i$ is the projection onto the subspace spanned by the basis vector e_i. The ranges of the P_{J_ℓ} then provide an orthogonal decomposition of $\mathbb{R}^d$ and the above mentioned simplifications also apply. We get

$$\|\beta\|_{\mathcal{M}} = \sum_{\ell=1}^{r} \|P_{J_\ell}\beta\|$$

and

$$\|z\|_{\mathcal{M}^*} = \max_{\ell=1}^{r} \|P_{J_\ell} z\|.$$

The algorithm which uses $\|\beta\|_{\mathcal{M}}$ as a regularizer is called the group Lasso (see, for example, Yuan and Lin 2006). It encourages vectors β whose support lies the union of a small number of groups J_ℓ

of coordinate indices. If we know that $\|P_{J_\ell}X\| \leq 1$ almost surely for all $\ell \in \{1,\dots,r\}$ then we get

$$\mathcal{R}_{\mathcal{M}}(\mathbf{X}) \leq \frac{2^{3/2}}{\sqrt{n}}\left(2+\sqrt{\ln r}\right), \tag{5}$$

in complete symmetry with the Lasso and essentially the same as given in Kakade et al. (2010). If r is prohibitively large or if different penalties are desired for different groups, the same remarks apply as in the previous two sections. Just as in the case of the Lasso the second moment condition (3) translates to the simple form $\mathbb{E}\|X\|_2^2 < \infty$.

2.5 Overlapping Groups

In the previous examples the members of $\mathcal{M}$ always had mutually orthogonal ranges, which gave a simple appearance to the norm $\|\beta\|_{\mathcal{M}}$. If the ranges are not mutually orthogonal, the norm has a more complicated form. For example, in the group Lasso setting, if the groups J_ℓ cover $\{1,\dots,d\}$, but are not disjoint, we obtain the regularizer of Obozinski et al. (2009), given by

$$\Omega_{\text{overlap}}(\beta) = \inf\left\{\sum_{\ell=1}^{r}\|v_\ell\| : (v_\ell)_{jk} = 0 \text{ if } k \notin J_\ell \text{ and } \sum_{\ell=1}^{r} v_\ell = \beta\right\}.$$

If $\|P_{J_\ell}X_i\| \leq 1$ almost surely for all $\ell \in \{1,\dots,r\}$ then the Rademacher complexity of the set of linear functionals with $\Omega_{\text{overlap}}(\beta) \leq 1$ is bounded as in (5), in complete equivalence to the bound for the group Lasso.

The same bound also holds for the class satisfying $\Omega_{\text{group}}(\beta) \leq 1$, where the function Ω_{group} is defined, for every $\beta \in \mathbb{R}^d$, as

$$\Omega_{\text{group}}(\beta) = \sum_{\ell=1}^{r}\|P_{J_\ell}\beta\|$$

which has been proposed by Jenatton et al. (2011) and Zhao et al. (2009). To see this we only have to show that $\Omega_{\text{overlap}} \leq \Omega_{\text{group}}$ which is accomplished by generating a disjoint partition $\{J'_\ell\}_{\ell=1}^{r}$ where $J'_\ell \subseteq J_\ell$, writing $\beta = \sum_{\ell=1}^{r} P_{J'_\ell}\beta$ and realizing that $\left\|P_{J'_\ell}\beta\right\| \leq \|P_{J_\ell}\beta\|$. The bound obtained from this simple comparison may however be quite loose.

2.6 Regularizers Generated from Cones

Our next example considers structured sparsity regularizers as in Micchelli et al. (2010). Let Λ be a nonempty subset of the open positive orthant in $\mathbb{R}^d$ and define a function $\Omega_\Lambda : \mathbb{R}^d \to \mathbb{R}$ by

$$\Omega_\Lambda(\beta) = \frac{1}{2}\inf_{\lambda\in\Lambda}\sum_{j=1}^{d}\left(\frac{\beta_j^2}{\lambda_j} + \lambda_j\right).$$

If Λ is a convex cone, then it is shown in Micchelli et al. (2011) that Ω_Λ is a norm and that the dual norm is given by

$$\|z\|_{\Lambda^*} = \sup\left\{\left(\sum_{j=1}^{d}\mu_j z_j^2\right)^{1/2} : \mu_j = \lambda/\|\lambda\|_1 \text{ with } \lambda \in \Lambda\right\}.$$

The supremum in this formula is evidently attained on the set $\mathcal{E}(\Lambda)$ of extreme points of the closure of $\{\lambda/\|\lambda\|_1 : \lambda \in \Lambda\}$. For $\mu \in \mathcal{E}(\Lambda)$ let M_μ be the diagonal matrix whose diagonal entries are those of the vector μ_j and let $\mathcal{M}_\Lambda$ be the collection of matrices $\mathcal{M}_\Lambda = \left\{M_\mu : \mu \in \mathcal{E}(\Lambda)\right\}$. Then

$$\|z\|_{\Lambda^*} = \sup_{M \in \mathcal{M}_\Lambda} \|Mz\|.$$

Clearly $\mathcal{M}_\Lambda$ is uniformly bounded in the operator norm, so if Λ is a cone and $\mathcal{E}(\Lambda)$ is at most countable, then $\|\cdot\|_{\Lambda^*} = \|\cdot\|_{\mathcal{M}^*}$, $\Omega_\Lambda = \|\cdot\|_{\mathcal{M}^*}$ and our bounds apply. If $\mathcal{E}(\Lambda)$ is finite and $\mathbf{x}$ is a sample then the Rademacher complexity of the class with $\Omega_\Lambda(\beta) \leq 1$ is bounded by

$$\frac{2^{3/2}}{n}\sqrt{\sum_{i=1}^{n}\|x_i\|_{\Lambda^*}^2}\left(2+\sqrt{\ln|\mathcal{E}(\Lambda)|}\right).$$

2.7 Kernel Learning

This is the most general case to which the simplification applies: Suppose that H is the direct sum $H = \oplus_{j\in\mathcal{J}} H_j$ of an at most countable number of Hilbert spaces H_j. We set $\mathcal{M} = \left\{P_j\right\}_{j\in\mathcal{J}}$, where $P_j : H \to H$ is the projection on H_j. Then

$$\|\beta\|_{\mathcal{M}} = \sum_{j\in\mathcal{J}} \left\|P_j\beta\right\|$$

and

$$\|z\|_{\mathcal{M}^*} = \sup_{j\in\mathcal{J}} \left\|P_j z\right\|.$$

Such a situation arises in multiple kernel learning (Bach et al., 2004; Lanckriet et al., 2004) or the nonparametric group Lasso (Meier et al., 2009) in the following way: One has an input space $\mathcal{X}$ and a collection $\left\{K_j\right\}_{j\in\mathcal{J}}$ of positive definite kernels $K_j : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. Let $\phi_j : \mathcal{X} \to H_j$ be the feature map representation associated with kernel K_j, so that, for every $x,t \in \mathcal{X}$ $K_j(x,t) = \langle\phi_j(x),\phi_j(t)\rangle$ (for background on kernel methods see, for example, Shawe-Taylor and Cristianini 2004).

