Journal of Machine Learning Research 7 (2006) 117–139 Submitted 9/05; Revised 12/05; Published 1/06

Bounds for Linear Multi-Task Learning

Andreas Maurer ANDREASMAURER@COMPUSERVE.COM
Adalbertstrasse 55
D-80799 München, Germany

Editor: Nello Cristianini

Abstract

We give dimension-free and data-dependent bounds for linear multi-task learning where a common linear operator is chosen to preprocess data for a vector of task specific linear-thresholding classifiers. The complexity penalty of multi-task learning is bounded by a simple expression involving the margins of the task-specific classifiers, the Hilbert-Schmidt norm of the selected preprocessor and the Hilbert-Schmidt norm of the covariance operator for the total mixture of all task distributions, or, alternatively, the Frobenius norm of the total Gramian matrix for the data-dependent version. The results can be compared to state-of-the-art results on linear single-task learning.

Keywords: learning to learn, transfer learning, multi-task learning

1. Introduction

Simultaneous learning of different tasks under some common constraint, often called *multi-task learning*, has been tested in practice with good results under a variety of different circumstances (see Baxter 1998, Caruana 1998, Thrun 1998, Ando and Zhang 2005). The technique has been analyzed theoretically and in some generality by Baxter (2000) and Ando and Zhang (2005). The latter reference appears to be the first to use Rademacher averages in this context. The purpose of this paper is to improve some of these theoretical results in a special case of practical importance, when input data are represented in a linear, potentially infinite dimensional space, and the common constraint is a linear preprocessor.

Finite systems provide simple examples illustrating the potential advantages of multi-task learning. Consider agnostic learning with an input space $\mathcal{X}$ and a finite set $\mathcal{F}$ of hypotheses $f : \mathcal{X} \to \{0,1\}$. For a hypothesis $f \in \mathcal{F}$ let $\mathrm{er}(f)$ be the expected error and $\hat{\mathrm{er}}(f)$ the empirical error on a training sample S of size n (drawn iid from the underlying task distribution) respectively. Combining Hoeffding's inequality with a union bound one shows (see e.g. Anthony and Bartlett 1999), that with probability greater than $1-\delta$ we have for every $f \in \mathcal{F}$ the error bound

$$\mathrm{er}(f) \leq \hat{\mathrm{er}}(f) + \frac{1}{\sqrt{2n}}\sqrt{\ln|\mathcal{F}| + \ln(1/\delta)}. \tag{1}$$

Suppose now that there are a set $\mathcal{Y}$, a finite but large set $\mathcal{G}$ of preprocessors $g : \mathcal{X} \to \mathcal{Y}$, and another set $\mathcal{H}$ of classifiers $h : \mathcal{Y} \to \{0,1\}$ with $|\mathcal{H}| \ll |\mathcal{F}|$. For a cleverly chosen preprocessor $g \in \mathcal{G}$ it will likely be the case that we find some $h \in \mathcal{H}$ such that $h \circ g$ has the same or even a smaller empirical error than the best $f \in \mathcal{F}$. But this will lead to an improvement of the bound above (replacing $|\mathcal{F}|$ by $|\mathcal{H}|$) only if we choose g before seeing the data, otherwise we incur a large estimation penalty for the selection of g (replacing $|\mathcal{F}|$ by $|\mathcal{H} \circ \mathcal{G}|$).

The situation is improved if we have a set of m different learning tasks with corresponding task distributions and samples $S_1,...,S_m$, each of size n and drawn iid from the corresponding distributions. We now consider solutions $h_1 \circ g,... \; h_m \circ g$ for each of the m tasks where the preprocessing map $g \in \mathcal{G}$ is *constrained to be the same for all tasks* and only the $h_l \in \mathcal{H}$ specialize to each task l at hand. Again Hoeffding's inequality and a union bound imply that with probability greater $1-\delta$ we have for all $(h_1,...,h_m) \in \mathcal{H}^m$ and every $g \in \mathcal{G}$

$$\frac{1}{m}\sum_{l=1}^{m} \mathrm{er}^l (h_l \circ g) \leq \frac{1}{m}\sum_{l=1}^{m} \hat{\mathrm{er}}^l (h_l \circ g) + \frac{1}{\sqrt{2n}}\sqrt{\ln \left|\mathcal{H}\right| + \frac{\ln |\mathcal{G}| + \ln (1/\delta)}{m}}. \tag{2}$$

Here $\mathrm{er}^l (f)$ and $\hat{\mathrm{er}}^l (f)$ denote the expected error in task l and the empirical error on training sample S_l respectively. The left hand side above is an average of the expected errors, so that the guarantee implied by the bound is a little weaker than the usual PAC guarantees (but see Ben-David, 2003, for bounds on the individual errors). The first term on the right is the average empirical error, which a multi-task learning algorithm seeks to minimize. We can take it as an operational definition of task-relatedness relative to $(\mathcal{H}, \mathcal{G})$ that we are able to obtain a very small value for this term. The remaining term, which bounds the estimation error, now exhibits the advantage of multi-task learning: Sharing the preprocessor implies sharing its cost of estimation, and the entropy contribution arising from the selection of $g \in \mathcal{G}$ decreases with the number of learning tasks. Since by assumption $\left|\mathcal{H}\right| \ll |\mathcal{F}|$, the estimation error in the multi-task bound (2) can become much smaller than in the single task case (1) if the number m of tasks becomes large.

The choice of the preprocessor $g \in \mathcal{G}$ can also be viewed as the selection of the hypothesis space $\mathcal{H} \circ g$. This leads to an alternative formulation of multi-task learning, where the common object is a hypothesis space chosen from a class of hypothesis spaces (in this case $\left\{\mathcal{H} \circ g : g \in \mathcal{G}\right\}$), and the classifiers for the individual tasks are all chosen from the selected hypothesis space. Here we prefer the functional formulation of selecting a preprocessor instead of a hypothesis space, because it is more intuitive and sufficient in the situations which we consider.

The arguments leading to (2) can be refined and extended to certain infinite classes to give general bounds for multi-task learning (Baxter 2000, Ando and Zhang 2005). In this paper we concentrate on the case where the input space $\mathcal{X}$ is a subset of the unit ball in a Hilbert space H, the class $\mathcal{G}$ of preprocessors is a set $\mathcal{A}$ of bounded symmetric linear operators on H, and the class $\mathcal{H}$ is the set of classifiers h_v obtained by 0-thresholding linear functionals v in H with $\|v\| \leq B$, that is

$$h_v (x) = \mathrm{sign}\left(\langle x, v\rangle\right) \text{ and } h_v \circ T\,(x) = \mathrm{sign}\left(\langle Tx, v\rangle\right), x \in H, T \in \mathcal{A}, \; \|v\| \leq B.$$

The learner now searches for a multi-classifier $\mathbf{h_v} \circ T = (h_{v^1} \circ T,...,h_{v^m} \circ T)$ where the preprocessing operator $T \in \mathcal{A}$ is the same for all tasks and only the vectors v^l specialize to each task l at hand. The desired multi-classifier $\mathbf{h_v} \circ T$ should have a small value of the average error

$$\mathrm{er}\left(\mathbf{h_v} \circ T\right) = \frac{1}{m}\sum_{l=1}^{m} \mathrm{er}^l (h_{v_l} \circ T) = \frac{1}{m}\sum_{l=1}^{m} \Pr\left\{\mathrm{sign}\left(\left\langle TX^l, v^l\right\rangle\right) \neq Y^l\right\},$$

where X^l and Y^l are the random variables modeling input-values and labels for the l-th task. To guide this search we look for bounds on $\mathrm{er}(\mathbf{h_v} \circ T)$ in terms of the total observed data for all tasks, valid uniformly for all $\mathbf{v} = \left(v^1,...,v^m\right)$ with $\left\|v^l\right\| \leq B$ and all $T \in \mathcal{A}$. We will prove the following :

Theorem 1 *Let $\delta \in (0,1)$. With probability greater than $1-\delta$ it holds for all $\mathbf{v} = \left(v^1,...,v^m\right) \in H$ with $\left\|v^l\right\| \leq 1$ and all bounded symmetric operators T on H with $\|T\|_{HS} \geq 1$, and for all $\gamma \in (0,1)$ that*

$$er\left(\mathbf{h}_{\mathbf{v}} \circ T\right) \leq \hat{er}_{\gamma}(\mathbf{v} \circ T) + \frac{8\|T\|_{HS}}{\gamma\sqrt{n}}\sqrt{\|C\|_{HS} + \frac{1}{m}} + \sqrt{\frac{\ln\frac{4\|T\|_{HS}}{\delta\gamma}}{2nm}}.$$

Here $\hat{\mathrm{er}}_{\gamma}(\mathbf{v} \circ T)$ is a margin-based empirical error estimate, bounded by the relative number of examples $\left(X_i^l, Y_i^l\right)$ in the total training sample for all tasks l, where $Y_i^l \left\langle TX_i^l, v^l \right\rangle < \gamma$ (see section 4).

The quantity $\|T\|_{HS}$ is the *Hilbert-Schmidt norm* of T, defined for symmetric T by

$$\|T\|_{HS} = \left(\sum \lambda_i^2\right)^{1/2},$$

where λ_i is the sequence of eigenvalues of T (counting multiplicities, see section 2).

C is *the total covariance operator* corresponding to the mixture of all the task-input-distributions in H. Since data are constrained to the unit ball in H we always have $\|C\|_{HS} \leq 1$ (see section 3).

The above theorem is the simplest, but not the tightest or most general form of our results. For example the factor 8 on the right hand side can be decreased to be arbitrarily close to 2, thereby incurring only a logarithmic penalty in the last term.

A special case results from restricting the set of candidate preprocessors to $\mathcal{P}_d$, the set of orthogonal projections in H with d-dimensional range. In this case learning amounts to the selection of a d-dimensional subspace M of H and of an m-tuple of vectors v^l in M (components of v^l orthogonal to M are irrelevant to the projected data). All operators $T \in \mathcal{P}_d$ satisfy $\|T\|_{HS} = \sqrt{d}$, which can then be substituted in the above bound. Identifying such a projection with the structural parameter Θ, this corresponds to the case considered by Ando and Zhang (2005), where a practical algorithm for this type of multi-task learning is presented. The identity $\|\Theta\|_{HS} = \sqrt{d}$ then expresses the regularization condition mentioned in (Ando, Zhang 2005).

