The Rademacher Complexity of Linear Transformation Classes

Andreas Maurer

Adalbertstr. 55
D-80799 München
andreasmaurer@compuserve.com

Abstract. Bounds are given for the empirical and expected Rademacher complexity of classes of linear transformations from a Hilbert space H to a finite dimensional space. The results imply generalization guarantees for graph regularization and multi-task subspace learning.

1 Introduction

Rademacher averages have been introduced to learning theory as an efficient complexity measure for function classes, motivated by tight, sample or distribution dependent generalization bounds ([10], [2]). Both the definition of Rademacher complexity and the generalization bounds extend easily from real-valued function classes to function classes with values in $\mathbb{R}^m$, as they are relevant to multi-task learning ([1], [12]).

There has been an increasing interest in multi-task learning which has shown to be very effective in experiments ([7], [1]), and there have been some general studies of its generalisation performance ([4], [5]). For a large collection of tasks there are usually more data available than for a single task and these data may be put to a coherent use by some constraint of 'relatedness'. A practically interesting case is linear multi-task learning, extending linear large margin classifiers to vector valued large-margin classifiers. Different types of constraints have been proposed: Evgeniou et al ([8], [9]) propose graph regularization, where the vectors defining the classifiers of related tasks have to be near each other. They also show that their scheme can be implemented in the framework of kernel machines. Ando and Zhang [1] on the other hand require the classifiers to be members of a common low dimensional subspace. They also give generalization bounds using Rademacher complexity, but these bounds increase with the dimension of the input space. This paper gives dimension free bounds which apply to both approaches.

1.1 Multi-task generalization and Rademacher complexity

Suppose we have m classification tasks, represented by m independent random variables $\left(X^l, Y^l\right)$ taking values in $\mathcal{X} \times \{-1, 1\}$, where X^l models the random

occurrence of input data in some input space $\mathcal{X}$, and Y^l models the corresponding binary output for learning task $l \in \{1, ..., m\}$. The draw of an iid sample for the l-th task is described by a sequence $\left(X_i^l, Y_i^l\right)_{i=1}^n$ of independent random variables, each identically distributed to $\left(X^l, Y^l\right)$. Write $(\mathbf{X}, \mathbf{Y})$ for the combined random variable taking values in $\mathcal{X}^{mn} \times \{-1, 1\}^{mn}$.

One now seeks a function $\mathbf{f} = \left(f^1, ..., f^m\right) : \mathcal{X} \to \mathbb{R}^m$ such that predicting Y^l to be $\operatorname{sgn}\left(f^l\right)$ is correct with high average probability. To this end one searches a function class $\mathcal{F}$ of functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^m$ for a member with a small average empirical error estimate. The choice of the function class $\mathcal{F}$ expresses the constraints of 'relatedness' which we want to impose. This procedure is justified by the following result. ([1], [12]):

Theorem 1. *Let ϕ be the function on $\mathbb{R}$ defined by*

$$\phi(t) = \begin{cases} 1 & \text{if} \quad t \leq 0 \\ 1 - t & \text{if} \, 0 \leq t \leq 1 \\ 0 & \text{if} \quad 1 \leq t \end{cases}.$$

Let $\mathcal{F}$ be a class of functions $\mathbf{f} = \left(f^1, ..., f^m\right) : \mathcal{X} \to \mathbb{R}^m$ and fix $\delta > 0$. Then with probability greater than $1 - \delta$ we have for all $\mathbf{f} \in \mathcal{F}$

$$\frac{1}{m} \sum_{l=1}^{m} \Pr\left\{ sgn\left(f^l\left(X^l\right)\right) \neq Y^l \right\} \leq \frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \phi\left(Y^l f^l\left(X_i^l\right)\right) + \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) + \sqrt{\frac{9 \ln(2/\delta)}{2mn}}.$$

The first term on the right hand side is an empirical large-margin error estimate. Selecting a function class $\mathcal{F}$ means that we make a bet that we will be able to find within $\mathcal{F}$ a solution with a reasonably low value for this term. The other two terms bound the estimation error. The last term decays quickly with the product mn and depends only logarithmically on the confidence parameter δ and will not concern us very much. The remaining term is a complexity measure of the class $\mathcal{F}$ when acting on the data set $\mathbf{X}$.

Definition 1. *For $l \in \{1, ..., m\}$ and $i \in \{1, ..., n\}$ let σ_i^l be independent random variables, distributed uniformly in $\{-1, 1\}$. The empirical Rademacher complexity of a class $\mathcal{F}$ of functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^m$ is the function $\hat{\mathcal{R}}_n^m(\mathcal{F})$ defined on $\mathcal{X}^{nm}$ by*

$$\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{x}) = E_\sigma \left[\sup_{\mathbf{f} = (f^1, ..., f^m) \in \mathcal{F}} \frac{2}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \sigma_i^l f^l\left(x_i^l\right) \right].$$

Theorem 1 above explains the value of bounds on this function, the principal subject of this paper. There is also a version of Theorem 1 involving the expectation $E_{\mathbf{X}}\left[\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X})\right]$ with a slightly better final term. We have restricted ourselves to classification for definiteness. Substitution of our results in

other generalization bounds using Rademacher complexities should make them applicable to multi-task regression.