Suppose that $\mathbf{x} = (x_1,\ldots,x_n) \in \mathcal{X}^n$ is a sample. Define the kernel matrix $\mathbf{K}_j = (K_j(x_i,x_k))_{i,k=1}^n$. Using this notation the bound in Theorem 2 reads

$$\mathcal{R}\left((\phi(x_1),\ldots,\phi(x_n))\right) \leq \frac{2^{3/2}}{n}\sqrt{\sup_{j\in\mathcal{J}}\mathrm{tr}\mathbf{K}_j}\left(2+\sqrt{\ln\frac{\sum_{j\in\mathcal{J}}\mathrm{tr}\mathbf{K}_j}{\sup_{j\in\mathcal{J}}\mathrm{tr}\mathbf{K}_j}}\right).$$

In particular, if $\mathcal{J}$ is finite and $K_j(x,x) \leq 1$ for every $x \in \mathcal{X}$ and $j \in \mathcal{J}$, then the the bound reduces to

$$\frac{2^{3/2}}{\sqrt{n}}\left(2+\sqrt{\ln|\mathcal{J}|}\right),$$

essentially in agreement with Cortes et al. (2010), Kakade et al. (2010) and Ying and Campbell (2009). Our leading constant of $2\sqrt{2}$ is slightly better than the constant of $2\sqrt{\frac{23}{22}e}$, given by Cortes et al. (2010).

For infinite or prohibitively large $\mathcal{J}$ the second moment condition now becomes

$$\mathbb{E}\sum_{j\in\mathcal{J}} K_j(X,X) < \infty.$$

We conclude this section by noting that, for every set $\mathcal{M}$ we may choose a set of kernels such that empirical risk minimization with the norm $\|\cdot\|_{\mathcal{M}}$ is equivalent to multiple kernel learning with kernels $K_M(x,t)=\langle Mx,Mt\rangle$, $M\in\mathcal{M}$. To see this, choose, for every $M\in\mathcal{M}$, $\phi_M(x)=Mx$. Note however, that this may yield an overparameterization of the problem. For example, the regularizers in Section 2.6 can be reformulated as a multiple kernel learning problem, but this requires $d|\mathcal{E}(\Lambda)|$ parameters instead of d.

3. Proofs

We first give some notation and auxiliary results, then we prove the results announced in the introduction.

3.1 Notation and Auxiliary Results

The Hilbert space H and the collection M are fixed throughout the following, as is the sample size $n\in\mathbb{N}$.

Recall that $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the norm and inner product in H, respectively. For a linear transformation $M:\mathbb{R}^n\to H$ the Hilbert-Schmidt norm is defined as

$$\|M\|_{HS}=\left(\sum_{i=1}^n\|Me_i\|^2\right)^{1/2}$$

where $\{e_i:i\in\mathbb{N}\}$ is the canonical basis of $\mathbb{R}^n$.

We use bold letters $(\mathbf{x},\ \mathbf{X},\ \boldsymbol{\varepsilon},\ \dots)$ to denote n-tuples of objects, such as vectors or random variables.

Let $\mathcal{X}$ be any space. For $\mathbf{x}=(x_1,\dots,x_n)\in\mathcal{X}^n$, $1\le k\le n$ and $y\in\mathcal{X}$ we use $\mathbf{x}_{k\leftarrow y}$ to denote the object obtained from $\mathbf{x}$ by replacing the k-th coordinate of $\mathbf{x}$ with y. That is

$$\mathbf{x}_{k\leftarrow y}=(x_1,\dots,x_{k-1},y,x_{k+1},\dots,x_n).$$

The following concentration inequality, known as the bounded difference inequality (see McDiarmid 1998), goes back to the work of Hoeffding (1963). We only need it in the weak form stated below.

Theorem 4 *Let $F:\mathcal{X}^n\to\mathbb{R}$ and write*

$$B^2=\sum_{k=1}^n\sup_{y_1,y_2\in\mathcal{X},\,\mathbf{x}\in\mathcal{X}^n}\left(F\left(\mathbf{x}_{k\leftarrow y_1}\right)-F\left(\mathbf{x}_{k\leftarrow y_2}\right)\right)^2.$$

Let $\mathbf{X}=(X_1,\dots,X_n)$ be a vector of independent random variables with values in $\mathcal{X}$, and let $\mathbf{X}'$ be iid to $\mathbf{X}$. Then for any $t>0$

$$\Pr\left\{F\left(\mathbf{X}\right)>\mathbb{E}F\left(\mathbf{X}'\right)+t\right\}\le e^{-2t^2/B^2}.$$

Finally we need a simple lemma on the normal approximation:

Lemma 5 *Let $a,\delta>0$. Then*

$$\int_\delta^\infty\exp\left(\frac{-t^2}{2a^2}\right)dt\le\frac{a^2}{\delta}\exp\left(\frac{-\delta^2}{2a^2}\right).$$

Proof For $t \geq \delta/a$ we have $1 \leq at/\delta$. Thus

$$\int_{\delta}^{\infty} \exp\left(\frac{-t^2}{2a^2}\right) dt = a \int_{\delta/a}^{\infty} e^{-t^2/2} dt \leq \frac{a^2}{\delta} \int_{\delta/a}^{\infty} t e^{-t^2/2} dt = \frac{a^2}{\delta} \exp\left(\frac{-\delta^2}{2a^2}\right).$$

■

3.2 Properties of the Regularizer

In this section, we show that the regularizer in Equation (1) is indeed a norm and we derive the associated dual norm. In parallel we treat an entire class of regularizers, which relates to $\|\cdot\|_{\mathcal{M}}$ as the ℓ_q-norm relates to the ℓ_1-norm. To this end, we fix an exponent $q \in [1, \infty]$. The conjugate exponent is denoted p, with $1/q + 1/p = 1$.