The bound in the above theorem is dimension free, it does not require the data distribution in H to be confined to a finite dimensional subspace. Almost to the contrary: Suppose that the input data are distributed uniformly on $M \cap S_1$ where M is a k-dimensional subspace in H and S_1 is the sphere consisting of vectors with unit norm in H. Then C has the k-fold eigenvalue $1/k$, the remaining eigenvalues being zero. Therefore $\|C\|_{HS} = 1/\sqrt{k}$, so part of the bound above decreases to zero as the dimensionality of the data-distribution increases, in contrast to the bound in (Ando, Zhang, 2005), which increases linearly in k. The fact that our bounds are dimension free allows their general use for multi-task learning in kernel-induced Hilbert spaces (see Cristianini and Shawe-Taylor 2000).

If we compare the second term on the right hand side to the estimation error bound in (2), we can recognize a certain similarity: Loosely speaking we can identify $\|T\|_{HS}^2/m$ with the cost of estimating the operator T, and $\|T\|_{HS}^2\|C\|_{HS}$ with the cost of finding the linear classifiers $v_1,...,v_m$. The order of dependence on the number of tasks m is the same in Theorem 1 as in (2).

In the limit $m \to \infty$ it is preferable to use a different bound (see Theorems 13 and 15), at the expense of slower convergence in m. The main inequality of the theorem then becomes

$$\mathrm{er}\left(\mathbf{h}_{\mathbf{v}} \circ T\right) \leq \hat{\mathrm{er}}_{\gamma}(\mathbf{v} \circ T) + \frac{2\left\|T^2\right\|_{HS}^{1/2}}{(1-\varepsilon)^2\gamma\sqrt{n}}\left(\|C\|_{HS}^2 + \frac{3}{m}\right)^{1/4} + \sqrt{\frac{\ln\frac{\left\|T^2\right\|_{HS}^{1/2}}{\delta\gamma\varepsilon^2}}{2nm}}. \tag{3}$$

for some very small $\varepsilon > 0$ to be fixed in advance. If T is an orthogonal projection with d-dimensional range then $\left\|T^2\right\|_{HS}^{1/2} = d^{1/4}$, so for a large number of tasks m the bound on the estimation error becomes approximately

$$\frac{2d^{1/4}\left\|C\right\|_{HS}^{1/2}}{\gamma\sqrt{n}}.$$

One of the best dimension-free bounds for linear single-task learning (see e.g. Bartlett and Mendelson 2002, or Lemma 11 below) would give $2/\left(\gamma\sqrt{n}\right)$ for this term, if all data are constrained to unit vectors. We therefore expect superior estimation for multi-task learning of d-dimensional projections with large m, whenever $d^{1/4}\left\|C\right\|_{HS}^{1/2} \ll 1$. If we again assume the data-distribution to be uniform on $M \cap S_1$ with M a k-dimensional subspace, this is the case whenever $d \ll k$, that is, qualitatively speaking, whenever the dimension of the utilizable part of the data is considerably smaller than the dimension of the total data distribution.

The results stated above give some insights, but they have the practical disadvantage of being unobservable, because they depend on the properties of the covariance operator C, which in turn depends on an unknown data distribution. One way to solve this problem is using the fact that the finite-sample approximations to the covariance operator have good concentration properties (see Theorem 8 below). The corresponding result is:

Theorem 2 *With probability greater than $1-\delta$ in the sample $\mathbf{X}$ it holds for all $v_1,...,v_m \in H$ with $\|v_l\| \leq 1$ and all bounded symmetric operators T on H with $\|T\|_{HS} \geq 1$, and for all $\gamma \in (0,1)$ that*

$$er\left(\mathbf{h}_{\mathbf{v}} \circ T\right) \leq \hat{er}_{\gamma}\left(\mathbf{v} \circ T\right) + \frac{8\left\|T\right\|_{HS}}{\gamma\sqrt{n}}\sqrt{\frac{1}{mn}\left\|\hat{C}\left(\mathbf{X}\right)\right\|_{Fr} + \frac{1}{m}} + \sqrt{\frac{9\ln\frac{8\|T\|_{HS}}{\delta\gamma}}{2nm}}.$$

where the $\left\|\hat{C}\left(\mathbf{X}\right)\right\|_{Fr}$ is the Frobenius norm of the gramian.

By definition

$$\left\|\hat{C}\left(\mathbf{X}\right)\right\|_{Fr} = \left(\sum_{l,r,i,j}\left\langle X_i^l, X_j^r\right\rangle^2\right)^{1/2}.$$

Here X_i^l is the random variable describing the i-th data point in the sample corresponding to the l-th task. The corresponding label Y_i^l enters only in the empirical margin error. The quantity $(mn)^{-1}\left\|\hat{C}\left(\mathbf{X}\right)\right\|_{Fr}$ can be regarded as an approximation to $\left\|C\right\|_{HS}$, valid with high probability, so that Theorem 2 is a sample based version of Theorem 1.

In section 2 we introduce the necessary terminology and results on Hilbert-Schmidt operators and in section 3 the covariance operator of random elements in a Hilbert space. Section 4 gives a formal definition of multi-task systems and a general PAC bound in terms of Rademacher complexities. For the readers benefit a proof of this bound is given in an appendix, where we follow the path prepared by Kolchinskii and Panchenko (2002) and Bartlett and Mendelson (2002). In section 5 we study the Rademacher complexities of linear multi-task systems. In section 6 we give bounds for non-interacting systems, which are essentially equivalent to single-task learning, and derive bounds for proper, interacting multi-task learning, including the above mentioned results. We conclude with

the construction of an example to demonstrate the advantages of multi-task learning. The appendix contains missing proofs and a convenient reference-table to the notation and definitions introduced in the paper.

2. Hilbert-Schmidt Operators

For a fixed real, separable Hilbert space H, with inner product $\langle .,. \rangle$ and norm $\|.\|$, we define a second real, separable Hilbert space consisting of *Hilbert-Schmidt operators*. With HS we denote the real vector space of operators T on H satisfying $\sum_{i=1}^{\infty} \|Te_i\|^2 \leq \infty$ for every orthonormal basis $(e_i)_{i=1}^{\infty}$ of H. Every $T \in HS$ is bounded. For $S,T \in HS$ and an orthonormal basis (e_i) the series $\sum_i \langle Se_i, Te_i \rangle$ is absolutely summable and independent of the chosen basis. The number $\langle S,T \rangle_{HS} = \sum \langle Se_i, Te_i \rangle$ defines an inner product on HS, making it into a Hilbert space. We denote the corresponding norm with $\|.\|_{HS}$ in contrast to the usual operator norm $\|.\|_{\infty}$. See Reed and Simon (1980) for background on functional analysis). We use HS^* to denote the set of symmetric Hilbert-Schmidt operators. For every member of HS^* there is a complete orthonormal basis of eigenvectors, and for $T \in HS^*$ the norm $\|T\|_{HS}$ is just the ℓ_2-norm of its sequence of eigenvalues. With HS^+ we denote the members of HS^* with only nonnegative eigenvalues.

We use two simple maps from H or H^2 to HS to relate the geometries of objects in H to the geometry in HS.

Definition 3 *Let $x,y \in H$. We define two operators Q_x and $G_{x,y}$ on H by*

$$\begin{aligned} Q_x z &= \langle z,x \rangle x, \forall z \in H \\ G_{x,y} z &= \langle x,z \rangle y, \forall z \in H. \end{aligned}$$

We will frequently use parts of the following lemma, the proof of which is very easy.

Lemma 4 *Let $x,y,x',y' \in H$ and $T \in HS$. Then*
(i) $Q_x \in HS^+$ and $\|Q_x\|_{HS} = \|x\|^2$.
(ii) $\langle Q_x, Q_y \rangle_{HS} = \langle x,y \rangle^2$.
(iii) $\langle T, Q_x \rangle_{HS} = \langle Tx,x \rangle$.
(iv) $\langle T^*T, Q_v \rangle_{HS} = \|Tv\|^2$.
(v) $Q_y Q_x = \langle x,y \rangle G_{x,y}$.
(vi) $G_{x,y} \in HS$ and $\|G_{x,y}\|_{HS} = \|x\| \|y\|$.
(vii) $\langle G_{x,y}, G_{x',y'} \rangle_{HS} = \langle x,x' \rangle \langle y,y' \rangle$
(viii) $\langle T, G_{x,y} \rangle_{HS} = \langle Tx,y \rangle$.
(ix) For $\alpha \in \mathbb{R}$, $Q_{\alpha x} = \alpha^2 Q_x$.

Proof For $x = 0$ (iii) is obvious. For $x \neq 0$ choose an orthonormal basis $(e_i)_1^{\infty}$, so that $e_1 = x/\|x\|$. Then e_1 is the only nonzero eigenvector of Q_x with eigenvalue $\|x\|^2 > 0$. Also

$$\langle T, Q_x \rangle_{HS} = \sum_i \langle Te_i, Q_x e_i \rangle = \langle Tx, Q_x x \rangle / \|x\|^2 = \langle Tx,x \rangle,$$

which gives (iii). (ii), (i) and (iv) follow from substitution of Q_y, Q_x and T^*T respectively for T. (v) follows directly from the definition when applied to any $z \in H$. Let $(e_k)_{k=1}^{\infty}$ be any orthonormal

basis. Then $x = \sum_k \langle x, e_k\rangle e_k$, so by boundedness of T

$$\begin{aligned}\langle Tx, y\rangle &= \left\langle T\sum_k \langle x, e_k\rangle e_k, y\right\rangle = \sum_k \langle Te_k, \langle x, e_k\rangle y\rangle = \sum_k \langle Te_k, G_{x,y}e_k\rangle \\ &= \langle T, G_{x,y}\rangle_{HS},\end{aligned}$$

which is (viii). Similarly

$$\begin{aligned}\left\langle G_{x,y}, G_{x',y'}\right\rangle_{HS} &= \sum_k \left\langle \langle x, e_k\rangle y, \left\langle x', e_k\right\rangle y'\right\rangle = \left\langle y, y'\right\rangle \sum_k \langle x, e_k\rangle \left\langle x', e_k\right\rangle \\ &= \left\langle x, x'\right\rangle \left\langle y, y'\right\rangle,\end{aligned}$$

which gives (vii) and (vi). (ix) is obvious. ■

The following application of Lemma 4 is the key to our main results.