1.2 Bounds on the Rademacher complexity

This paper assumes that the input space $\mathcal{X}$ is contained in the closed unit ball of a real separable Hilbert space H (fixed from now on) and that $\mathcal{F}$ is a class of bounded linear transformations $V : H \to \mathbb{R}^m$. Such transformations correspond to m-tuples $\left(v^1, ..., v^m\right) \in H^m$ of vectors in H such that the l-th component of V is given by $V(x)_l = \left\langle v^l, x \right\rangle$ (we denote this by $V \leftrightarrow \left(v^1, ..., v^m\right)$). Thresholding the functional $x \to \left\langle v^l, x \right\rangle$ gives the classifier for the l-th task. The assumption $\left\|X^l\right\| \leq 1$ is also a notational convenience, but we would always need $E\left[\left\|X^l\right\|^2\right] < \infty$ for part (I) and $E\left[\left\|X^l\right\|^4\right] < \infty$ for part (II) of the following theorem, which is the main contribution of this work.

Theorem 2. *Let $\mathcal{F}$ be a set of linear transformations $V : H \to \mathbb{R}^m$ and $\mathbf{x} \in H^{mn}$ with $\left\|\mathbf{x}_i^l\right\|^2 \leq 1$.*

(I) Then for every positive definite operator A on $\mathbb{R}^m$

$$\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{x}) \leq \frac{2}{\sqrt{n}} \sup_{V \in \mathcal{F}} \left(\frac{\left\|A^{1/2} V\right\|_2}{\sqrt{m}} \right) \sqrt{\frac{tr\left(A^{-1}\right)}{m}}.$$

(II) Let $p \in [4, \infty]$ and let q be the conjugate exponent, that is $p^{-1} + q^{-1} = 1$. Then

$$\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{x}) \leq \frac{2}{\sqrt{n}} \sup_{V \in \mathcal{F}} \left(\frac{\|V\|_q}{\sqrt{m}} \right) \sqrt{\left\|\hat{C}(\mathbf{x})\right\|_{p/2} + \sqrt{\frac{2}{m}}}$$

and

$$E\left[\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X})\right] \leq \frac{2}{\sqrt{n}} \sup_{V \in \mathcal{F}} \left(\frac{\|V\|_q}{\sqrt{m}} \right) \sqrt{\|C_m\|_{p/2} + \sqrt{\frac{3}{m}}}$$

Here $\hat{C}(\mathbf{x})$ is the empirical covariance operator of the data set $\mathbf{x}$, that is the covariance operator corresponding to the empirical distribution $1/(mn) \sum_{l,i} \delta_{x_i^l}$ on H, while C_m is the covariance operator corresponding to the mixture of data-distributions[1]. $C_m = (1/m) C^l$, where C^l is the covariance operator for the data-distribution of the l-th task. The Schatten-norms $\|...\|_p$ are defined for compact operators T by

$$\|T\|_p = \left(\sum_i \mu_i^p \right)^{1/p},$$

[1] Here δ_x is the unit mass concentrated at x

where the μ_i are the singular values of T. The norms $\|T\|_p$ are a decreasing function in p. See section 2 for more detailed definitions. For $V \leftrightarrow \left(v^1, ..., v^m\right) : H \to \mathbb{R}^m$ we have

$$\|V\|_2 = \left(\sum_l \left\|v^l\right\|^2\right)^{1/2}.$$

1.3 Interpretation of the bounds

Each of the bounds in Theorem 2 has been grouped in three factors: The factor $2/\sqrt{n}$ is important, because it insures learnability as long as the other two factors remain bounded or increase slowly enough with the sample size n.

Next comes a regularization factor depending on $\mathcal{F}$ and a norm, which encodes the relatedness-constraint. Equating the supremum to some chosen constant B defines a maximal function class $\mathcal{F}$ which is necessarily convex. The first bound for example gives rise to the function class $\mathcal{F} = \left\{V : m^{-1/2} \left\|A^{1/2}V\right\|_2 < B\right\}$. For such classes the constant B can of course be substituted for the supremum, giving the bounds a simpler appearance. The $m^{-1/2}$ will typically be cancelled by allowing the individual functional components v^i of $V \leftrightarrow \left(v^1, ..., v^m\right) \in \mathcal{F}$ to have norm of unit order on average, that is $\|V\|_2^2 = \sum_l \left\|v^l\right\|^2 = O(m)$ or $\left\|A^{1/2}V\right\|_2^2 = O(m)$.

The third factor gives the bound proper and depends on the situation studied. It will typically decrease to some limiting positive value, as the number of tasks m increases.

If we set $A = I$ then part (I) above can be recognized as a trival extension of existing bounds ([2]) for single task linear large margin classifiers. It corresponds to the noninteracting case, essentially equivalent to single task learning. If we set $A = L + \eta I$, where L is the Laplacian on a graph with m vertices, and $\eta > 0$ a small regularization constant, then we obtain bounds to justify the graph regularization schemes in [9], concisely relating generalization to the spectrum of the Laplacian. This will be explained in some detail in section 3.