Recall that $|||\cdot|||$ denotes the operator norm. We first state the general conditions on the set $\mathcal{M}$ of operators.

Condition 6 *$\mathcal{M}$ is an at most countable set of symmetric bounded linear operators on a real separable Hilbert space H such that*

(a) *For every $x \in H$ with $x \neq 0$, there exists $M \in \mathcal{M}$ such that $Mx \neq 0$*

(b) $\sup_{M \in \mathcal{M}} |||M||| < \infty$ *if $q = 1$ and* $\sum_{M \in \mathcal{M}} |||M|||^p < \infty$ *if $q > 1$.*

Now we define $\ell_q(\mathcal{M})$ to be the set of those vectors $\beta \in H$ for which the quantity

$$\|\beta\|_{\mathcal{M}_q} = \inf\left\{ \left(\sum_{M \in \mathcal{M}} \|v_M\|^q \right)^{1/q} : v_M \in H \text{ and } \sum_{M \in \mathcal{M}} M v_M = \beta \right\}$$

is finite. If $q = 1$ we drop the subscript in $\|\cdot\|_{\mathcal{M}_q}$ to lighten notation. Observe that the case $q = 1$ coincides with the definition given in the introduction.

Theorem 7 *$\ell_q(\mathcal{M})$ is a Banach space with norm $\|\cdot\|_{\mathcal{M}_q}$, and $\ell_q(\mathcal{M})$ is dense in H. If $\mathcal{M}$ is finite or H is finite-dimensional, then $\ell_q(\mathcal{M}) = H$. For $z \in H$ the norm of the linear functional $\beta \in \ell_q(\mathcal{M}) \mapsto \langle \beta, z \rangle$ is*

$$\|z\|_{\mathcal{M}_{q^*}} = \begin{cases} \sup_{M \in \mathcal{M}} \|Mz\|, & \text{if } q = 1, \\ \left(\sum_{M \in \mathcal{M}} \|Mz\|^p \right)^{1/p}, & \text{if } q > 1. \end{cases}$$

Proof Let $\mathcal{V}_q(\mathcal{M}) = \{v : v = (v_M)_{M \in \mathcal{M}},\ v_M \in H\}$ be the set of those H-valued sequences indexed by $\mathcal{M}$, for which the function

$$v \mapsto \|v\|_{\mathcal{V}_q(\mathcal{M})} = \left(\sum_{M \in \mathcal{M}} \|v_M\|^q \right)^{1/q}$$

is finite. Then $\|\cdot\|_{\mathcal{V}_q(\mathcal{M})}$ defines a complete norm on $\mathcal{V}_q(\mathcal{M})$, making $\mathcal{V}_q(\mathcal{M})$ a Banach space. If $w = (w_M)_{M\in\mathcal{M}}$ is an H-valued sequence indexed by $\mathcal{M}$, then the linear functional

$$v \in \mathcal{V}_q(\mathcal{M}) \mapsto \sum_{M\in\mathcal{M}} \langle v_M, w_M\rangle$$

has norm

$$\|w\|_{\mathcal{V}_q(\mathcal{M})^*} = \begin{cases} \sup_{M\in\mathcal{M}} \|Mw_M\|, & \text{if } q = 1, \\ \left(\sum_{M\in\mathcal{M}} \|v_M\|^p\right)^{1/p}, & \text{if } q > 1. \end{cases}$$

The verification of these claims parallels that of the standard results on Lebesgue spaces.

Now define a map

$$A : v \in \mathcal{V}_q(\mathcal{M}) \mapsto \sum_{M\in\mathcal{M}} Mv_M \in H.$$

We have

$$\|Av\| \leq \sum_{M\in\mathcal{M}} |||M|||\,\|v_M\|.$$

By Condition 6(b) and Hölder's inequality A is a bounded linear transformation whose kernel $\mathcal{K}$ is therefore closed, making the quotient space $\mathcal{V}_q(\mathcal{M})/\mathcal{K}$ into a Banach space with quotient norm $\|w+\mathcal{K}\|_Q = \inf\left\{\|v\|_{\mathcal{V}_q(\mathcal{M})} : w - v \in \mathcal{K}\right\}$. The map A induces an isomorphism

$$\hat{A} : w + \mathcal{K} \in \mathcal{V}_q(\mathcal{M})/\mathcal{K} \mapsto Aw \in H.$$

The range of $\hat{A}$ is $\ell_q(\mathcal{M})$ and becomes a Banach space with the norm $\left\|\hat{A}^{-1}(\beta)\right\|_Q$. But

$$\begin{aligned} \left\|\hat{A}^{-1}(\beta)\right\|_Q &= \inf\left\{\|v\|_{\mathcal{V}_q(\mathcal{M})} : \hat{A}^{-1}(\beta) - v \in \mathcal{K}\right\} \\ &= \inf\left\{\|v\|_{\mathcal{V}_q(\mathcal{M})} : \beta = Av\right\} = \|\beta\|_{\mathcal{M}_q}, \end{aligned}$$

so $\|.\|_{\mathcal{M}_q}$ is a norm making $\ell_q(\mathcal{M})$ into a Banach space.

Suppose that $w \in H$ is orthogonal to $\ell_q(\mathcal{M})$. Let $M_0 \in \mathcal{M}$ be arbitrary and define $v = (v_M)$ by $v_{M_0} = M_0 w$ and $v_M = 0$ for all other M. Then

$$0 = \langle w, Av\rangle = \left\langle w, M_0^2 w\right\rangle = \|M_0 w\|^2,$$

so $M_0 = 0$. This holds for any $M_0 \in \mathcal{M}$, so Condition 6(a) implies that $w = 0$. By the Hahn Banach Theorem $\ell_q(\mathcal{M})$ is therefore dense in H. If $\mathcal{M}$ is finite or H is finite-dimensional, then $\ell_q(\mathcal{M})$ is also finite-dimensional and closed and thus $\ell_q(\mathcal{M}) = H$.