Lemma 5 *Let $T \in HS$ and $w_1, \ldots, w_m$ and $v_1, \ldots, v_m$ vectors in H with $\|v_i\| \le B$. Then*

$$\sum_{l=1}^m \langle Tw_l, v_l\rangle \le B\|T\|_{HS}\left(\sum_{l,r}|\langle w_l, w_r\rangle|\right)^{1/2}$$

and

$$\sum_{l=1}^m \langle Tw_l, v_l\rangle \le Bm^{1/2}\|T^*T\|_{HS}^{1/2}\left(\sum_{l,r}\langle w_l, w_r\rangle^2\right)^{1/4}$$

Proof Without loss of generality assume $B = 1$. Using Lemma 4 (viii), Schwarz' inequality in HS and Lemma 4 (vii) we have

$$\begin{aligned}\sum_{l=1}^m \langle Tw_l, v_l\rangle &= \left\langle T, \sum_{l=1}^m G_{w_l,v_l}\right\rangle_{HS} \le \|T\|_{HS}\left\|\sum_{l=1}^m G_{w_l,v_l}\right\|_{HS} \\ &= \|T\|_{HS}\left(\sum_{l,r}^m \langle w_l, w_r\rangle\langle v_l, v_r\rangle\right)^{1/2} \\ &\le \|T\|_{HS}\left(\sum_{l,r}^m |\langle w_l, w_r\rangle|\right)^{1/2}.\end{aligned}$$

This proves the first inequality. Also, using Schwarz' inequality in H and $\mathbb{R}^m$, Lemma 4 (iv) and Schwarz' inequality in HS

$$\begin{aligned}\sum_{l=1}^m \langle Tw_l, v_l\rangle &\le \left(\sum_{l=1}^m \|v_l\|^2\right)^{1/2}\left(\sum_{l=1}^m \|Tw_l\|^2\right)^{1/2} \le \sqrt{m}\left\langle T^*T, \sum_{l=1}^m Q_{w_l}\right\rangle_{HS}^{1/2} \\ &\le \sqrt{m}\|T^*T\|_{HS}^{1/2}\left\|\sum_{l=1}^m Q_{w_l}\right\|_{HS}^{1/2} = \sqrt{m}\|T^*T\|_{HS}^{1/2}\left(\sum_{l,r}\langle w_l, w_r\rangle^2\right)^{1/4}\end{aligned}$$

■

The set of d-dimensional, orthogonal projections in H is denoted with $\mathcal{P}_d$. We have $\mathcal{P}_d \subset HS^*$ and if $P \in \mathcal{P}_d$ then $\|P\|_{HS} = \sqrt{d}$ and $P^2 = P$.

An operator T is called *trace-class* if $\sum_{i=1}^{\infty} \langle Te_i, e_i \rangle$ is an absolutely convergent series for every orthonormal basis $(e_i)_{i=1}^{\infty}$ of H. In this case the number $tr(T) = \sum_{i=1}^{\infty} \langle Te_i, e_i \rangle$ is called the *trace* of T and it is independent of the chosen basis.

If $\mathcal{A} \subset HS^*$ is a set of symmetric and bounded operators in H we use the notation

$$\|\mathcal{A}\|_{HS} = \sup \{\|T\|_{HS} : T \in \mathcal{A}\} \text{ and } \mathcal{A}^2 = \left\{T^2 : T \in \mathcal{A}\right\}.$$

3. Vector- and Operator-Valued Random Variables

Let (Ω, Σ, μ) be a probability space with expectation $E[F] = \int_\Omega F d\mu$ for a random variable $F : \Omega \to \dot{\mathbb{R}}$. Let X be a random variable with values in H, such that $E[\|X\|] < \infty$. The linear functional $v \in H \mapsto E[\langle X, v \rangle]$ is bounded by $E[\|X\|]$ and thus defines (by the Riesz Lemma) a unique vector $E[X] \in H$ such that $E[\langle X, v \rangle] = \langle E[X], v \rangle, \forall v \in H$, with $\|E[X]\| \leq E[\|X\|]$.

We now look at the random variable Q_X, with values in HS. Suppose that $E\left[\|X\|^2\right] < \infty$. Passing to the space HS of Hilbert-Schmidt operators the above construction can be carried out again: By Lemma 4 (i) $E[\|Q_X\|_{HS}] = E\left[\|X\|^2\right] < \infty$, so there is a unique operator $E[Q_X] \in HS$ such that $E[\langle Q_X, T \rangle_{HS}] = \langle E[Q_X], T \rangle_{HS}, \forall T \in HS$.

Definition 6 *The operator $E[Q_X]$ is called the covariance operator of X.*

Lemma 7 *The covariance operator $E[Q_X] \in HS^+$ has the following properties.*
(i) $\|E[Q_X]\|_{HS} \leq E[\|Q_X\|_{HS}]$.
(ii) $\langle E[Q_X]y, z \rangle = E[\langle y, X \rangle \langle z, X \rangle], \forall y, z \in H$.
(iii) $tr(E[Q_X]) = E\left[\|X\|^2\right]$.
(iv) For H-valued independent X_1 and X_2 with $E\left[\|X_i\|^2\right] \leq \infty$ we have

$$\langle E[Q_{X_1}], E[Q_{X_2}] \rangle_{HS} = E\left[\langle X_1, X_2 \rangle^2\right].$$

(v) Under the same hypotheses, if $E[X_2] = 0$ then

$$E[Q_{X_1+X_2}] = E[Q_{X_1}] + E[Q_{X_2}]$$

Proof (i) follows directly from the construction, (iv) from the identity $\langle E[Q_{X_1}], E[Q_{X_2}] \rangle_{HS} = E\left[\langle Q_{X_1}, Q_{X_2} \rangle_{HS}\right]$. Let $y, z \in H$. Then using 4 (viii) we get

$$\begin{aligned} \langle E[Q_X]y, z \rangle &= \langle E[Q_X], G_{y,z} \rangle_{HS} = E\left[\langle Q_X, G_{y,z} \rangle_{HS}\right] = E[\langle Q_X y, z \rangle] \\ &= E[\langle y, X \rangle \langle z, X \rangle] \end{aligned}$$

and hence (ii). We have with orthonormal basis $(e_k)_{k=1}^{\infty}$ and using (ii)

$$tr(E[Q_X]) = \sum_k \langle E[Q_X]e_k, e_k \rangle = \sum_k E[\langle e_k, X \rangle \langle e_k, X \rangle] = E\left[\|X\|^2\right],$$

which gives (iii). Substitution of an eigenvector v for both y and z in (ii) shows that the corresponding eigenvalue must be nonnegative, so $E[Q_X] \in HS^+$.

Finally (v) holds because for any $y,z \in H$ we have, using independence and the mean-zero condition for X_2, that

$$\begin{aligned}
&\langle E[Q_{X_1+X_2}]y,z\rangle \\
&= E[\langle y,X_1+X_2\rangle\langle X_1+X_2,z\rangle] \\
&= E[\langle y,X_1\rangle\langle X_1,z\rangle] + E[\langle y,X_2\rangle\langle X_2,z\rangle] + E[\langle y,X_1\rangle\langle X_2,z\rangle] + E[\langle y,X_2\rangle\langle X_1,z\rangle] \\
&= \langle (E[Q_{X_1}]+E[Q_{X_2}])y,z\rangle + \langle y,E[X_1]\rangle\langle E[X_2],z\rangle + \langle y,E[X_2]\rangle\langle E[X_1],z\rangle \\
&= \langle (E[Q_{X_1}]+E[Q_{X_2}])y,z\rangle
\end{aligned}$$

■

Property (ii) above is sometimes taken as the defining property of the covariance operator (see Baxendale 1976).

If X is distributed uniformly on $M \cap S_1$, where M is a k-dimensional subspace and S_1 the unit sphere in H, then $E\left[\langle X,y\rangle^2\right] = \langle E[Q_X]y,y\rangle$ is zero if and only if $y \in M^\perp$, so the range of $E[Q_X]$ is M, so there are exactly k-eigenvectors corresponding to non-zero eigenvalues of $E[Q_X]$. By symmetry these eigenvalues must all be equal, and by (iii) above they sum up to 1, so $E[Q_X]$ has the k-fold eigenvalue $1/k$, with zero as the only other eigenvalue. It follows that $\|E[Q_X]\|_{HS} = 1/\sqrt{k}$. We have given this derivation to illustrate the tendency of the Hilbert-Schmidt norm of the covariance operator of a distribution concentrated on unit vectors to decrease with the effective dimensionality of the distribution. This idea is relevant to the interpretation of our results.

The fact that HS is a separable Hilbertspace just as H allows to define an operator $E[T]$ whenever T is a random variable with values in HS and $E[\|T\|_{HS}] < \infty$. Also any result valid in H has a corresponding analogue valid in HS. We quote a corresponding operator-version of a Theorem of Cristianini and Shawe-Taylor (2004) on the concentration of independent vector-valued random variables.

Theorem 8 *Suppose that $T_1,\ldots,T_m$ are independent random variables in H with $\|T_i\| \leq 1$. Then for all $\delta > 0$ with probability greater than δ we have*

$$\left\| \frac{1}{m}\sum_{i=1}^m E[T_i] - \frac{1}{m}\sum_{i=1}^m T_i \right\|_{HS} \leq \frac{2}{\sqrt{m}}\left(1+\sqrt{\frac{\ln(1/\delta)}{2}}\right).$$

Apply this with $T_i = Q_{X_i}$ where the X_i are iid H-valued with $\|X_i\| \leq 1$. The theorem then shows that the covariance operator $E[Q_X]$ can be approximated in HS-norm with high probability by the empirical estimates $(1/m)\sum_i Q_{X_i}$. The quantity

$$\left\| \sum_i Q_{X_i} \right\|_{HS} = \left(\sum_{i,j} \langle X_i,X_j\rangle^2 \right)^{1/2}$$

is the Frobenius norm of the Gramian (or kernel-) matrix $\hat{C}(\mathbf{X})_{ij} = \langle X_i,X_j\rangle$, denoted $\left\|\hat{C}(\mathbf{X})\right\|_{Fr}$. An immediate corollary to the above is, that $(1/m)\left\|\hat{C}(\mathbf{X})\right\|_{Fr}$ is with high probability a good approximation of $\|E[Q_X]\|_{HS}$. In the proof of Theorem 2 we will not need this fact however.

4. Multi-Task Problems and General Bounds

For our discussion of multi-task learning we concentrate on binary labeled data. Let (Ω, Σ, μ) be a probability space. We assume that there is a *multi-task problem* modeled by m independent random variables $Z^l = \left(X^l, Y^l\right) : \Omega \to \mathcal{X} \times \{-1,1\}$, where

- $l \in \{1,...,m\}$ identifies one of the m learning tasks,
- X^l models the input data of the l-th task, distributed in a set $\mathcal{X}$, called the *input space*.
- $Y^l \in \{-1,1\}$ models the output-, or label-data of the l-th task.
- For each $l \in \{1,...,m\}$ there is an n-tuple of independent random variables $\left(Z_i^l\right)_{i=1}^n = \left(X_i^l, Y_i^l\right)_{i=1}^n$, where each Z_i^l is identically distributed to Z^l.