Part (II) of the theorem can be applied to subspace learning. The norms $\|T\|_p$ can be viewed as combined measures of amplitude and dimensionality (or rank if T has finite dimensional range), and imposing a bound on $\|V\|_q$ is a combined form of amplitude and dimensional regularization. The conceptually simplest way to do this is to consider the class $\mathcal{F}_{B,d}$ of transformations $V \leftrightarrow \left(v^1, ..., v^m\right)$ such that $\|V\|_2^2 = \sum_l \left\|v^l\right\|^2 \leq B^2 m$ and $rank(V) \leq d$ (a notation which extends to the case $d = \infty$). Then all the individual linear classifiers v^l are constrained to lie in some d-dimensional subspace of H. This subspace can be freely chosen after seeing the data, so the above bounds become generalisation guarantees for subspace selection through multitask learning. This corresponds to the regularization in [1]. In this case the regularization factor can be shown

to be equal to

$$\sup_{V \in \mathcal{F}_{B,d}} \left(\frac{\|V\|_q}{\sqrt{m}} \right) = B d^{\frac{2-q}{2q}} = B d^{\frac{p-2}{2p}}.$$

If $p = q = 2$ then this is just B and there will be no penalty on dimension. Correspondingly the bound will exhibit no benefit from constraining d. If $p = 4$ and $q = 4/3$ then we obtain $Bd^{1/4}$ and for $p = \infty$ and $q = 1$ it is $Bd^{1/2}$, corresponding to increasing penalties on the dimensionality. The class $\mathcal{F}_{B,d}$ is practical and corresponds to the scheme in [1], but it is not convex, while setting $\mathcal{F} = \left\{ V : \|V\|_q \leq B \right\}$ and replacing ϕ by the hinge loss always results in a convex optimization problem.

The data- or distribution dependent third factor in (II) contains two terms. The decrease in m with essentially the fourth root may be an artifact of our proof. As the number of tasks increases the norm of the covariances becomes dominant. Since we restricted ourselves to data in the unit ball, we will have $\left\| \hat{C}(\mathbf{x}) \right\|_1 \leq 1$ and $\|C_m\|_1 \leq 1$, so the amplitude is essentially normalized. Let us assume that the mixture of data-distributions is uniform on a k-dimensional unit sphere in H. Then C_m has k eigenvalues, all equal to $1/k$, so $\|C_m\|_{p/2} = k^{\frac{2-p}{p}}$. If we combine this with the $\mathcal{F}_{B,d}$ regularization considered above we obtain the bound

$$E\left[\hat{\mathcal{R}}_n^m \left(\mathcal{F}_{B,d} \right) \mathbf{X} \right] \leq \frac{2B}{\sqrt{n}} \left(\left(\frac{d}{k} \right)^{\frac{p-2}{p}} + d^{\frac{p-2}{p}} \sqrt{\frac{3}{m}} \right)^{1/2}.$$

The limiting value as the number m of tasks increases depends only on the fraction $\rho = d/k$, which might be viewed as the ratio of utilized information to totally present information k. If $\rho < 1$ multi-task learning will always be an improvement over single task learning for sufficiently large m (modulo the important requirement that the tasks are sufficiently related to arrive at a small empirical error despite the regularisation). In the limit $m \to \infty$ the best exponent is $p = \infty$ leading to the bound

$$\limsup_{m \to \infty} E\left[\hat{\mathcal{R}}_n^m \left(\mathcal{F}_{B,d} \right) (\mathbf{X}) \right] \leq \frac{2B}{\sqrt{n}} \rho^{1/2}.$$

For small values of m and large d smaller values of p will give a better bound.

If $\rho \geq 1$ constraining to $\mathcal{F}_{B,d}$ will bring no improvement over $\mathcal{F}_{B,\infty}$. This is understandable because constraining to at most d-dimensional subspaces has little effect when the data-distribution is already less than d-dimensional. Normally we expect that there exist low-dimensional subspaces expressing the relevant information in a chaos of data, which is the same as assuming $\rho \ll 1$.

A precursor of this paper is [12], where a result like part (II) of Theorem 2 is given for the case $p = 4$. It does not extend to larger values of p however, nor is it directly applicable to graph regularization.

The next section gives missing definitions and some important preliminary result. Section 3 gives a proof of part (I) of Theorem 2 and applies it to graph regularization. Section 4 is dedicated to the proof of part (II) of Theorem 2.

2 Definitions, Schatten-norms and Hoelders inequality

Throughout this paper we will use superscripts $l, r \in \{1, ..., m\}$ to index one of m learning tasks and we fix a real, separable Hilbert space H with inner product $\langle ., . \rangle$ and norm $\|.\|$, and assume the random variables X_i^l to be as described in the introduction. For a bounded operator T on H we generally use T^* to denote the adjoint and write $|T|^2 = T^*T$.

Let T be a compact operator from H to a Hilbert space H' and $\mu_i = \mu_i(T)$ its sequence of singular values in descending order, counting multiplicities. The μ_i are just the necessarily nonnegative eigenvalues of $(T^*T)^{1/2} = |T|$ (see [14] for background). For such operators T and $p \geq 1$ define

$$\|T\|_p = \left(\sum_i \mu_i^p \right)^{1/p} \text{ and } \mathcal{I}_p = \left\{ T : \|T\|_p < \infty \right\}.$$

We also define $\|T\|_\infty = \sup_i \mu_i = \mu_0$ and $\mathcal{I}_\infty$ as the set of all compact operators from H to another Hilbert space H'. As the notation indicates, $\|.\|_p$ does indeed define a norm making $\mathcal{I}_p$ into a Banach space. For $1 \leq p_1 \leq p_2 \leq \infty$ we have $\|T\|_{p_2} \leq \|T\|_{p_1}$. For $T \in \mathcal{I}_1$ the trace $tr(T)$ is defined by

$$tr(T) = \sum_i \langle Te_i, e_i \rangle,$$

where (e_i) is an orthonormal basis of H. This series converges absolutely and its limit is independent of the choice of basis. If A and B are in $\mathcal{I}_2$ then A^*B is in $\mathcal{I}_1$ and the inner product

$$\langle A, B \rangle_2 = tr(A^*B)$$

makes $\mathcal{I}_2$ into a Hilbert space, the space of Hilbert-Schmidt operators. This work will rely on Hoelder's inequality for compact operators, a beautiful classical theorem (see e.g. Reed-Simon [13]).