For the last assertion let $z \in H$. Then

$$\begin{aligned} \|z\|_{\mathcal{M}_q^*} &= \sup\left\{\langle z, \beta\rangle : \|\beta\|_{\mathcal{M}_q} \leq 1\right\} \\ &= \sup\left\{\langle z, Av\rangle : \|v\|_{\mathcal{V}_q(\mathcal{M})} \leq 1\right\} \\ &= \sup\left\{\langle A^* z, v\rangle : \|v\|_{\mathcal{V}_q(\mathcal{M})} \leq 1\right\} \\ &= \|A^* z\|_{\mathcal{V}_q(\mathcal{M})^*} \\ &= \sup_{M\in\mathcal{M}} \|Mz\| \text{ if } q = 1 \text{ or } \left(\sum_{M\in\mathcal{M}} \|Mz\|^p\right)^{1/p} \text{ if } q > 1. \end{aligned}$$

■

Proposition 8 *If the ranges of the members of $\mathcal{M}$ are mutually orthogonal then for $\beta \in \ell_1(\mathcal{M})$*

$$\|\beta\|_{\mathcal{M}} = \sum_{M \in \mathcal{M}} \left\|M^+\beta\right\|,$$

where M^+ is the pseudoinverse of M.

Proof The ranges of the members of $\mathcal{M}$ provide an orthogonal decomposition of H, so

$$\beta = \sum_{M \in \mathcal{M}} M\left(M^+\beta\right),$$

where we used the fact that MM^+ is the orthogonal projection onto the range of M. Taking $v_M = M^+\beta$ this implies that $\|\beta\|_{\mathcal{M}} \leq \sum_{M \in \mathcal{M}} \|M^+\beta\|$. On the other hand, if $\beta = \sum_{N \in \mathcal{M}} N v_N$, then, applying M^+ to this identity we see that $M^+ M v_M = M^+\beta$ for all M, so

$$\sum_{M \in \mathcal{M}} \|v_M\| \geq \sum_{M \in \mathcal{M}} \left\|M^+ M v_M\right\| = \sum_{M \in \mathcal{M}} \left\|M^+\beta\right\|,$$

which shows the reverse inequality. ■

3.3 Bounds for the $\ell_1(\mathcal{M})$-Norm Regularizer

We use the bounded difference inequality to derive a concentration inequality for linearly transformed random vectors.

Lemma 9 *Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ be a vector of independent real random variables with $-1 \leq \varepsilon_i \leq 1$, and ε' iid to ε. Suppose that M is a linear transformation $M : \mathbb{R}^n \to H$.*

(i) *Then for $t > 0$ we have*

$$\Pr\left\{\|M\varepsilon\| \geq \mathbb{E}\left\|M\varepsilon'\right\| + t\right\} \leq \exp\left(\frac{-t^2}{2\,\|M\|_{HS}^2}\right).$$

(ii) *If ε is orthonormal (satisfying $\mathbb{E}\varepsilon_i\varepsilon_j = \delta_{ij}$), then*

$$\mathbb{E}\,\|M\varepsilon\| \leq \|M\|_{HS}. \tag{6}$$

and, for every $r > 0$,

$$\Pr\{\|M\varepsilon\| > t\} \leq e^{1/r} \exp\left(\frac{-t^2}{(2+r)\,\|M\|_{HS}^2}\right).$$

Proof (i) Define $F:[-1,1]^n\to\mathbb{R}$ by $F(\mathbf{x})=\|M\mathbf{x}\|$. By the triangle inequality

$$\begin{aligned}
&\sum_{k=1}^{n}\sup_{y_1,y_2\in[-1,1],\,\mathbf{x}\in[-1,1]^n}\left(F\left(\mathbf{x}_{k\leftarrow y_1}\right)-F\left(\mathbf{x}_{k\leftarrow y_2}\right)\right)^2\\
\leq\;&\sum_{k=1}^{n}\sup_{y_1,y_2\in[-1,1],\,\mathbf{x}\in[-1,1]^n}\left\|M\left(\mathbf{x}_{k\leftarrow y_1}-\mathbf{x}_{k\leftarrow y_2}\right)\right\|^2\\
=\;&\sum_{k=1}^{n}\sup_{y_1,y_2\in[-1,1]}\left(y_1-y_2\right)^2\left\|Me_k\right\|^2\\
\leq\;&4\left\|M\right\|_{HS}^2.
\end{aligned}$$

The result now follows from the bounded difference inequality (Theorem 4).

(ii) If ε is orthonormal then it follows from Jensen's inequality that

$$\mathbb{E}\left\|M\varepsilon\right\|\leq\left(\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i Me_i\right\|^2\right)^{1/2}=\left(\sum_{i=1}^{n}\left\|Me_i\right\|^2\right)^{1/2}=\left\|M\right\|_{HS}.$$

For the second assertion of (ii) first note that from calculus we get $(t-1)^2/2-t^2/(2+r)\geq-1/r$ for all $t\in\mathbb{R}$. This implies that

$$e^{-(t-1)^2/2}\leq e^{1/r}e^{-t^2/(2+r)}. \tag{7}$$

Since $1/r\geq1/(2+r)$ the inequality to be proved is trivial for $t\leq\|M\|_{HS}$. If $t>\|M\|_{HS}$ then, using $\mathbb{E}\|M\varepsilon\|\leq\|M\|_{HS}$, we have $t-E\|M\varepsilon\|\geq t-\|M\|_{HS}>0$, so by part (i) and (7) we obtain

$$\begin{aligned}
\Pr\{\|M\varepsilon\|\geq t\}&=\Pr\{\|M\varepsilon\|\geq E\|M\varepsilon\|+(t-E\|M\varepsilon\|)\}\\
&\leq\exp\left(\frac{-(t-E\|M\varepsilon\|)^2}{2\|M\|_{HS}^2}\right)\leq\exp\left(\frac{-(t-\|M\|_{HS})^2}{2\|M\|_{HS}^2}\right)\\
&=\exp\left(\frac{-(t/\|M\|_{HS}-1)^2}{2}\right)\leq e^{1/r}e^{-(t/\|M\|_{HS})^2/(2+r)}\\
&=e^{1/r}\exp\left(\frac{-t^2}{(2+r)\|M\|_{HS}^2}\right).
\end{aligned}$$

■

We now use integration by parts, a union bound and the above concentration inequality to derive a bound on the expectation of the supremum of the norms $\|M\varepsilon\|$. This is the essential step in the proof of Theorem 2. It is by no means a new technique, in fact it appears many times in the book by Ledoux and Talagrand (1991), but compared to the combinatorial approach by Cortes et al. (2010) it seems more suited to the study of the problem at hand, and gives insights into the fine structure of the logarithmic factor appearing in bounds for Lasso-like methods.