The random variable $\mathbf{Z} = \left(Z_i^l\right)_{(i,l)=(1,1)}^{(n,m)}$ is called the *training sample* or training data. We also write $\mathbf{X} = \left(X_i^l\right)_{(i,l)=(1,1)}^{(n,m)}$. We use the superscript ^{l} to identify learning tasks running from 1 to m, the subscript $_i$ to identify data points in the sample, running from 1 to n. We will use the notations $\mathbf{x} = \left(x_i^l\right)_{(i,l)=(1,1)}^{(n,m)}$ for generic members of $(\mathcal{X}^n)^m$ and $\mathbf{z} = \left(z_i^l\right)_{(i,l)=(1,1)}^{(n,m)} = (\mathbf{x}, \mathbf{y}) = \left(x_i^l, y_i^l\right)_{(i,l)=(1,1)}^{(n,m)}$ for generic members of $((\mathcal{X} \times \{-1,1\})^n)^m$.

A *multiclassifier* is a map $\mathbf{h} : \mathcal{X} \to \{-1,1\}^m$. We write $\mathbf{h} = \left(h^1, ..., h^m\right)$ and interpret $h^l(x)$ as the label assigned to the vector x when the task is known to be l. The average error of a multiclassifier $\mathbf{h}$ is the quantity

$$\mathrm{er}(\mathbf{h}) = \frac{1}{m} \sum_{l=1}^{m} \Pr\left\{h^l\left(X^l\right) \neq Y^l\right\},$$

which is just the misclassification probability averaged over all tasks. Typically a classifier is chosen from some candidate set minimizing some error estimate based on the training data $\mathbf{Z}$. Here we consider zero-threshold classifiers $\mathbf{h_f}$ which arise as follows:

Suppose that $\mathcal{F}$ is a class of vector valued functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^m$ with $\mathbf{f} = \left(f^1, ..., f^m\right)$. A function $\mathbf{f} \in \mathcal{F}$ defines a multi-classifier $\mathbf{h_f} = \left(h_\mathbf{f}^1, ..., h_\mathbf{f}^m\right)$ through $h_\mathbf{f}^l(x) = \mathrm{sign}\left(f^l(x)\right)$. To give uniform error bounds for such classifiers in terms of empirical estimates, we define for $\gamma > 0$ the margin functions

$$\phi_\gamma(t) = \begin{cases} 1 & \text{if} \quad t \leq 0 \\ 1 - t/\gamma & \text{if} \quad 0 < t < \gamma \\ 0 & \text{if} \quad \gamma \leq t \end{cases},$$

and for $\mathbf{f} \in \mathcal{F}$ the random variable

$$\hat{\mathrm{er}}_\gamma(\mathbf{f}) = \frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \phi_\gamma\left(Y_i^l f^l\left(X_i^l\right)\right),$$

called the *empirical γ-margin error* of $\mathbf{f}$. The following Theorem gives a bound on $\mathrm{er}(\mathbf{h_f})$ in terms of $\hat{\mathrm{er}}_\gamma(\mathbf{f})$, valid with high probability uniformly in $\mathbf{f} \in \mathcal{F}$ and γ.

Theorem 9 *Let* $\varepsilon, \delta \in (0,1)$

(i) With probability greater than $1-\delta$ *it holds for all* $\mathbf{f} \in \mathcal{F}$ *and all* $\gamma \in (0,1)$ *that*

$$er(\mathbf{h_f}) \leq \hat{er}_\gamma(\mathbf{f}) + \frac{1}{\gamma(1-\varepsilon)} E\left[\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X})\right] + \sqrt{\frac{\ln(1/(\delta\gamma\varepsilon))}{2nm}}.$$

(ii) With probability greater than $1-\delta$ *it holds for all* $\mathbf{f} \in \mathcal{F}$ *and all* $\gamma \in (0,1)$ *that*

$$er(\mathbf{h_f}) \leq \hat{er}_\gamma(\mathbf{f}) + \frac{1}{\gamma(1-\varepsilon)} \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) + \sqrt{\frac{9\ln(2/(\delta\gamma\varepsilon))}{2nm}}.$$

Here $\hat{\mathcal{R}}_n^m(\mathcal{F})$ is the *empirical Rademacher complexity* in the sense of the following

Definition 10 *Let* $\left\{\sigma_i^l : l \in \{1,...,m\}, i \in \{1,...,n\}\right\}$ *be a collection of 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* $(\mathcal{X}^n)^m$ *by*

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

For the readers convenience we give a proof of Theorem 9 in the appendix.

The bounds in the Theorem each involve three terms. The last one expresses the dependence of the estimation error on the confidence parameter δ and a model-selection penalty $\ln(1/(\gamma\varepsilon))$ for the choice of the margin γ. Note that it generally decreases as $1/\sqrt{nm}$. This is not an a priori advantage of multi-task learning, but a trivial consequence of the fact that we estimate an average of m probabilities (in contrast to Ben David, 2003, where bounds are valid for each individual task - of course under more restrictive assumptions). The $1/\sqrt{nm}$ decay however implies that even for moderate values of m and n the parameter ε in Theorem 9 can be chosen very small, so that the factor $1/(1-\varepsilon)$ in the second term on the right of the two bounds is very close to unity.

The second term involves the complexity of the function class $\mathcal{F}$, either as measured in terms of the distribution of the random variable $\mathbf{X}$ or in terms of the observed sample. Since the distribution of $\mathbf{X}$ is unobservable in practice, the bound (i) is primarily of theoretical importance, while (ii) can be used to drive an algorithm which selects the multi-classifier $\mathbf{h}_{\mathbf{f}^*}$, where $(\mathbf{f}^*, \gamma) \in \mathcal{F} \times (0,1)$ are chosen to minimize the right side of the bound with given δ, ε. It is questionable if minimizing upper bounds is a good strategy, but it can serve as a motivating guideline.

Of key importance in the analysis of these algorithms is the empirical Rademacher complexity $\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X})$, as observed on the sample $\mathbf{X}$, and its expectation, measuring respectively the sample- and distribution-dependent complexities of the function class $\mathcal{F}$. Bounds on these quantities can be substituted in Theorem 9 to give corresponding error bounds.

5. The Rademacher Complexity of Linear Multi-Task Learning

We will now concentrate on multi-task learning in the linear case, when the data live in a real, separable Hilbert space H, by means of some kernel-induced embedding (see Cristianini and Shawe-Taylor 2000), the details of which will not concern us at this point. We therefore take H as input space $\mathcal{X}$, so that the random variables X^l take values in H for all $l \in \{1,...,m\}$, and we generally

require $\left\|X^l\right\| \leq 1$. The case $\left\|X^l\right\| = 1$ where the data are constrained to the unit sphere in H is of particular interest, corresponding to a class of radial basis function kernels.

We write C^l for the covariance operator $E\left[Q_{X^l}\right]$ and C for the total covariance operator $C = (1/m)\sum_l C^l$, corresponding to a uniform mixture of distributions. By Lemma 7 we have $\left\|C^l\right\|_{HS} \leq tr\left(C^l\right) = E\left[\left\|X^l\right\|^2\right] \leq 1$.

Let $B > 0$, let T be a fixed symmetric, bounded linear operator on H with $\|T\|_\infty \leq 1$, and let $\mathcal{A}$ be a set of symmetric, bounded linear operators T on H, all satisfying $\|T\|_\infty \leq 1$. We will consider the vector-valued function classes

$$\begin{aligned}
\mathcal{F}_B &= \{x \in H \mapsto (v_1,...,v_m)(x) := (\langle x, v_1\rangle,...,\langle x, v_m\rangle) : \|v_i\| \leq B\} \\
\mathcal{F}_B \circ T &= \{x \in H \mapsto (v_1,...,v_m) \circ T(x) := (\langle Tx, v_1\rangle,...,\langle Tx, v_m\rangle) : \|v_i\| \leq B\} \\
\mathcal{F}_B \circ \mathcal{A} &= \{x \in H \mapsto (v_1,...,v_m) \circ T(x) : \|v_i\| \leq B, T \in \mathcal{A}\}.
\end{aligned}$$

The algorithms which choose from $\mathcal{F}_B$ and $\mathcal{F}_B \circ T$ are essentially trivial extensions of linear single-task learning, where the tasks do not interact in the selection of the individual classifiers v_i, which are chosen independently. In the case of $\mathcal{F}_B \circ T$ the preprocessing operator T is chosen before seeing the training data. Since $\|T\|_\infty \leq 1$ we have $\mathcal{F}_B \circ T \subseteq \mathcal{F}_B$, so that we can expect a reduced complexity for $\mathcal{F}_B \circ T$ and the key question becomes if the choice of T (possibly based on experience with other data) was lucky enough to allow for a sufficiently low empirical error.

The non-interacting classes $\mathcal{F}_B$ and $\mathcal{F}_B \circ T$ are important for comparison to $\mathcal{F}_B \circ \mathcal{A}$ which represents proper multi-task learning. Here the preprocessing operator T is selected from $\mathcal{A}$ in response to the data. The constraint that T be the same for all tasks forces an interaction of tasks in the choice of T and $(v_1,...,v_m)$, deliberately aiming for a low empirical error. At the same time we also have $\mathcal{F}_B \circ \mathcal{A} \subseteq \mathcal{F}_B$, so that again a reduced complexity is to be expected, giving a smaller contribution to the estimation error. The promise of multi-task learning is based on the combination of these two ideas: Aiming for a low empirical error, using a function class of reduced complexity.

We first look at the complexity of the function class $\mathcal{F}_B$. The proof of the following lemma is essentially the same as the proof of Lemma 22 in Bartlett and Mendelson (2002).