Theorem 3. *Let $1 \leq p \leq \infty$ and $q^{-1} + p^{-1} = 1$. If $A \in \mathcal{I}_p$ and $B \in \mathcal{I}_q$ then $AB \in \mathcal{I}_1$ and $|\langle A, B \rangle_2| = |tr(A^*B)| \leq \|A\|_p \|B\|_q$.*

Let V be a bounded operator $V : H \to \mathbb{R}^m$. Let $\left(e^l\right)_{l=1}^m$ be the canonical basis for $\mathbb{R}^m$. By the Riesz theorem there is an m-tuples $\left(v^1, ..., v^m\right) \in H^m$ of vectors in H such that

$$\left\langle Vx, e^l \right\rangle = \left\langle v^l, x \right\rangle \tag{1}$$

holds for all l. Conversely, if $\left(v^1, ..., v^m\right) \in H^m$ then the formula

$$Vx = \sum_{l=1}^m \left\langle v^l, x \right\rangle e^l$$

defines a bounded linear transformation V such that (1) holds. We will just write $V \leftrightarrow \left(v^1, ..., v^m\right)$ for this bijection. Observe that if $V, W : H \to \mathbb{R}^m$ with

$V \leftrightarrow \left(v^1, ..., v^m\right)$ and $W \leftrightarrow \left(w^1, ..., w^m\right)$, then

$$tr\left(W^* V\right) = \sum_{l=1}^{m} \left\langle v^l, w^l \right\rangle. \tag{2}$$

Definition 2. *For a configuration* $\boldsymbol{\sigma} = \left(\sigma_i^l\right)_{(l,i)=1}^{(m,n)} \in \{-1,1\}^{nm}$ *of the Rademacher variables and* $\mathbf{x} = \left(x_i^l\right)_{(l,i)=1}^{(m,n)} \in H^{nm}$ *with* $\left\|x_i^l\right\| \leq 1$ *define a linear transformation* $W\left(\boldsymbol{\sigma}, \mathbf{x}\right) : H \to \mathbb{R}^m$ *by* $W\left(\boldsymbol{\sigma}, \mathbf{x}\right) \leftrightarrow \left(w^1\left(\boldsymbol{\sigma}, \mathbf{x}\right), ..., w^m\left(\boldsymbol{\sigma}, \mathbf{x}\right)\right)$ *and*

$$w^l\left(\boldsymbol{\sigma}, \mathbf{x}\right) = \sum_{i=1}^{n} \sigma_i^l x_i^l.$$

When there is no ambiguity we drop the explicit dependance on either $\boldsymbol{\sigma}$ or $\mathbf{x}$ or both. W is thus an operator-valued random variable and its components $w^l = \sum_{i=1}^{n} \sigma_i^l x_i^l$ are vector valued random variables.

In our context the beauty of Hoelder's inequality is that it immediately splits the Rademacher complexity into a regularizing factor, depending on the function class $\mathcal{F}$ used for learning, and a data dependent factor:

Lemma 1. *For conjugate exponents* $p, q \geq 1$ *with* $1/p + 1/q = 1$ *and* $\mathbf{x} = \left(x_i^l\right)_{(l,i)=1}^{(m,n)} \in H^{nm}$ *and a class* $\mathcal{F}$ *of bounded linear transformations* $V : H \to \mathbb{R}^m$*, we have*

$$\hat{\mathcal{R}}_n^m\left(\mathcal{F}\right)\left(\mathbf{x}\right) \leq \frac{2}{\sqrt{n}} \left(\sup_{V \in \mathcal{F}} \frac{\left\|V\right\|_q}{\sqrt{m}}\right) \left(\frac{E_\sigma\left[\left\|W\left(\boldsymbol{\sigma}, \mathbf{x}\right)\right\|_p\right]}{\sqrt{mn}}\right).$$

Proof. Using the trace formula (2) and Hoelder's inequality (Theorem 3) we obtain

$$\begin{aligned}
\hat{\mathcal{R}}_n^m\left(\mathcal{F}\right)\left(\mathbf{x}\right) &= E_\sigma\left[\sup_{V \in \mathcal{F}} \frac{2}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \sigma_i^l \left\langle x_i^l, v^l \right\rangle\right] \\
&= E_\sigma\left[\sup_{V \in \mathcal{F}} \frac{2}{mn} \sum_{l=1}^{m} \left\langle w^l, v^l \right\rangle\right] \\
&= E_\sigma\left[\sup_{V \in \mathcal{F}} \frac{2}{mn} tr\left(W^* V\right)\right] \\
&\leq \frac{2}{mn} E_\sigma\left[\sup_{V \in \mathcal{F}} \left\|V\right\|_q \left\|W\right\|_p\right].
\end{aligned}$$

□

For every $x \in H$ we define an operator Q_x by $Q_x y = \langle y, x \rangle x$ for $y \in H$. The following facts are easily verified:

Lemma 2. *Let $x, y \in H$ and $p \in [1, \infty]$. Then*
(i) $Q_x \in \mathcal{I}_p$ and $\|Q_x\|_p = \|x\|^2$.
(ii) $\langle Q_x, Q_y \rangle_2 = \langle x, y \rangle^2$.
(iii) If $V \hookrightarrow (v^1, ..., v^m) : H \to \mathbb{R}^m$ then $|V|^2 = \sum_{l=1}^m Q_{v^l}$.