Lemma 10 *Let $\mathcal{M}$ be an at most countable set of linear transformations $M : \mathbb{R}^n \to H$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ a vector of orthonormal random variables (satisfying $\mathbb{E}\varepsilon_i\varepsilon_j = \delta_{ij}$) with values in $[-1,1]$. Then*

$$\mathbb{E} \sup_{M\in\mathcal{M}} \|M\varepsilon\| \le \sqrt{2} \sup_{M\in\mathcal{M}} \|M\|_{HS} \left(2 + \sqrt{\ln \frac{\sum_{M\in\mathcal{M}} \|M\|_{HS}^2}{\sup_{M\in\mathcal{M}} \|M\|_{HS}^2}}\right).$$

Proof To lighten notation we abbreviate $\mathcal{M}_\infty := \sup_{M\in\mathcal{M}} \|M\|_{HS}$ below. We now use integration by parts

$$\begin{aligned}
\mathbb{E} \sup_{M\in\mathcal{M}} \|M\varepsilon\| &= \int_0^\infty \Pr\left\{\sup_{M\in\mathcal{M}} \|M\varepsilon\| > t\right\} dt \\
&\le \mathcal{M}_\infty + \delta + \int_{\mathcal{M}_\infty+\delta}^\infty \Pr\left\{\sup_{M\in\mathcal{M}} \|M\varepsilon\| > t\right\} dt \\
&\le \mathcal{M}_\infty + \delta + \sum_{M\in\mathcal{M}} \int_{\mathcal{M}_\infty+\delta}^\infty \Pr\{\|M\varepsilon\| > t\}\, dt,
\end{aligned}$$

where we have introduced a parameter $\delta \ge 0$. The first inequality above follows from the fact that probabilities never exceed 1, and the second from a union bound. Now for any $M \in \mathcal{M}$ we can make a change of variables and use (6), which gives $\mathbb{E}\|M\varepsilon\| \le \|M\|_{HS} \le \mathcal{M}_\infty$, so that

$$\begin{aligned}
\int_{\mathcal{M}_\infty+\delta}^\infty \Pr\{\|M\varepsilon\| > t\}\, dt &\le \int_\delta^\infty \Pr\{\|M\varepsilon\| > \mathbb{E}\|M\varepsilon\| + t\}\, dt \\
&\le \int_\delta^\infty \exp\left(\frac{-t^2}{2\|M\|_{HS}^2}\right) dt \\
&\le \frac{\|M\|_{HS}^2}{\delta} \exp\left(\frac{-\delta^2}{2\|M\|_{HS}^2}\right),
\end{aligned}$$

where the second inequality follows from Lemma 9-(i), and the third from Lemma 5. Substitution in the previous chain of inequalities and using Hoelder's inequality (in the ℓ_1/ℓ_∞-version) give

$$\mathbb{E} \sup_{M\in\mathcal{M}} \|M\varepsilon\| \le \mathcal{M}_\infty + \delta + \frac{1}{\delta}\left(\sum_{M\in\mathcal{M}} \|M\|_{HS}^2\right) \exp\left(\frac{-\delta^2}{2\mathcal{M}_\infty^2}\right). \tag{8}$$

We now set

$$\delta = \mathcal{M}_\infty \sqrt{2 \ln\left(e \frac{\sum_{M\in\mathcal{M}} \|M\|_{HS}^2}{\mathcal{M}_\infty^2}\right)}.$$

Then $\delta \ge 0$ as required. The substitution makes the last term in (8) smaller than $\mathcal{M}_\infty / \left(e\sqrt{2}\right)$, and since $1 + 1/\left(e\sqrt{2}\right) < \sqrt{2}$, we obtain

$$\mathbb{E} \sup_{M\in\mathcal{M}} \|M\varepsilon\| \le \sqrt{2}\mathcal{M}_\infty \left(1 + \sqrt{\ln\left(\frac{e\sum_{M\in\mathcal{M}} \|M\|_{HS}^2}{\mathcal{M}_\infty^2}\right)}\right).$$

Finally we use $\sqrt{\ln es} \leq 1 + \sqrt{\ln s}$ for $s \geq 1$. ■

Proof of Theorem 2 Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ be a vector of iid Rademacher variables. For $M \in \mathcal{M}$ we use $M\mathbf{x}$ to denote the linear transformation $M\mathbf{x} : \mathbb{R}^n \to H$ given by $(M\mathbf{x})\,\mathbf{y} = \sum_{i=1}^n (Mx_i)\, y_i$. We have

$$\mathcal{R}_{\mathcal{M}}(\mathbf{x}) = \frac{2}{n}\mathbb{E} \sup_{\beta : \|\beta\|_{\mathcal{M}} \leq 1} \left\langle \beta, \sum_{i=1}^n \varepsilon_i x_i \right\rangle \leq \frac{2}{n}\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_{\mathcal{M}^*} = \frac{2}{n}\mathbb{E} \sup_{M \in \mathcal{M}} \|M\mathbf{x}\varepsilon\|.$$

Applying Lemma 10 to the set of transformations $\mathcal{M}\mathbf{x} = \{M\mathbf{x} : M \in \mathcal{M}\}$ gives

$$\mathcal{R}_{\mathcal{M}}(\mathbf{x}) \leq \frac{2^{3/2} \sup_{M \in \mathcal{M}} \|M\mathbf{x}\|_{HS}}{n} \left(2 + \sqrt{\ln \frac{\sum_{M \in \mathcal{M}} \|M\mathbf{x}\|_{HS}^2}{\sup_{M \in \mathcal{M}} \|M\mathbf{x}\|_{HS}^2}} \right).$$