Lemma 11 *We have*

$$\begin{aligned}
\hat{\mathcal{R}}_n^m(\mathcal{F}_B)(\mathbf{x}) &\leq \frac{2B}{nm}\sum_{l=1}^m\left(\sum_{i=1}^n\left\|x_i^l\right\|^2\right)^{1/2} \\
E\left[\hat{\mathcal{R}}_n^m(\mathcal{F}_B)(\mathbf{X})\right] &\leq \frac{2B}{\sqrt{n}}\frac{1}{m}\sum_{l=1}^m\left(E\left[\left\|X^l\right\|^2\right]\right)^{1/2} = \frac{2B}{\sqrt{n}}\frac{1}{m}\sum_{l=1}^m tr\left(C^l\right)^{1/2}
\end{aligned}$$

Proof Using Schwarz' and Jensen's inequality and the independence of the σ_i^l we get

$$\begin{aligned}
\hat{\mathcal{R}}_n^m(\mathcal{F}_B)(\mathbf{x}) &= E_\sigma\left[\sup_{v_1,\ldots,v_m,\|v_l\|\le B}\frac{2}{nm}\sum_{l=1}^m\left\langle\sum_{i=1}^n\sigma_i^l x_i^l, v_l\right\rangle\right] \\
&\le BE_\sigma\left[\frac{2}{nm}\sum_{l=1}^m\left\|\sum_{i=1}^n\sigma_i^l x_i^l\right\|\right] \\
&\le \frac{2B}{nm}\sum_{l=1}^m\left(E_\sigma\left[\left\|\sum_{i=1}^n\sigma_i^l x_i^l\right\|^2\right]\right)^{1/2} \\
&= \frac{2B}{nm}\sum_{l=1}^m\left(\sum_{i=1}^n\left\|x_i^l\right\|^2\right)^{1/2}.
\end{aligned}$$

Jensen's inequality then gives the second conclusion ■

The first bound in the lemma is just the average of the bounds given by Bartlett and Mendelson in on the empirical complexities for the various task-components of the sample. For inputs constrained to the unit sphere in H, when $\left\|X^l\right\| = 1$, both bounds become $2B/\sqrt{n}$, which sets the mark for comparison with the interacting case $\mathcal{F}_B \circ \mathcal{A}$. For motivation we next look at the case $\mathcal{F}_B \circ T$, working with a fixed linear preprocessor T of operator norm bounded by 1. Using the above bound we obtain

$$\hat{\mathcal{R}}_n^m(\mathcal{F}_B \circ T)(\mathbf{x}) = \hat{\mathcal{R}}_n^m(\mathcal{F}_B)(T\mathbf{x}) \le \frac{2B}{nm}\sum_{l=1}^m\left(\sum_{i=1}^n\left\|Tx_i^l\right\|^2\right)^{1/2}, \tag{4}$$

which is always bounded by $B/\sqrt{n}$, because $\|Tx\| \le \|x\|, \forall x$. Using Lemma 4 (iv) we can rewrite the right side above as

$$\frac{2B}{\sqrt{n}}\frac{1}{m}\sum_{l=1}^m\left\langle T^2, \frac{1}{n}\sum_{i=1}^n Q_{x_i^l}\right\rangle_{HS}^{1/2}.$$

Taking the expectation and using the concavity of the root function gives, with two applications of Jensen's inequality and an application of Schwarz' inequality (in HS),

$$E\left[\hat{\mathcal{R}}_n^m(\mathcal{F}_B \circ T)(\mathbf{X})\right] \le \frac{2B}{\sqrt{n}}\left\|T^2\right\|_{HS}^{1/2}\|C\|_{HS}^{1/2},$$

which can be significantly smaller than $B/\sqrt{n}$, for example if T is a d-dimensional projection, and the data-distribution is spread well over a much more than d-dimensional submanifold of the unit ball in H, as explained in the introduction and section 3. If we substitute the bound above in Theorem 9 we obtain an inequality which looks like (3) in the limit $m \to \infty$.

We now consider the case where T is chosen from some set $\mathcal{A}$ of (symmetric, bounded) candidate operators on the basis of the same sample $\mathbf{X}$, simultaneous to the determination of the classification vectors $v_1, \ldots, v_l$. We give two bounds each for the Rademacher complexity and its expectation. One is somewhat similar to other bounds for multi-task learning (e.g. (2)) and another one is tighter in the limit when the number of tasks m goes to infinity.

Theorem 12 *The following inequalities hold*

$$\hat{\mathcal{R}}_{\mathbf{x}}^{m}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) \leq \frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\sqrt{\frac{1}{mn}\left\|\hat{C}(\mathbf{x})\right\|_{Fr}+\frac{1}{m}} \tag{5}$$

$$\hat{\mathcal{R}}_{\mathbf{x}}^{m}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) \leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{\sqrt{n}}\left(\left(\frac{1}{mn}\left\|\hat{C}(\mathbf{x})\right\|_{Fr}\right)^2+\frac{2}{m}\right)^{1/4} \tag{6}$$

$$E\left[\hat{\mathcal{R}}_{\mathbf{x}}^{m}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{X})\right] \leq \frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\sqrt{\left\|C\right\|_{HS}+\frac{1}{m}} \tag{7}$$

$$E\left[\hat{\mathcal{R}}_{\mathbf{x}}^{m}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{X})\right] \leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{\sqrt{n}}\left(\left\|C\right\|_{HS}^2+\frac{3}{m}\right)^{1/4}. \tag{8}$$

Proof Fix $\mathbf{x}$ and define vectors $w_l = w_l(\sigma) = \sum_{i=1}^{n}\sigma_i^l x_i^l$ depending on the Rademacher variables σ_i^l. Then by Lemma 5 and Jensen's inequality

$$\begin{aligned}\hat{\mathcal{R}}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) &= E_\sigma\left[\sup_{T\in\mathcal{A}}\sup_{v_1,\ldots,v_m,\left\|v_i\right\|\leq B}\frac{2}{nm}\sum_{l=1}^{m}\left\langle Tw_l, v_l\right\rangle\right] \\ &\leq \frac{2B}{nm}\left\|\mathcal{A}\right\|_{HS}E_\sigma\left[\left(\sum_{l,r}\left|\left\langle w_l, w_r\right\rangle\right|\right)^{1/2}\right] \\ &\leq \frac{2B}{nm}\left\|\mathcal{A}\right\|_{HS}\left(\sum_{l,r}E_\sigma\left[\left|\left\langle w_l, w_r\right\rangle\right|\right]\right)^{1/2}.\end{aligned} \tag{9}$$

Now we have

$$E_\sigma\left[\left\|w_l\right\|^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}E_\sigma\left[\sigma_i^l\sigma_j^l\right]\left\langle x_i^l, x_j^l\right\rangle = \sum_{i=1}^{n}\left\|x_i^l\right\|^2. \tag{10}$$

Also, for $l \neq r$, we get, using Jensen's inequality and independence of the Rademacher variables,

$$\begin{aligned}\left(E_\sigma\left[\left|\left\langle w_l, w_r\right\rangle\right|\right]\right)^2 &\leq E_\sigma\left[\left\langle w_l, w_r\right\rangle^2\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{i'=1}^{n}\sum_{j'=1}^{n}E_\sigma\left[\sigma_i^l\sigma_j^r\sigma_{i'}^l\sigma_{j'}^r\right]\left\langle x_i^l, x_j^r\right\rangle\left\langle x_{i'}^l, x_{j'}^r\right\rangle \\ &= \sum_{i.j=1}^{n}\left\langle x_i^l, x_j^r\right\rangle^2.\end{aligned} \tag{11}$$

Taking the square-root and inserting it together with (10) in (9) we obtain the following intermediate bound

$$\hat{\mathcal{R}}_{\mathbf{x}}^{m}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) \leq \frac{2B\left\|\mathcal{A}\right\|_{HS}}{nm}\left(\sum_{l=1}^{m}\sum_{i=1}^{n}\left\|x_i^l\right\|^2+\sum_{l\neq r}\left(\sum_{i,j=1}^{n}\left\langle x_i^l, x_j^r\right\rangle^2\right)^{1/2}\right)^{1/2} \tag{12}$$

By Jensen's inequality we have

$$\frac{1}{m^2}\sum_{l\neq r}\left(\sum_{i,j=1}^{n}\left\langle x_i^l, x_j^r\right\rangle^2\right)^{1/2} \leq \left(\frac{1}{m^2}\sum_{l,r=1}^{m}\sum_{i,j=1}^{n}\left\langle x_i^l, x_j^r\right\rangle^2\right)^{1/2} = \frac{1}{m}\left\|\hat{C}(\mathbf{x})\right\|_{Fr},$$

which together with (12) and $\left\|x_i^l\right\| \leq 1$ implies (5).

To prove (6) first use the second part of Lemma 5 and Jensen's inequality to get

$$\hat{\mathcal{R}}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) \leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}} \left(\sum_{l,r} E_\sigma\left[\left\langle w_l, w_r\right\rangle^2\right]\right)^{1/4}. \tag{13}$$

Now we have $E_\sigma\left[\sigma_i^l \sigma_j^l \sigma_{i'}^l \sigma_{j'}^l\right] \leq \delta_{ij}\delta_{i'j'} + \delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{ji'}$ so

$$\begin{aligned} E_\sigma\left[\left\langle w_l, w_l\right\rangle^2\right] &\leq \sum_{i,j=1}^n \left(\left\|x_i^l\right\|^2 \left\|x_j^l\right\|^2 + 2\left\langle x_i^l, x_j^l\right\rangle^2\right) \\ &\leq 2\left(\sum_{i=1}^n \left\|x_i^l\right\|^2\right)^2 + \sum_{i,j=1}^n \left\langle x_i^l, x_j^l\right\rangle^2 \leq 2n^2 + \sum_{i,j=1}^n \left\langle x_i^l, x_j^l\right\rangle^2, \end{aligned}$$

where we used $\left\|x_i^l\right\| \leq 1$. Inserting this together with (11) in (13) gives

$$\begin{aligned} \hat{\mathcal{R}}_{\mathbf{u}}^m\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) &\leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}} \left(\sum_{l,r:l\neq r} E_\sigma\left[\left\langle w_l, w_r\right\rangle^2\right] + \sum_{l=1}^m E_\sigma\left[\left\langle w_l, w_l\right\rangle^2\right]\right)^{1/4} \\ &\leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}} \left(\sum_{l,r=1}^m \sum_{i,j=1}^n \left\langle x_i^l, x_j^r\right\rangle^2 + 2mn^2\right)^{1/4}, \end{aligned} \tag{14}$$

which is (6).