Let X be a random variable with values in H, such that $E[\|X\|] \leq \infty$. The linear functional $y \in H \mapsto E[\langle X, y \rangle]$ is bounded by $E[\|X\|]$ and thus defines (by the Riesz Lemma) a unique vector $E[X] \in H$ such that $E[\langle X, y \rangle] = \langle E[X], y \rangle, \forall y \in H$, with $\|E[X]\| \leq E[\|X\|]$.

If we also have $E\left[\|X\|^2\right] \leq \infty$ then we can apply the same construction to the random variable Q_X with values in the Hilbert space $\mathcal{I}_2$: By Lemma 2 (i) $E[\|Q_X\|_2] = E\left[\|X\|^2\right] \leq \infty$, so there is a unique operator $E[Q_X] \in \mathcal{I}_2$ such that $E[\langle Q_X, T \rangle_2] = \langle E[Q_X], T \rangle_2, \forall T \in \mathcal{I}_2$.

Definition 3. *The operator $E[Q_X]$ is called the covariance operator of X.*

We summarize some of its properties in the following lemma (see e.g. [12]). Property (ii) is sometimes taken as the defining property of the covariance operator.

Lemma 3. *The covariance operator $E[Q_X] \in \mathcal{I}_2$ has the following properties.*
(i) $\|E[Q_X]\|_2 \leq E[\|Q_X\|_2]$.
(ii) $\langle E[Q_X] y, z \rangle = E[\langle y, X \rangle \langle z, X \rangle], \forall y, z \in H$.
(iii) $tr(E[Q_X]) = E\left[\|X\|^2\right]$.

If $\mathbf{x} \in H^{mn}$ with $\mathbf{x} = \left(x_i^l : l \in \{1, ..., m\}, i \in \{1, ..., n\}\right)$ is a data-set, $\hat{E}$ be the expectation corresponding to the empirical distribution $1/(mn) \sum_{l,i=1}^{m,n} \delta_{x_i^l}$. The corresponding empirical covariance $\hat{C}(\mathbf{x})$ is the operator

$$\hat{C}(\mathbf{x}) = \hat{E}[Q_X] = \frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} Q_{x_i^l}.$$

3 Graph Regularization

We give a proof of part (I) of Theorem 2 and sketch how it applies to graph regularization as described in [8] and [9].

Proof (of Theorem 2, part (I)). Beginning as in the proof of Lemma 1 we obtain from Hoelder's inequality in the simplest (Schwarz-) case $p = q = 2$

$$\begin{aligned}
\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{x}) &= E_\sigma \left[\sup_{V \in \mathcal{F}} \frac{2}{mn} tr(W^* V) \right] \\
&= E_\sigma \left[\sup_{V \in \mathcal{F}} \frac{2}{mn} tr\left(W^* A^{-1/2} A^{1/2} V \right) \right] \\
&\leq \frac{2}{\sqrt{n}} \sup_{V \in \mathcal{F}} \left(\frac{\left\| A^{1/2} V \right\|_2}{\sqrt{m}} \right) \frac{E_\sigma \left[\left\| W^* A^{-1/2} \right\|_2 \right]}{\sqrt{mn}}.
\end{aligned}$$

To prove part (I) it therefore suffices to show that $E_\sigma\left[\left\|W^*A^{-1/2}\right\|_2\right] \leq \left(n\ tr\left(A^{-1}\right)\right)^{1/2}$. Using Jensen's inequality, independence and symmetry of the Rademacher variables, we obtain

$$\begin{aligned} E_\sigma\left[\left\|W^*A^{-1/2}\right\|_2\right]^2 &\leq E\left[\left\|W^*A^{-1/2}\right\|_2^2\right] = E\left[tr\left(W^*A^{-1}W\right)\right] \\ &= \sum_{l=1}^m\sum_{r=1}^m A_{lr}^{-1}\sum_{i=1}^n\sum_{j=1}^n E\left[\sigma_i^l\sigma_j^r\right]\left\langle x_i^l, x_j^r\right\rangle \\ &= \sum_{l=1}^m A_{ll}^{-1}\sum_{i=1}^n\left\|x_i^l\right\|^2 \\ &\leq n\ tr\left(A^{-1}\right), \end{aligned}$$

as required.□

Suppose that we have some way to quantify the 'relatedness' ω_{lr} of any pair (l,r) of distinct learning tasks, where we require symmetry $\omega_{lr} = \omega_{rl}$ and nonnegativity $\omega_{lr} \geq 0$. For simplicity we will assume connectivity in the sense that for all pairs (l,r), there is a sequence of indices $(l_i)_{i=0}^K$ such that $l = l_0$ and $r = l_K$ and $\omega_{l_{k-1}l_k} > 0$ for all $1 \leq k \leq K$.

The idea of graph regularization ([8], [9]) is to use a regularizer $J(V) = J\left(v^1, ..., v^m\right)$, which forces the classifiers of related tasks to be near each other, penalizing the squared distance $\left\|v^l - v^r\right\|^2$ proportional to ω_{lr}. Such a regularizer is

$$\begin{aligned} J(V) &= \frac{1}{2m}\sum_{l,r}\omega_{lr}\left\|v^l - v^r\right\|^2 + \frac{\eta}{m}\sum_{l=1}^m\left\|v^l\right\|^2 \\ &= \frac{1}{m}\sum_{l,r}\left(L + \eta I\right)_{lr}\left\langle v^l, v^r\right\rangle, \end{aligned}$$

where L is the Laplacian of the graph with m vertices and edge-weights ω, and I is the identity in $\mathbb{R}^m$. We have slightly departed from the form given in [9] by adding the term in η. We will however see, that a large number m of tasks allows η to be chosen small.