Substitution of $\|M\mathbf{x}\|_{HS}^2 = \sum_{i=1}^n \|Mx_i\|^2$ gives the first inequality of Theorem 2 and

$$\sup_{M \in \mathcal{M}} \|M\mathbf{x}\|_{HS}^2 \leq \sum_{i=1}^n \sup_{M \in \mathcal{M}} \|Mx_i\|^2 = \sum_{i=1}^n \|x_i\|_{\mathcal{M}^*}^2$$

gives the second inequality. ■

Proof of Corollary 3 From calculus we find that $t \ln t \geq -1/e$ for all $t > 0$. For $A, B > 0$ and $n \in \mathbb{N}$ this implies that

$$A \ln \frac{B}{A} = n\left[(A/n)\ln(B/n) - (A/n)\ln(A/n)\right] \leq A \ln(B/n) + n/e. \tag{9}$$

Now multiply out the first inequality of Theorem 2 and use (9) with

$$A = \sup_{M \in \mathcal{M}} \sum_{i=1}^n \|Mx_i\|^2 \text{ and } B = \sum_{M \in \mathcal{M}} \sum_{i=1}^n \|Mx_i\|^2.$$

Finally use $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ and the fact that $2^{3/2}/\sqrt{e} \leq 2$. ■

4. The $\ell_q(\mathcal{M})$ Case

In this section we give bounds for the $\ell_q(\mathcal{M})$-norm regularizers, with $q > 1$.

We give two results, which can be applied to cases analogous to those in Section 2. The first result is essentially equivalent to Cortes et al. (2010), Kakade et al. (2010) and Kloft et al. (2011) and is presented for completeness. The second result is not dimension free, but it approaches the bound in Theorem 2 for arbitrarily large dimensions. The proofs are analogous to the proof of Theorem 2.

Theorem 11 *Let* $\mathbf{x}$ *be a sample and* $\mathcal{R}_{\mathcal{M}_q}(\mathbf{x})$ *the empirical Rademacher complexity of the class of linear functions parameterized by* β *with* $\|\beta\|_{\mathcal{M}_q} \leq 1$*. Then for* $1 < q \leq 2$

$$\mathcal{R}_{\mathcal{M}_q}(\mathbf{x}) \leq \frac{2}{n}\left(1 + \left(\frac{\pi}{2}\right)^{\frac{1}{2p}} \sqrt{p}\right) \sqrt{\sum_{i=1}^n \|x_i\|_{\mathcal{M}_q^*}^2}.$$

The proof is based on the following

Lemma 12 *Let $\mathcal{M}$ be an at most countable set of linear transformations $M : \mathbb{R}^n \to H$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ a vector of orthonormal random variables (satisfying $\mathbb{E}\varepsilon_i\varepsilon_j = \delta_{ij}$) with values in $[-1,1]$. Then for $p \geq 2$*

$$\mathbb{E}\left[\left(\sum_{M\in\mathcal{M}}\|M\varepsilon\|^p\right)^{1/p}\right] \leq \left(1+\left(\frac{\pi}{2}\right)^{\frac{1}{2p}}\sqrt{p}\right)\left(\sum_{M\in\mathcal{M}}\|M\|_{HS}^p\right)^{1/p}.$$

Proof We first note, by Jensen inequality, that

$$\mathbb{E}\left[\left(\sum_{M\in\mathcal{M}}\|M\varepsilon\|^p\right)^{1/p}\right] \leq \left(\mathbb{E}\left[\sum_{M\in\mathcal{M}}\|M\varepsilon\|^p\right]\right)^{1/p}. \tag{10}$$

We rewrite the expectation appearing in the right hand side using integration by parts and a change of variable as

$$\mathbb{E}\left[\|M\varepsilon\|^p\right] = \int_0^\infty \Pr\left\{\|M\varepsilon\|^p > t\right\} dt = A^p + p\int_0^\infty \Pr\left\{\|M\varepsilon\|^p > s^p + A^p\right\} s^{p-1} ds \tag{11}$$

where $A \geq 0$. Next, we use convexity of the function $x \mapsto x^p$, $x \geq 0$, which gives for $\lambda \in (0,1)$

$$\left(\lambda^{\frac{p-1}{p}} s + (1-\lambda)^{\frac{p-1}{p}} A\right)^p \leq \lambda\left(\frac{s}{\lambda^{1/p}}\right)^p + (1-\lambda)\left(\frac{A}{(1-\lambda)^{1/p}}\right)^p = s^p + A^p.$$

This allows us to bound

$$\begin{aligned}\Pr\left\{\|M\varepsilon\|^p > s^p + A^p\right\} &\leq \Pr\left\{\|M\varepsilon\|^p > \left(\lambda^{\frac{p-1}{p}} s + (1-\lambda)^{\frac{p-1}{p}} A\right)^p\right\} \\ &= \Pr\left\{\|M\varepsilon\| > \lambda^{\frac{p-1}{p}} s + (1-\lambda)^{\frac{p-1}{p}} A\right\}.\end{aligned} \tag{12}$$

Combining Equations (11) and (12), choosing $A = (1-\lambda)^{\frac{1-p}{p}}\|M\|_{HS}$ and making the change of variable $t = \lambda^{\frac{p-1}{p}} s$, gives

$$\begin{aligned}\int_0^\infty \Pr\left\{\|M\varepsilon\|^p > t\right\} dt &\leq (1-\lambda)^{1-p}\|M\|_{HS}^p + p\lambda^{1-p}\int_0^\infty \Pr\left\{\|M\varepsilon\| > \|M\|_{HS} + t\right\} t^{p-1} dt \\ &\leq (1-\lambda)^{1-p}\|M\|_{HS}^p + p\lambda^{1-p}\int_0^\infty t^{1-p} e^{-t^2/2\|M\|_{HS}^2} dt \\ &\leq (1-\lambda)^{1-p}\|M\|_{HS}^p + p\lambda^{1-p}\|M\|_{HS}^p\sqrt{\frac{\pi}{2}}p^{p/2-1} \\ &= \|M\|_{HS}^p\left((1-\lambda)^{1-p} + \lambda^{1-p}\sqrt{\frac{\pi}{2}}p^{p/2}\right)\end{aligned}$$

where the second inequality follows by Lemma 9-(i) and the third inequality follows by a standard result on the moments of the normal distribution, namely

$$\int_0^\infty t^{p-1}\exp\left(\frac{-t^2}{2}\right) dt \leq \sqrt{\frac{\pi}{2}}(p-2)!! \leq \sqrt{\frac{\pi}{2}}(1\cdot 3\cdot\ldots\cdot p-2) \leq \sqrt{\frac{\pi}{2}}p^{p/2-1}.$$