Taking the expectation of (12), using Jensen's inequality, $\left\|X^l\right\| \leq 1$ and independence of X^l and X^r for $l \neq r$, and Jensen's inequality again, we get

$$\begin{aligned} &E\left[\hat{\mathcal{R}}_{\mathbf{u}}^m\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{X})\right] \\ &\leq \frac{2B\left\|\mathcal{A}\right\|_{HS}}{nm}\left(nm + \sum_{l\neq r}\left(E\left[\sum_{i,j=1}^n \left\langle X_i^l, X_j^r\right\rangle^2\right]\right)^{1/2}\right)^{1/2} \\ &= \frac{2B\left\|\mathcal{A}\right\|_{HS}}{nm}\left(\sum_{l\neq r}\left(E\left[\left\langle \sum_{i=1}^n Q_{X_i^l}, \sum_{j=1}^n Q_{X_j^r}\right\rangle_{HS}\right]\right)^{1/2} + nm\right)^{1/2} \\ &= \frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\left(\frac{1}{m^2}\sum_{l\neq r}\left\langle E\left[Q_{X^l}\right], E\left[Q_{X^r}\right]\right\rangle_{HS}^{1/2} + \frac{1}{m}\right)^{1/2} \\ &\leq \frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\left(\left\langle \frac{1}{m}\sum_{l=1}^m E\left[Q_{X^l}\right], \frac{1}{m}\sum_{r=1}^m E\left[Q_{X^r}\right]\right\rangle_{HS}^{1/2} + \frac{1}{m}\right)^{1/2}, \end{aligned}$$

which gives (7). In a similar way we obtain from (14)

$$\begin{aligned}
&E\left[\hat{\mathcal{R}}_{\mathbf{a}}^{m}\left(\mathcal{F}_B\circ\mathcal{A}\right)(\mathbf{X})\right]\\
&\leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}}\left(mn^2+\sum_{l\neq r}^{m}\sum_{i,j=1}^{n}\left\langle E\left[Q_{X_i^l}\right],E\left[Q_{X_j^r}\right]\right\rangle_{HS}+2mn^2\right)^{1/4}\\
&\leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}}\left(m^2n^2\left\|E\left[\frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}Q_{X_i^l}\right]\right\|_{HS}^{2}+3mn^2\right)^{1/4},
\end{aligned}$$

which gives (8) ■

6. Bounds for Linear Multi-Task Learning

Inserting the bounds of Theorem 12 in Theorem 9 immediately gives

Theorem 13 *Let $\mathcal{A}$ be a be set of bounded, symmetric operators in H and $\varepsilon,\delta\in(0,1)$*

(i) With probability greater than $1-\delta$ it holds for all $\mathbf{f}=(v_1,...,v_m)\circ T\in\mathcal{F}_B\circ\mathcal{A}$ and all $\gamma\in(0,1)$ that

$$er\left(\mathbf{h}_{\mathbf{f}}\right)\leq e\hat{r}_{\gamma}\left(\mathbf{f}\right)+\frac{1}{\gamma(1-\varepsilon)}A+\sqrt{\frac{\ln\left(1/\left(\delta\gamma\varepsilon\right)\right)}{2nm}},$$

where A is either

$$A=\frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\sqrt{\left\|C\right\|_{HS}+\frac{1}{m}} \tag{15}$$

or

$$A=\frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{\sqrt{n}}\left(\left\|C\right\|_{HS}^{2}+\frac{3}{m}\right)^{1/4}. \tag{16}$$

(ii) With probability greater than $1-\delta$ it holds for all $\mathbf{f}=(v_1,...,v_m)\circ T\in\mathcal{F}_B\circ\mathcal{A}$ and for all $\gamma\in(0,1)$ that

$$er\left(\mathbf{h}_{\mathbf{f}}\right)\leq e\hat{r}_{\gamma}\left(\mathbf{f}\right)+\frac{1}{\gamma(1-\varepsilon)}A\left(\mathbf{X}\right)+\sqrt{\frac{9\ln\left(2/\left(\delta\gamma\varepsilon\right)\right)}{2nm}},$$

where the random variable $A(\mathbf{X})$ is either

$$A\left(\mathbf{X}\right)=\frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\sqrt{\frac{1}{mn}\left\|\hat{C}\left(\mathbf{x}\right)\right\|_{Fr}+\frac{1}{m}}$$

or

$$A\left(\mathbf{X}\right)=\frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{\sqrt{n}}\left(\left(\frac{1}{mn}\left\|\hat{C}\left(\mathbf{x}\right)\right\|_{Fr}\right)^{2}+\frac{2}{m}\right)^{1/4}.$$

We finally extend this result from uniformly bounded sets $\mathcal{A}$ of operators to the set HS^* of all symmetric Hilbert-Schmidt operators. This is done following the techniques described in (Anthony, Bartlett, 1999), using the following lemma (a copy of Lemma 15.5 from this reference):

Lemma 14 *Suppose*

$$\{F(\alpha_1,\alpha_2,\delta) : 0 < \alpha_1,\alpha_2,\delta \leq 1\}$$

is a set of events such that:

(i) For all $0 < \alpha \leq 1$ *and* $0 < \delta \leq 1$,

$$\Pr\{F(\alpha,\alpha,\delta)\} \leq \delta.$$

(ii) For all $0 < \alpha_1 \leq \alpha \leq \alpha_2 \leq 1$ *and* $0 < \delta_1 \leq \delta \leq 1$,

$$F(\alpha_1,\alpha_2,\delta_1) \subseteq F(\alpha,\alpha,\delta).$$

Then for $0 < a,\delta < 1$,

$$\Pr\left(\bigcup_{\alpha\in(0,1]} F(\alpha a,\alpha,\delta\alpha(1-a))\right) \leq \delta.$$

Applications of this lemma follow a standard pattern, as explained in detail in (Anthony, Bartlett, 1999). Let ε,δ,B be as in the previous theorem. For $\alpha \in (0,1]$ set

$$\mathcal{A}(\alpha) = \{T \in HS^* : \|T\|_{HS} \leq 1/\alpha\}$$

and consider the events

$$F(\alpha_1,\alpha_2,\delta) = \{\exists \mathbf{f} \in \mathcal{F}_B \circ \mathcal{A}(\alpha_2) \text{ such that}$$
$$\operatorname{er}(\mathbf{h_f}) > \hat{\operatorname{er}}_\gamma(\mathbf{f}) + \frac{2B}{\alpha_1\gamma(1-\varepsilon)\sqrt{n}}\sqrt{\|C\|_{HS}+\frac{1}{m}} + \sqrt{\frac{\ln(1/(\delta\gamma\varepsilon))}{2nm}}\Bigg\}.$$

By the first conclusion of Theorem 13 the events $F(\alpha_1,\alpha_2,\delta)$ satisfy hypothesis (i) of Lemma 14, and it is easy to see that (ii) also holds. If we set $a = 1-\varepsilon$ and replace α by $1/\|T\|_{HS}$, then the conclusion of Lemma 14 reads as follows:

With probability greater $1-\delta$ it holds for every $\mathbf{f} = (v_1,...,v_m)\circ T$ with $(v_1,...,v_m) \in \mathcal{F}_B$ and $T \in HS^*$ with $\|T\|_{HS} \geq 1$ and all $\gamma \in (0,1)$ that

$$\operatorname{er}(\mathbf{h_f}) \leq \hat{\operatorname{er}}_\gamma(\mathbf{f}) + \frac{2B\|T\|_{HS}}{\gamma(1-\varepsilon)^2\sqrt{n}}\sqrt{\|C\|_{HS}+\frac{1}{m}} + \sqrt{\frac{\ln\left(\frac{\|T\|_{HS}}{\delta\gamma\varepsilon^2}\right)}{2nm}}.$$

Applying the same technique to the other conclusions of Theorem 13 gives the following result, which we state in abbreviated fashion:

Theorem 15 *Theorem 13 holds with the following modifications:*

- *The class* $\mathcal{F}_B \circ \mathcal{A}$ *is replaced by all* $\mathbf{f} = (v_1,...,v_m)\circ T \in \mathcal{F}_B \circ HS^*$ *with* $\|T\|_{HS} \geq 1$ *(or* $\left\|T^2\right\|_{HS} \geq 1$*).*
- $\|\mathcal{A}\|_{HS}$ *(or* $\left\|\mathcal{A}^2\right\|_{HS}$*) is replaced by* $\|T\|_{HS}$ *(or* $\left\|T^2\right\|_{HS}$*).*
- $(1-\varepsilon)$ *and* $1/(\delta\gamma\varepsilon)$ *are replaced by* $(1-\varepsilon)^2$ *and* $\|T\|_{HS}/\left(\delta\gamma\varepsilon^2\right)$ *(or* $\left\|T^2\right\|_{HS}^{1/2}/\left(\delta\gamma\varepsilon^2\right)$*) respectively.*

The requirement $\|T\|_{HS} \geq 1$ (or $\left\|T^2\right\|_{HS} \geq 1$) is an artifact introduced by the stratification lemma 14. Setting $\varepsilon = 1/2$ and $B = 1$ gives Theorem 1 and Theorem 2.

7. An Example

We conclude with an example illustrating the behavior of our bounds when learning from noisy data. We start with a fairly generic situation and subsequently introduce several idealized assumptions.

Suppose that $\left(X^l, Y^l\right)_{l=1}^m$ are random variables modelling a multi-task problem as above. In the following let M be the smallest closed subspace $M \subset H$ such that $X^l \in M$ a.s..

We now mix the data variables X^l with noise, modeled by a random variable X^N with values in H, such that $\left\|X^N\right\| \leq 1$ and $E\left[X^N\right] = 0$. The mixture is controlled by a parameter $s \in (0,1]$ and replaces the original data-variables X^l with the contaminated random variable $\hat{X}^l = sX^l + (1-s)X^N$. We now make the assumption that

- X^N is independent of X^l and Y^l for all l.