Fix $B > 0$. We will bound the Rademacher complexity of the function class $\mathcal{F} = \left\{V : J(V) \leq B^2\right\}$. Substitution of our bound in Theorem 1 will then lead to generalisation guarantees for graph regularisation.

To bound $\hat{\mathcal{R}}_n^m(\mathcal{F})$ note that a transformation $V \leftrightarrow \left(v^1, ..., v^m\right)$ belongs to $\mathcal{F}$ if and only if

$$\begin{aligned} \sum_{l,r}\left(L + \eta I\right)_{lr}\left\langle v^l, v^r\right\rangle \leq mB^2 &\iff tr\left(V^*\left(L + \eta I\right)V\right) \leq mB^2 \\ &\iff m^{-1/2}\left\|\left(L + \eta I\right)^{1/2}V\right\|_2 \leq B. \end{aligned}$$

Using Theorem 2 (I) with $A = L + \eta I$ therefore gives

$$\begin{aligned}\hat{\mathcal{R}}_n^m(\mathcal{F}) &\leq \frac{2B}{\sqrt{n}}\sqrt{\frac{tr\left((L+\eta I)^{-1}\right)}{m}} \\ &= \frac{2B}{\sqrt{n}}\sqrt{\frac{1}{m}\sum_{i=2}^{m}\frac{1}{\lambda_i+\eta}+\frac{1}{m\eta}} \qquad (3) \\ &\leq \frac{2B}{\sqrt{n}}\sqrt{\frac{1}{\lambda_2}+\frac{1}{m\eta}}, \qquad (4)\end{aligned}$$

where $\lambda_2, ..., \lambda_m$ are the nonzero eigenvalues of the Laplacian in (now) ascending order (with $\lambda_1 = 0$ having multiplicity 1 - it is here that we used connectivity). For a large number of tasks m we can choose η small, say $\eta = O(1/\sqrt{m})$, and the contribution of the Laplacian becomes dominant. Which of the bounds (3) or (4) is preferable depends on the nature of the Laplacian, which in turn depends on the coupling constants ω_{lr}.

For a particularly simple example consider $\omega_{lr} = c/m$ for all distinct tasks l and r, where c is some positive constant. The regularizer then becomes

$$\begin{aligned}J(V) &= \frac{1}{2m}\sum_{l,r}\frac{c}{m}\left\|v^l - v^r\right\|^2 + \frac{\eta}{m}\sum_{l=1}^{m}\left\|v^l\right\|^2 \\ &= \frac{c}{m}\sum_{l}\left\|v^l - \frac{1}{m}\sum_{r}v^r\right\|^2 + \frac{\eta}{m}\sum_{l=1}^{m}\left\|v^l\right\|^2,\end{aligned}$$

and can be recognized as the regularizer in section 3.1.1 in [9]. The corresponding Laplacian is $L_{lr} = c(\delta_{lr} - 1/m)$, and the nonzero eigenvalues are all equal to c. Substitution in the bound (4) then gives

$$\hat{\mathcal{R}}_n^m(\mathcal{F}) \leq \frac{2B}{\sqrt{n}}\sqrt{\frac{1}{c}+\frac{1}{m\eta}},$$

exhibiting both the benefit of assuming a large 'relatedness' c of the tasks, and the increasing irrelevance of the general regularization parameter η for a large number m of tasks.

4 Bounding the expected norm of $W(\boldsymbol{\sigma}, \mathbf{x})$

Now we prove part (II) of Theorem 2. Hoelders inequality essentially reduces the problem of the proof to the analysis of the expected norm of $W = W(\sigma, \mathbf{x})$, $W \leftrightarrow \left(w^1, ..., w^m\right)$. Our idea of proof is to instead study $|W|^2$, which is easier to deal with. We compute the expectation and bound the variance of this random variable

Lemma 4. *We have the two identities*
(i) $E_\sigma\left[\left|W\left(\sigma,\mathbf{x}\right)\right|^2\right]=mn\hat{C}\left(\mathbf{x}\right)$
(ii) $E_{\mathbf{X}}E_\sigma\left[\left|W\left(\sigma,\mathbf{X}\right)\right|^2\right]=mnC_m$

Proof. For a fixed configuration $\mathbf{x}$ and any $y,z\in H$ we have, by independence and symmetry of the Rademacher variables,

$$\begin{aligned}\left\langle E_\sigma\left[W^*W\right]y,z\right\rangle &= E_\sigma\left[\left\langle Wy,Wz\right\rangle\right]=\sum_{l=1}^{m}E_\sigma\left[\left\langle w^l,y\right\rangle\left\langle w^l,z\right\rangle\right]\\ &=\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}E_\sigma\left[\sigma_i^l\sigma_j^l\right]\left\langle x_i^l,y\right\rangle\left\langle x_j^l,z\right\rangle\\ &=\sum_{l=1}^{m}\sum_{i=1}^{n}\left\langle x_i^l,y\right\rangle\left\langle x_i^l,z\right\rangle=\left\langle\sum_{l=1}^{m}\sum_{i=1}^{n}Q_{x_i^l}y,z\right\rangle\\ &=mn\left\langle\hat{C}\left(\mathbf{x}\right)y,z\right\rangle.\end{aligned}$$