Summing both sides of Equation (13) over M we obtain that

$$\mathbb{E}\left[\sum_{M\in\mathcal{M}}\|M\varepsilon\|^{p}\right]\leq\sum_{M}\|M\|_{HS}^{p}\left((1-\lambda)^{1-p}+\lambda^{1-p}\sqrt{\frac{\pi}{2}}p^{p/2}\right).$$

A direct computation gives that the right hand side of the above equation attains its minimum at

$$\lambda=\frac{\left(\frac{\pi}{2}\right)^{\frac{1}{2p}}p^{\frac{1}{2}}}{1+\left(\frac{\pi}{2}\right)^{\frac{1}{2p}}p^{\frac{1}{2}}}.$$

The result now follows by Equation (10). ■

Proof of Theorem 11 Let $\alpha=1+\left(\frac{\pi}{2}\right)^{\frac{1}{2p}}\sqrt{p}$. As in the proof of Theorem 2 we proceed using duality and apply Lemma 12 to the set of transformations $\mathcal{M}\mathbf{x}=\left\{M\mathbf{x}:M\in\mathcal{M}\right\}$,

$$\begin{aligned}
\mathcal{R}_{\mathcal{M}_q}(\mathbf{x}) &\leq \frac{2}{n}\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i x_i\right\|_{\mathcal{M}_q^*}=\frac{2}{n}\mathbb{E}\left[\left(\sum_{M\in\mathcal{M}}\|M\mathbf{x}\varepsilon\|^{p}\right)^{1/p}\right]\\
&\leq \frac{2\alpha}{n}\left(\sum_{M\in\mathcal{M}}\|M\mathbf{x}\|_{HS}^{p}\right)^{1/p}=\frac{2\alpha}{n}\sqrt{\left(\sum_{M\in\mathcal{M}}\left(\sum_{i=1}^{n}\|Mx_i\|^{2}\right)^{p/2}\right)^{2/p}}\\
&\leq \frac{2\alpha}{n}\sqrt{\sum_{i=1}^{n}\left(\sum_{M\in\mathcal{M}}\left(\|Mx_i\|^{2}\right)^{p/2}\right)^{2/p}}=\frac{2\alpha}{n}\sqrt{\sum_{i=1}^{n}\|x_i\|_{\mathcal{M}_{q^*}}^{2}},
\end{aligned}$$

where the last inequality is just the triangle inequality in $\ell_{p/2}$. ■

One can verify that the leading constant in our bound is smaller than the one in Cortes et al. (2010) for $p>12$. Note that the bound in Theorem 11 diverges for q going to 1 since in this case p grows to infinity.

We conclude this section with a result, which shows that the bound in Theorem 2 has a stability property in the following sense: If $\mathcal{M}$ is finite, then we can give a bound on the Rademacher complexity of the unit ball in $\ell_q(\mathcal{M})$ which converges to the bound in Theorem 2 as $q\to1$, regardless of the size of $\mathcal{M}$. Only the rate of convergence is dimension dependent.

Theorem 13 *Under the conditions of Theorem 11*

$$\mathcal{R}_{\mathcal{M}_q}(\mathbf{x})\leq\frac{4\left|\mathcal{M}\right|^{1/p}}{n}\sqrt{\sup_{M\in\mathcal{M}}\sum_{i}\|Mx_i\|^{2}}\left(2+\sqrt{\ln\sum_{M}\frac{\sum_i\|Mx_i\|^{2}}{\sup_{N\in\mathcal{M}}\sum_i\|Nx_i\|^{2}}}\right).$$

So, as q goes to 1, $p\to\infty$ and we recover the bound in Theorem 2 up to a small multiplicative constant. The key step in the proof of Theorem 13 is the following

Lemma 14 *Let $\mathcal{M}$ be a finite set of linear transformations $M : \mathbb{R}^n \to H$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ a vector of orthonormal random variables with values in $[-1,1]$. Then*

$$\mathbb{E}\left[\left(\sum_M \|M\varepsilon\|^p\right)^{1/p}\right] \leq 2\left|\mathcal{M}\right|^{1/p} \sup_{M\in\mathcal{M}} \|M\|_{HS}\left(2+\sqrt{\ln\frac{\sum_M \|M\|_{HS}^2}{\sup_{N\in\mathcal{M}}\|N\|_{HS}^2}}\right).$$

Proof If $t \geq 0$ and $\sum_M \|M\varepsilon\|^p > t^p$, then there must exist some $M \in \mathcal{M}$ such that $\|M\varepsilon\|^p > t^p/\left|\mathcal{M}\right|$, which in turn implies that $\|M\varepsilon\| > t/\left|\mathcal{M}\right|^{1/p}$. It then follows from a union bound that

$$\Pr\left\{\sum_M \|M\varepsilon\|^p > t^p\right\} \leq \sum_M \Pr\left\{\|M\varepsilon\| > t/\left|\mathcal{M}\right|^{1/p}\right\} \leq \exp\left(\frac{-t^2}{4\left|\mathcal{M}\right|^{2/p}\|M\|_{HS}^2}\right),$$

where we used the subgaussian concentration inequality Lemma 9-(ii) with $r = 2$. Using integration by parts we have with $\delta \geq 0$ that