Let us call s the signal amplitude and $1-s$ the noise amplitude. The case $s = 1$ corresponds to the original multi-task problem. Decreasing s, and adding more noise, clearly makes learning more difficult up to the case $s = 0$ (which we exclude), where the data variables become independent of the labels and learning becomes impossible. We will look at the behavior of both of our bounds for non-interacting (single-task) and interacting (multi-task) learners as we decrease the signal amplitude s. The bounds which we use are implied by Lemma 11 and Theorem 9 for the non-interacting case and Theorem 13 for the interacting case, and state that each of the following two statements holds with probability at least $1-\delta$:

1. **Non-interacting bound**. $\forall \mathbf{v} \in \mathcal{F}_1$, $\forall \gamma$,
   $\mathrm{er}(\mathbf{h}_\mathbf{v}) \leq \hat{\mathrm{er}}_\gamma(\mathbf{v}) + \frac{2}{\gamma(1-\varepsilon)\sqrt{n}} + \sqrt{\frac{\ln(1/(\delta\gamma\varepsilon))}{2nm}}$
2. **Interacting bound**. $\forall \mathbf{v} \circ T \in \mathcal{F}_1 \circ \mathcal{A}$, $\forall \gamma$,
   $\mathrm{er}(\mathbf{h}_\mathbf{v} \circ T) \leq \hat{\mathrm{er}}_\gamma(\mathbf{v} \circ T) + \frac{2\|\mathcal{A}\|_{HS}}{\gamma(1-\varepsilon)\sqrt{n}} \sqrt{\|C\|_{HS} + \frac{1}{m}} + \sqrt{\frac{\ln(1/(\delta\gamma\varepsilon))}{2nm}}$

The first damage done by decreasing s is that the margin γ must be also decreased to $s\gamma$ to obtain a comparable empirical margin error for the mixed problem as for the original problem. Replacing γ by γs is very crude and normally insufficient, because of interfering noise, but the replacement can be justified if one is willing to accept the orthogonality assumption:

- $X^N \perp M$ a.s.

The assumption that the noise to be mixed in is orthogonal to the signal is somewhat artificial. We will later assume that the dimension d of the signal space M is small and that X^N is distributed homogeneously on a high-dimensional sphere. This implies weak orthogonality in the sense that $\left\langle X^l, X^N \right\rangle^2$ is small with high probability, a statement which could also be used, but at the expense of considerable complications. To immediately free us from consideration of the empirical term (and only for this purpose), we make the orthogonality assumption. By projecting to M we can then find for any sample $\left(X_i^l, Y_i^l\right)$ and any preprocessor T and any multi-classifying vector $\mathbf{v}$ some T' and $\mathbf{v}'$ such that $\hat{\mathrm{er}}_{\gamma s}(\mathbf{v}')$ and $\hat{\mathrm{er}}_{\gamma s}(\mathbf{v}' \circ T')$ for the mixed sample $\left(\hat{X}_i^l, Y_i^l\right)$ are the same as $\hat{\mathrm{er}}_\gamma(\mathbf{v})$ and $\hat{\mathrm{er}}_\gamma(\mathbf{v} \circ T)$ for the original sample. We can therefore regard the empirical terms as equal in both bounds and for all values of s as long as $\mathcal{A}$ is stable under projection to M and γ is replaced by γs.

This will cause a logarithmic penalty in the last term depending on the confidence parameter δ, but we will neglect this entire term on the grounds of rapid decay with the product nm. The remaining term, which depends on s and is different for both bounds, is then

$$\frac{2}{\gamma s(1-\varepsilon)\sqrt{n}} \tag{17}$$

for the non-interacting and

$$\frac{2\|\mathcal{A}\|_{HS}}{\gamma s(1-\varepsilon)\sqrt{n}}\sqrt{\|C_s\|_{HS}+\frac{1}{m}} \tag{18}$$

for the interacting case. Here $C_s = (1/m)\sum_l E\left[Q_{sX^l+(1-s)X^N}\right]$ is the total covariance operator for the noisy mixture problem. By independence of X^N and X^l and the mean-zero assumption for X^N we obtain from Lemma 4 and 7 that

$$C_s = (1/m)\sum_l \left(s^2 E\left[Q_{X^l}\right] + (1-s)^2 E\left[Q_{X^N}\right]\right) = s^2 C + (1-s)^2 E\left[Q_{X^N}\right],$$

where C would be the total covariance operator for the original problem. To bound the HS-norm of this operator we now introduce a simplifying assumption of homogeneity for the noise distribution:

- X^N is distributed uniformly on a k-dimensional unit-sphere centered at the origin.

This implies that $\|E\left[Q_{X^N}\right]\|_{HS} = 1/\sqrt{k}$ so that

$$\|C_s\|_{HS}^2 \leq s^2\|C\|_{HS} + \|E\left[Q_{X^N}\right]\|_{HS} \leq s^2\|C\|_{HS} + 1/\sqrt{k},$$

and substitution in (18) gives the new term

$$\frac{2\|\mathcal{A}\|_{HS}}{\gamma(1-\varepsilon)\sqrt{n}}\sqrt{\|C\|_{HS}+\frac{1}{s^2}\left(\frac{1}{\sqrt{k}}+\frac{1}{m}\right)} \tag{19}$$

for the interacting bound. The dependence on the inverse signal amplitude in the first factor has disappeared, and as k and m increase, the bound for the noisy problem tends to the same limiting value

$$\frac{2\|\mathcal{A}\|_{HS}\|C\|_{HS}^{1/2}}{\gamma(1-\varepsilon)\sqrt{n}}$$

as the bound for the original 'noise free' problem, for any fixed positive value of s. This contrasts the behavior of all bounds which depend linearly on the dimension of the input space (such as in) and diverge as $k \to \infty$.

The quotient of (19) to the non-interacting (17) is

$$\|\mathcal{A}\|_{HS}\sqrt{s^2\|C\|_{HS}+\frac{1}{\sqrt{k}}+\frac{1}{m}},$$

and the interacting bound will be better than the non-interacting bound whenever this expression is less than unity. This is more likely to happen when the signal amplitude s is small, and the dimension k of the noise distribution and the number of tasks m are large.

An intuitive explanation of the fact, that for multi-task learning a large dimension k of the noise-distribution has a positive effect on the bound, is that for large k a sample of homogeneously distributed random unit vectors is less likely to lie in a common low-dimensional subspace, a circumstance which could mislead the multi-task learner.

Of course there are many situations, when multi-task learning doesn't give any advantage over single-task learning. To make a quantitative comparison we make more three more simplifying assumptions on the data distribution of the X^l:

- $\dim(M) = d < \infty$

The signal space M is of course unknown to the learner, but we assume that

- we know its dimension d.

Multi-task learning can then select from an economically chosen set $\mathcal{A} \subset HS^*$ of preprocessors such that $\mathcal{A}$ contains the set of d-dimensional projections $\mathcal{P}_d$ and $\|\mathcal{A}\|_{HS} = \sqrt{d}$. We assume knowledge of d mainly for simplicity, without it we could invoke Theorem 15 instead of Theorem 13 above, causing some complications, which we seek to avoid.

- The mixture of the distributions of the X^l is homogeneous on $S_1 \cap M$.

This implies $\|C\|_{HS} = 1/\sqrt{d}$, and, with $\|\mathcal{A}\|_{HS} = \sqrt{d}$, the multi-task bound will improve over the non-interacting bound if

$$\sqrt{d}s^2 + \frac{d}{\sqrt{k}} + \frac{d}{m} < 1.$$

From this condition we conclude with four cook-book-rules to decide when it is worthwhile to go through the computational trouble of multi-task learning instead of the simpler single-task learning.

1. The problem is very noisy (s is expected to be small)
2. The noise is high-dimensional (k is expected to be large)
3. There are many learning tasks (m is large)
4. We suspect that the relevant information for all m tasks lies in a low-dimensional (d is small)

If one believes these criteria to be met, then one can use an algorithm as the one developed in (Ando, Zhang, 2005) to minimize the interacting bound above, with $\mathcal{A} = \mathcal{P}_d$.

Appendix

We give a proof of Theorem 9 for the readers convenience. Most of this material is combined from Anthony and Bartlett (1999), Bartlett and Mendelson (2002), Bartlett et al (2005) and Ando and Zhang (2005), and we make no claim to originality for any of it. A preliminary result is

Theorem 16 *Let $\mathcal{F}$ be a $[0,1]^m$-valued function class on a space $\mathcal{X}$, and $\mathbf{X} = \left(X_i^l\right)_{(l,i)=(1,1)}^{(m,n)}$ a vector of $\mathcal{X}$-valued independent random variables where for fixed l and varying i all the X_i^l are identically distributed. Fix $\delta > 0$. Then with probability greater than $1-\delta$ we have for all $\mathbf{f} = \left(f^1, ..., f^m\right) \in \mathcal{F}$*

$$\frac{1}{m}\sum_{l=1}^{m} E\left[f^l\left(X_1^l\right)\right] \leq \frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n} f^l\left(X_i^l\right) + \mathcal{R}_n^m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2mn}}.$$

We also have with probability greater than $1-\delta$ for all $\mathbf{f} = \left(f^1, ..., f^m\right) \in \mathcal{F}$, that

$$\frac{1}{m}\sum_{l=1}^{m} E\left[f^l\left(X_1^l\right)\right] \leq \frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n} f^l\left(X_i^l\right) + \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) + \sqrt{\frac{9\ln(2/\delta)}{2mn}}.$$

Proof Let Ψ be the function on $\mathcal{X}^{mn}$ given by

$$\Psi(\mathbf{x}) = \sup_{\mathbf{f}\in\mathcal{F}} \frac{1}{m}\sum_{l=1}^{m}\left(E\left[f^l\left(X_1^l\right)\right] - \frac{1}{n}\sum_{i=1}^{n} f^l\left(X_i^l\right)\right)$$

and let $\mathbf{X}'$ be an iid copy of the $\mathcal{X}^{mn}$-valued random variable $\mathbf{X}$. Then

$$\begin{aligned} E[\Psi(\mathbf{X})] &= E_{\mathbf{X}}\left[\sup_{\mathbf{f}\in\mathcal{F}} \frac{1}{mn} E_{\mathbf{X}'}\left[\sum_{l=1}^{m}\sum_{i=1}^{n}\left(f^l\left(\left(X_i^l\right)'\right) - f^l\left(X_i^l\right)\right)\right]\right] \\ &\leq E_{\mathbf{X}\mathbf{X}'}\left[\sup_{\mathbf{f}\in\mathcal{F}} \frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}\left(f^l\left(\left(X_i^l\right)'\right) - f^l\left(X_i^l\right)\right)\right] \\ &= E_{\mathbf{X}\mathbf{X}'}\left[\sup_{\mathbf{f}\in\mathcal{F}} \frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}\sigma_i^l\left(f^l\left(\left(X_i^l\right)'\right) - f^l\left(X_i^l\right)\right)\right], \end{aligned}$$

for any realization $\sigma = \left(\sigma_i^l\right)$ of the Rademacher variables, because the expectation $E_{\mathbf{X}\mathbf{X}'}$ is symmetric under the exchange $\left(X_i^l\right)' \longleftrightarrow X_i^l$. Hence

$$E[\Psi(\mathbf{X})] \leq E_{\mathbf{X}} E_{\sigma}\left[\sup_{\mathbf{f}\in\mathcal{F}} \frac{2}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}\sigma_i^l f^l\left(X_i^l\right)\right] = \mathcal{R}_n^m(\mathcal{F}).$$

Now fix $\mathbf{x} \in \mathcal{X}^{mn}$ and let $\mathbf{x}' \in \mathcal{X}^{mn}$ be as $\mathbf{x}$, except for one modified coordinate $\left(x_i^l\right)'$. Since each f^l has values in $[0,1]$ we have $|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')| \leq 1/(mn)$. So by the one-sided version of the bounded difference inequality (see McDiarmid, 1998)

$$\Pr\left\{\Psi(\mathbf{X}) > E_{\mathbf{X}'}\left[\Psi\left(\mathbf{X}'\right)\right] + \sqrt{\frac{\ln(1/\delta)}{2mn}}\right\} \leq \delta.$$

Together with the above bound on $E[\Psi(\mathbf{X})]$ and the definition of Ψ this gives the first conclusion.