The second equation follows from replacing $\mathbf{x}$ by $\mathbf{X}$ in the first one and applying $E_{\mathbf{X}}$.□

Lemma 5. *For fixed* $\mathbf{x}\in H^{mn}$ *with* $\left\|x_i^l\right\|\leq 1$ *we have*

$$E_\sigma\left[\left\|\left|W\left(\sigma,\mathbf{x}\right)\right|^2-E_\sigma\left[\left|W\left(\sigma,\mathbf{x}\right)\right|^2\right]\right\|_2\right]\leq n\sqrt{2m}.$$

Also

$$E_{\mathbf{X}}E_\sigma\left[\left\|\left|W\left(\sigma,\mathbf{X}\right)\right|^2-E_{\mathbf{X}}E_\sigma\left[\left|W\left(\sigma,\mathbf{X}\right)\right|^2\right]\right\|_2\right]\leq n\sqrt{3m}$$

Proof. We use the representation $\left|W\right|^2=\left|W\left(\sigma,\mathbf{x}\right)\right|^2=\sum_{l=1}^{m}Q_{w^l}$ as introduced in section 2, Lemma 2, with

$$w^l=w^l\left(\sigma,\mathbf{x}\right)=\sum_{i=1}^{n}\sigma_i^lx_i^l.$$

To prove the first inequality we keep $\mathbf{x}$ fixed. Let the η_i^l be iid copies of σ_i^l and write $\hat{W}=W\left(\eta,\mathbf{x}\right)$ and $\hat{w}^l=w^l\left(\eta,\mathbf{x}\right)$. Then

$$\begin{aligned}E_\sigma\left[\left\|\left|W\right|^2-E_\sigma\left[\left|W\right|^2\right]\right\|_2^2\right] &= E_{\sigma,\eta}\left[\left\langle\left|W\right|^2,\left|W\right|^2\right\rangle_2-\left\langle\left|W\right|^2,\left|\hat{W}\right|^2\right\rangle_2\right]\\ &=\sum_{l=1}^{m}\sum_{r=1}^{m}E_{\sigma,\eta}\left[\left\langle Q_{w^l},Q_{w^r}\right\rangle_2-\left\langle Q_{w^l},Q_{\hat{w}^r}\right\rangle_2\right]\\ &=\sum_{l=1}^{m}E_{\sigma,\eta}\left[\left\langle Q_{w^l},Q_{w^l}\right\rangle_2-\left\langle Q_{w^l},Q_{\hat{w}^l}\right\rangle_2\right],\end{aligned}$$

because of the independence of w^l and w^r for $l \neq r$. The l-th term in the last expression is equal to

$$E_{\sigma,\eta}\left[\left\|w^l\right\|^4 - \left\langle w^l, \hat{w}^l\right\rangle^2\right] = \sum_{i,j,i',j'} E_{\sigma,\eta}\left[\sigma_i^l\sigma_j^l\sigma_{i'}^l\sigma_{j'}^l - \sigma_i^l\eta_j^l\sigma_{i'}^l\eta_{j'}^l\right]\left\langle x_i^l, x_j^l\right\rangle\left\langle x_{i'}^l, x_{j'}^l\right\rangle.$$

By independence and symmetry of the σ variables the expectation on the right will have the value one if $i = j$ and $i' = j'$ or if $i = j'$ and $i' = j$. If $i = i'$ and $j = j'$ a cancellation will occur, so the expectation will be zero. In all other cases it vanishes because there will be some factor of the σ_k^l occurring only once. We conclude that

$$E_{\sigma,\eta}\left[\left\|w^l\right\|^4 - \left\langle w^l, \hat{w}^l\right\rangle^2\right] = \sum_{i,j}\left(\left\|x_i^l\right\|^2\left\|x_j^l\right\|^2 + \left\langle x_i^l, x_j^l\right\rangle^2\right) \leq 2n^2.$$

Summing over l we get with Jensen's inequality,

$$E_\sigma\left[\left\||W|^2 - E_\sigma\left[|W|^2\right]\right\|_2\right] \leq \left(E_\sigma\left[\left\||W|^2 - E_\sigma\left[|W|^2\right]\right\|_2^2\right]\right)^{1/2} \leq \left(2mn^2\right)^{1/2},$$

which is the first inequality.

To prove the second inequality we also introduce iid copies $\hat{X}_i^l$ of X_i^l and write $W = W(\sigma, \mathbf{X})$ and $\hat{W} = W\left(\eta, \hat{\mathbf{X}}\right)$. Proceeding as before we obtain

$$E_{\sigma,\mathbf{X}}\left[\left\||W|^2 - E_{\sigma,\mathbf{X}}\left[|W|^2\right]\right\|_2^2\right] = \sum_{l=1}^m E_{\sigma,\eta,\mathbf{X},\hat{\mathbf{X}}}\left[\left\|w^l\right\|^4 - \left\langle w^l, \hat{w}^l\right\rangle^2\right].$$

Now we have

$$\begin{aligned} E_{\sigma,\eta,\mathbf{X},\hat{\mathbf{X}}}\left[\left\|w^l\right\|^4 - \left\langle w^l, \hat{w}^l\right\rangle^2\right] &\leq E_{\sigma,\mathbf{X}}\left[\left\|w^l\right\|^4\right] \\ &= \sum_{i,j,i',j'} E_\sigma\left[\sigma_i^l\sigma_j^l\sigma_{i'}^l\sigma_{j'}^l\right] E_{\mathbf{X}}\left[\left\langle x_i^l, x_j^l\right\rangle\left\langle x_{i'}^l, x_{j'}^l\right\rangle\right]. \end{aligned}$$