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_M \|M\varepsilon\|^p\right)^{1/p}\right] &\leq \delta + \int_\delta^\infty \Pr\left\{\sum_M \|M\varepsilon\|^p > t^p\right\} dt \\
&\leq \delta + 2\sum_M \int_\delta^\infty \exp\left(\frac{-t^2}{4\left|\mathcal{M}\right|^{2/p}\|M\|_{HS}^2}\right) dt \\
&\leq \delta + \frac{4\left|\mathcal{M}\right|^{2/p}}{\delta}\sum_M \|M\|_{HS}^2 \exp\left(\frac{-\delta^2}{4\left|\mathcal{M}\right|^{2/p}\|M\|_{HS}^2}\right) \\
&\leq \delta + \frac{4\left|\mathcal{M}\right|^{2/p}}{\delta}\left(\sum_M \|M\|_{HS}^2\right)\exp\left(\frac{-\delta^2}{4\left|\mathcal{M}\right|^{2/p}\sup_{M\in\mathcal{M}}\|M\|_{HS}^2}\right),
\end{aligned}$$

where the third inequality follows from Lemma 5 and the fourth from Hölder's inequality. We now substitute

$$\delta = 2\left|\mathcal{M}\right|^{1/p}\sup_{M\in\mathcal{M}}\|M\|_{HS}\sqrt{\ln\frac{e\sum_M \|M\|_{HS}^2}{\sup_{N\in\mathcal{M}}\|N\|_{HS}^2}}$$

and use $1 + 1/e \leq 2$ to arrive at the conclusion. ■

Proof of Theorem 13 Apply Lemma 12 to the set of transformations $\mathcal{M}\mathbf{x} = \left\{M\mathbf{x} : M \in \mathcal{M}\right\}$. This gives

$$\mathbb{E}\left[\left(\sum_M \|M\mathbf{x}\varepsilon\|^p\right)^{1/p}\right] \leq 2\left|\mathcal{M}\right|^{1/p}\sup_{M\in\mathcal{M}}\|M\mathbf{x}\|_{HS}\left(2+\sqrt{\ln\frac{\sum_M\|M\mathbf{x}\|_{HS}^2}{\sup_{N\in\mathcal{M}}\|N\mathbf{x}\|_{HS}^2}}\right).$$

We now proceed as in the proof of Theorem 11 to obtain the result. ■

5. Conclusion and Future Work

We have presented a bound on the Rademacher average for linear function classes described by infimum convolution norms which are associated with a class of bounded linear operators on a Hilbert space. We highlighted the generality of the approach and its dimension independent features.

When the bound is applied to specific cases (ℓ_2, ℓ_1, mixed ℓ_1/ℓ_2 norms) it recovers existing bounds (up to small changes in the constants). The bound is however more general and allows for the possibility to remove the "$\log d$" factor which appears in previous bounds. Specifically, we have shown that the bound can be applied in infinite dimensional settings, provided that the moment condition (3) is satisfied. We have also applied the bound to multiple kernel learning. While in the standard case the bound is only slightly worse in the constants, the bound is potentially smaller and applies to the more general case in which there is a countable set of kernels, provided the expectation of the sum of the kernels is bounded.

An interesting question is whether the bound presented is tight. As noted in Cortes et al. (2010) the "$\log d$" is unavoidable in the case of the Lasso. This result immediately implies that our bound is also tight, since we may choose $R^2 = d$ in Equation (3).

A potential future direction of research is the application of our results in the context of sparsity oracle inequalities. In particular, it would be interesting to modify the analysis in Lounici et al. (2011), in order to derive dimension independent bounds. Another interesting scenario is the combination of our analysis with metric entropy.

Acknowledgments

We wish to thank Andreas Argyriou, Bharath Sriperumbudur, Alexandre Tsybakov and Sara van de Geer, for useful comments. A valuable contribution was made by one of the anonymous referees. Part of this work was supported by EPSRC Grant EP/H027203/1 and Royal Society International Joint Project Grant 2012/R2.

References

F.R. Bach, G.R.G. Lanckriet and M.I. Jordan. Multiple kernels learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004)*, pages 6–13, 2004.

L. Baldassarre, J. Morales, A. Argyriou, M. Pontil. A general framework for structured sparsity via proximal optimization. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, forthcoming.

R. Baraniuk, V. Cevher, M.F. Duarte, C. Hedge. Model based compressed sensing. *IEEE Trans. on Information Theory*, 56:1982–2001, 2010.

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3:463–482, 2002.

C. Cortes, M. Mohri, A. Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the Twenty-seventh International Conference on Machine Learning (ICML 2010*, pages 247–254, 2010.

C. De Mol, E. De Vito, L. Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

W. Hoeffding, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association*, 58:13–30, 1963.

J. Huang, T. Zhang, D. Metaxa. Learning with structured sparsity. In *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML 2009)*, pages 417–424, 2009.

L. Jacob, G. Obozinski, J.-P. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML 2009)*, pages 433–440, 2009.

R. Jenatton, J.-Y. Audibert, F.R. Bach. Structured variable selection with sparsity inducing norms. *Journal of Machine Learning Research*, 12:2777-2824, 2011.

S.M. Kakade, S. Shalev-Shwartz, A. Tewari. Regularization techniques for learning with matrices. *ArXiv preprint arXiv0910.0610*, 2010.

M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien. ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

V. Koltchinskii and D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *Annals of Statistics*, 30(1):1–50, 2002.

G.R.G. Lanckriet, N. Cristianini, P.L. Bartlett, L. El Ghaoui, M.I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

M. Ledoux, M. Talagrand. *Probability in Banach Spaces*, Springer, 1991.

K. Lounici, M. Pontil, A.B. Tsybakov and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39(4):2164–2204, 2011.

C. McDiarmid. Concentration. In *Probabilistic Methods of Algorithmic Discrete Mathematics*”, pages 195–248, Springer, 1998.

L. Meier, S.A. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Annals of Statistics*, 37(6B):3779–3821, 2009.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

C.A. Micchelli, J.M. Morales, M. Pontil. A family of penalty functions for structured sparsity. In *Advances in Neural Information Processing Systems 23*, pages 1612–1623, 2010.

C.A. Micchelli, J.M. Morales, M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, forthcoming.

C.A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, 66:297–319, 2007.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

T. Shimamura, S. Imoto, R. Yamaguchi and S. Miyano. Weighted Lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics*, 19:142–153, 2007.

Y. Ying and C. Campbell. Generalization bounds for learning the kernel problem. In *Proceedings of the 23rd Conference on Learning Theory (COLT 2009)*, pages 407–416, 2009.

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.

P. Zhao and G. Rocha and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.