With $\mathbf{x}$ and $\mathbf{x}'$ as above we have $\left|\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{x}) - \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{x}')\right| \leq 2/(mn)$, so by the other tail of the bounded difference inequality

$$\Pr\left\{\mathcal{R}_n^m(\mathcal{F}) < \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) + \sqrt{\frac{4\ln(1/\delta)}{2mn}}\right\} \leq \delta,$$

which, combined with the first conclusion in a union bound, gives the second conclusion. ■

We quote the following folklore theorem (see for example Bartlett et al, 2005) bounding the Rademacher complexity of a function class composed with a fixed Lipschitz function.

Theorem 17 *Let $\mathcal{F}$ be an $\mathbb{R}^m$-valued function class on a space $\mathcal{X}$ and suppose that $\phi : \mathbb{R} \to \mathbb{R}$ has Lipschitz constant L. Let*

$$\phi \circ \mathcal{F} = \left\{ \left(\phi \circ f^1, ..., \phi \circ f^m\right) : \left(f^1, ..., f^m\right) \in \mathcal{F} \right\}.$$

Then

$$\hat{\mathcal{R}}_n^m(\phi \circ \mathcal{F}) \leq L\, \hat{\mathcal{R}}_n^m(\mathcal{F}).$$

Suppose now that $\mathcal{F}$ is an $\mathbb{R}^m$-valued function class on $\mathcal{X}$. For $\mathbf{f} = \left(f^1, ..., f^m\right)$ define functions $\mathbf{f}' = \left(f'^1, ..., f'^m\right)$ and $\mathbf{f}'' = \left(f''^1, ..., f''^m\right)$, from $\mathcal{X} \times \{-1,1\}$ to $\mathbb{R}^m$ or $[0,1]^m$ respectively, by

$$f'^l(x,y) = y f^l(x) \text{ and } f''^l(x,y) = \phi_\gamma \circ f'^l(x,y) = \phi_\gamma(y f(x))$$

and function classes $\mathcal{F}' = \{\mathbf{f}' : \mathbf{f} \in \mathcal{F}\}$ and $\mathcal{F}'' = \{\mathbf{f}'' : \mathbf{f} \in \mathcal{F}\}$. It follows from the definition of $\hat{\mathcal{R}}$ that $\hat{\mathcal{R}}_n^m(\mathcal{F}')(\mathbf{x},\mathbf{y}) = \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{x})$ for all $(\mathbf{x},\mathbf{y}) \in (\mathcal{X} \times \{-1,1\})^{nm}$. Since ϕ_γ is Lipschitz with constant γ^{-1}, the previous theorem implies that

$$\hat{\mathcal{R}}_n^m\left(\mathcal{F}''\right)(\mathbf{X},\mathbf{Y}) \leq \gamma^{-1} \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) \text{ and } \mathcal{R}_n^m\left(\mathcal{F}''\right) \leq \gamma^{-1} \mathcal{R}_n^m(\mathcal{F}). \tag{20}$$

On the other hand, for every $\mathbf{f} = \left(f^1, ..., f^m\right) \in \mathcal{F}$ we have

$$\begin{aligned} \mathrm{er}(\mathbf{h_f}) &= \frac{1}{m} \sum E\left[1_{(-\infty,0]}\left(Y_1^l f^l\left(X_1^l\right)\right)\right] \\ &\leq \frac{1}{m} \sum E\left[\phi_\gamma \circ \left(f'\right)^l \left(X_1^l, Y_1^l\right)\right] \\ &= \frac{1}{m} \sum E\left[\left(f''\right)^l \left(X_1^l, Y_1^l\right)\right] \end{aligned} \tag{21}$$

and

$$\frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} f''^l\left(X_i^l, Y_i^l\right) = \frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \phi_\gamma\left(Y_i^l f^l\left(X_i^l\right)\right) = \hat{\mathrm{er}}_\gamma(\mathbf{f}). \tag{22}$$

Applying Theorem 16 to the class $\mathcal{F}''$ and substitution of (21), (22) and (20) yield

Theorem 18 *Let $\mathcal{F}$ be a $\mathbb{R}^m$-valued function class on a space $\mathcal{X}$, $\gamma \in (0,1)$ and*

$$(\mathbf{X},\mathbf{Y}) = \left(X_i^l, Y_i^l\right)_{(l,i)=(1,1)}^{(m,n)}$$

a vector of $\mathcal{X} \times \{-1,1\}$-valued independent random variables where for fixed l and varying i all the $\left(X_i^l, Y_i^l\right)$ are identically distributed. Fix $\delta > 0$. Then with probability greater than $1-\delta$ we have for all $\mathbf{f} \in \mathcal{F}$

$$er(\mathbf{h_f}) \leq \hat{er}_\gamma(\mathbf{f}) + \gamma^{-1} \mathcal{R}_n^m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2mn}}.$$

We also have with probability greater than $1-\delta$ for all $\mathbf{f} \in \mathcal{F}$, that

$$er(\mathbf{h_f}) \leq \hat{er}_\gamma(\mathbf{f}) + \gamma^{-1} \hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) + \sqrt{\frac{9\ln(2/\delta)}{2mn}}.$$

To arrive at Theorem 9 we still need to convert this into a statement valid with high probability for all margins $\gamma \in (0,1)$. This is done with Lemma 14, which we now apply to the event

$$F(\alpha_1,\alpha_2,\delta) = \left\{ \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } \mathrm{er}(\mathbf{h_f}) > \hat{\mathrm{er}}_{\alpha_2}(\mathbf{f}) + \alpha_1^{-1}\mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2mn}} \right\}.$$

Hypothesis (i) of Lemma 14 follows from the previous theorem, hypothesis (ii) from the fact that the right side in the inequality increases if we decrease δ and α_1 and increase α_2. If we replace a by $1-\varepsilon$ and α by γ, then the conclusion of the lemma becomes the first conclusion of Theorem 9. The second conclusion of Theorem 9 is handled similarly.

The following table is intended as an index and a quick reference to the notation and definitions introduced in the paper.

Notation	Short Description	Section
H	real, separable Hilbert space	2
$\langle .,. \rangle$ and $\|.\|$	inner product and norm on H	2
S_1	unit-sphere in H	3
HS	Hilbert-Schmidt operators on H	2
$\langle .,. \rangle_{HS}$ and $\|.\|_{HS}$	inner product and norm on HS	2
HS^*	symmetric operators in HS	2
$\mathcal{P}_d$	d-dimensional orthogonal projections in H	2
$\mathcal{A}$	a subset of HS^*	2
$\|\mathcal{A}\|_{HS}$	$\sup_{T\in\mathcal{A}} \|T\|_{HS}$	2
$\left\|\mathcal{A}^2\right\|_{HS}$	$\sup_{T\in\mathcal{A}} \left\|T^2\right\|_{HS}$	2
Q_x, for $x \in H$	operator $Q_x z = \langle z,x\rangle x$, $\forall z \in H$	2
$G_{x,y}$, for $x,y \in H$	operator $G_{x,y} z = \langle x,z\rangle y$, $\forall z \in H$	2
$tr(T)$	trace of the operator T	2
$E[Q_X]$	covariance operator of H-valued r.v. X	3
$\mathcal{X}$	generic input space	4
(X^l, Y^l)	random variables for multi-task problem	4
(X_i^l, Y_i^l)	random variables for multi-task sample	4
$\mathrm{er}(\mathbf{h})$	average error of multiclassifier $\mathbf{h}$	4
$\mathbf{h_f}$	multiclassifier obtained by thresholding $\mathbf{f}$	4
ϕ_γ	margin function	4
$\hat{\mathrm{er}}_\gamma(\mathbf{f})$	empirical margin error of vector function $\mathbf{f}$	4
$\hat{\mathcal{R}}_m(\mathcal{F})$	empirical Rademacher complexity	4
C^l	covariance operator for l-th task	5
C	total covariance operator	5
$\hat{C}(\mathbf{X})$	Gramian of data-sample $\mathbf{X}$	3
$\mathcal{F}_B$, $\mathcal{F}_B \circ \mathcal{A}$	fctn. classes for linear multi-task learning	5

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] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.

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

[4] P. Bartlett, O. Bousquet and S. Mendelson. Local Rademacher complexities. Available online: http://www.stat.berkeley.edu/˜bartlett/papers/bbm-lrc-02b.pdf.

[5] P. Baxendale. Gaussian measures on function spaces. *Amer. J. Math.*, 98:891-952, 1976.

[6] J. Baxter. Theoretical Models of Learning to Learn, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998.

[7] J. Baxter. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research* 12: 149-198, 2000.

[8] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *COLT 03*, 2003.

[9] R. Caruana. Multitask Learning, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998.

[10] Nello Cristianini and John Shawe-Taylor. Support Vector Machines. *Cambridge University Press*, 2000.

[11] T. Evgeniou and M. Pontil. Regularized multi-task learning. *Proc. Conference on Knowledge Discovery and Data Mining*, 2004.

[12] 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.

[13] Colin McDiarmid. Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195-248. Springer, Berlin, 1998.

[14] C. A. Micchelli and M. Pontil. Kernels for multi-task learning. Available online, 2005.

[15] S.Mika, B.Schölkopf, A.Smola, K.-R.Müller, M.Scholz and G.Rätsch. Kernel PCA and Denoising in Feature Spaces. *Advances in Neural Information Processing Systems* 11, 1998.

[16] J. Shawe-Taylor, N. Cristianini. Estimating the moments of a random vector. *Proceedings of GRETSI 2003 Conference*, I: 47–52, 2003.

[17] Michael Reed and Barry Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.

[18] S. Thrun. Lifelong Learning Algorithms, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998