Now $E_\sigma\left[\sigma_i^l\sigma_j^l\sigma_{i'}^l\sigma_{j'}^l\right]$ will be nonzero and equal to one if either $i = j$ and $i' = j'$ or $i = i'$ and $j = j'$ or $i = j'$ and $j = i'$, which gives a bound of $3n^2$ on the above expectation. Summing over l we obtain

$$E_{\sigma,\mathbf{X}}\left[\left\|W^*W - E\left[W^*W\right]\right\|_2^2\right] \leq 3mn^2$$

and the conclusion follows from Jensen's inequality.□

Proof (of Theorem 2, part (II)). We have from Lemma 4, the triangle inequality, the nonincreasing nature of spectral norms $\|.\|_q$ and Lemma 5 for any $q \geq 2$

$$\begin{aligned} E_\sigma\left[\left\||W|^2\right\|_q\right] - mn\left\|\hat{C}(\mathbf{x})\right\|_q &= E_\sigma\left[\left\||W|^2\right\|_q - \left\|E_\sigma\left[|W|^2\right]\right\|_q\right] \\ &\leq E_\sigma\left[\left\||W|^2 - E_\sigma\left[|W|^2\right]\right\|_q\right] \\ &\leq E_\sigma\left[\left\||W|^2 - E_\sigma\left[|W|^2\right]\right\|_2\right] \\ &\leq n\sqrt{2m}. \end{aligned}$$

Similarly we obtain $E_{\mathbf{X}}E_\sigma\left[\left\||W|^2\right\|_q\right] - mn\left\|C_m\right\|_q \leq n\sqrt{3m}$. It follows from Jensen's inequality that for $p \geq 4$

$$\begin{aligned}
(mn)^{-1/2} E_\sigma\left[\|W\|_p\right] &= (mn)^{-1/2} E_\sigma\left[\left\||W|^2\right\|_{p/2}^{1/2}\right] \\
&\leq (mn)^{-1/2}\left(E_\sigma\left[\left\||W|^2\right\|_{p/2}\right]\right)^{1/2} \\
&\leq (mn)^{-1/2}\left(mn\left\|\hat{C}\left(\mathbf{x}\right)\right\|_{p/2} + n\sqrt{2m}\right)^{1/2} \\
&= \left(\left\|\hat{C}\left(\mathbf{x}\right)\right\|_{p/2} + \sqrt{\frac{2}{m}}\right)^{1/2}.
\end{aligned}$$

In the same way we obtain

$$(mn)^{-1/2} E_{\mathbf{X}}E_\sigma\left[\|W\|_p\right] \leq \left(\left\|C_m\right\|_{p/2} + \sqrt{\frac{3}{m}}\right)^{1/2}.$$

Substitution in Lemma 1 completes the proof. □

5 Conclusion

We showed that an application of Hoelder's inequality to bound the Rademacher complexity of linear transformation classes leads to generalization bounds for various regularization schemes of multi-task learning. Two major defects of the results presented are the following:

- The decrease in the r.h.s of the bound in part II of Theorem 2 with $O\left(m^{-1/4}\right)$. Is this a necessary feature or an artifact of a clumsy proof? In [12] there is a similar bound with $O\left(m^{-1/2}\right)$, but it requires that the transformations can be factored $V = ST$ where $S : H \to \mathbb{R}^m$ has the property $\left\|S^* e^l\right\| \leq 1$ for the canonical basis $\left(e^l\right)$ of $\mathbb{R}^m$. Also the result in [12] is worse in the limit $m \to \infty$, diverging for constant ρ and $d \to \infty$ (in context and notation of section 1.3).
- Part II of Theorem 2 might well be valid for all $p \in [2, \infty]$, instead of just $p \in \{2\} \cup [4, \infty]$ (the case $p = 2$ follows trivially from part I). This would follow if something like Lemma 5 was true also for the 1-norm instead of the 2-norm.

References

1. R. K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, *Journal of Machine Learning Research*, 6: 1817-1853, 2005.

2. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3: 463-482, 2002.
3. P.Bartlett, O.Bousquet and S.Mendelson. Local Rademacher complexities. Available online: http://www.stat.berkeley.edu/~bartlett/papers/bbm-lrc-02b.pdf.
4. J.Baxter, Theoretical Models of Learning to Learn, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998
5. J.Baxter, A Model of Inductive Bias Learning, *Journal of Artificial Intelligence Research* 12: 149-198, 2000
6. S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *COLT 03*, 2003.
7. R.Caruana, Multitask Learning, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998.
8. T. Evgeniou and M. Pontil, Regularized multi-task learning. *Proc. Conference on Knowledge Discovery and Data Mining*, 2004.
9. T. Evgeniou, C. Micchelli and M. Pontil, Learning multiple tasks with kernel methods. *JMLR*, 6: 615-637, 2005.
10. V. Koltchinskii and D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *The Annals of Statistics*, Vol. 30, No 1, 1-50.
11. M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.
12. A. Maurer, Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.
13. Michael Reed and Barry Simon. *Fourier Analysis, Self-Adjointness*, part II of *Methods of Mathematical Physics, Academic Press*, 1980.
14. Michael Reed and Barry Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.
15. S.Thrun, Lifelong Learning Algorithms, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998