Bounds for Linear Multi-Task Learning

Andreas Maurer

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

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.

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 [4], [7], [14], [15]). The technique has been analyzed theoretically and in some generality (see Baxter [5] and Zhang[15]). The purpose of this paper is to improve and clarify some of these theoretical results in the case, when input data is represented in a linear, potentially infinite dimensional space, and the common constraint is a linear preprocessor.

The simplest conceptual model to understand multi-task learning and its potential advantages is perhaps 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 $\mathrm{er}(f)$ and $\hat{\mathrm{er}}(f)$ be the expected error and 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. [1]) 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 $\mathcal{F}$ factors in the following sense: There is a set $\mathcal{Y}$ and a set $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathcal{Y}$ and a set $\mathcal{H}$ of functions $h : \mathcal{Y} \to \{0, 1\}$ such that $\mathcal{F} = \mathcal{H} \circ \mathcal{G}$ and $|\mathcal{H}| \ll |\mathcal{F}|$. Imagine that 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 distribution. We now seek a multiple

solution $h_1 \circ g, \ldots h_m \circ g$ for each of the m task 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 every possible multiple solution $(h_1 \circ g, ..., h_m \circ g)$

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

Here $\mathrm{er}_l\left(f\right)$ and $\hat{\mathrm{er}}_l\left(f\right)$ 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 results (but see Ben-David [6] 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: The contribution coming from estimating the common preprocessor $g \in \mathcal{G}$ decreases with the number of learning tasks. Since by assumption $|\mathcal{H}| \ll |\mathcal{F}|$, 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. This behavior can be intuitively explained by the fact that much more data is being considered in the selection of the common preprocessor $g \in \mathcal{G}$.

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 $\{\mathcal{H} \circ g : g \in \mathcal{G}\}$), and the classifiers for the individual tasks are all chosen from the selected hypothesis space. The two formulations are equivalent: If $\mathbb{F}$ is a class of hypothesis spaces $\mathcal{F}$ of functions $f : \mathcal{X} \to \{0,1\}$ define $\mathcal{Y} = \mathbb{F} \times \mathcal{X}$ and

$$\mathcal{G} = \{x \in \mathcal{X} \mapsto (\mathcal{F}, x) \in \mathbb{F} \times \mathcal{X} : \mathcal{F} \in \mathbb{F}\}$$

and $\mathcal{H} = \{(\mathcal{F}, x) \in \mathbb{F} \times \mathcal{X} \mapsto f\left(x\right) : f \in \mathcal{F}\}$. Then choosing $\mathcal{F}$ from $\mathbb{F}$ and f from F is equivalent to choosing g from $\mathcal{G}$ and h from $\mathcal{H}$ and using the hypothesis $f = h \circ g$. Here we prefer the formulation of selecting a preprocessor instead of a hypothesis space, because it is more intuitive in the situations which we consider.

Using covering numbers the arguments leading to (2) can be extended to certain infinite classes to give general bounds for multi-task learning ([5] and [15]). 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 of 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 $\|v\| \leq B$. This contains the case

considered by Zhang ([15]) where $\mathcal{G} = \mathcal{P}_d$, the set of orthogonal projections in H with d-dimensional range. We will prove the following :

Theorem 1. *Let $\delta \in (0,1)$. With probability greater than $1-\delta$ 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*

$$\frac{1}{m}\sum_{l=1}^{m} er\left(h_{v_l} \circ T\right) \leq \frac{1}{m}\sum_{l=1}^{m} \hat{er}_{\gamma}\left(v_l \circ T\right) + \frac{8\|T\|_{HS}}{\gamma\sqrt{n}}\sqrt{\|C\|_{HS} + \frac{1}{m}} + \sqrt{\frac{\ln\frac{8}{\delta\gamma}}{2nm}}.$$

Here $\hat{\text{er}}_{\gamma}\left(v_l \circ T\right)$ is an *empirical margin error*, bounded by the relative number of examples $\left(X_i^l, Y_i^l\right)$ in the training sample for task l, when $Y_i^l \left\langle TX_i^l, v_l \right\rangle < \gamma$. The quantity $\|T\|_{HS}$ is the *Hilbert-Schmidt norm* the operator 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). C is *the total covariance operator* corresponding to the mixture of all the task-input-distributions in H. The above Theorem is the simplest, but not the tightest or most general form of our main 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.

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 is 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. The fact that our bounds are dimension free (in contrast to those in [15], for example) allows their general use for multi-task learning in kernel-induced Hilbert spaces (see [8]).

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 'university-limit' $m \to \infty$ it is preferable to use a different bound (see Theorems 6 and 7), at the expense of slower convergence in m. The second term on the right is then replaced by

$$\frac{(2+\epsilon)\left\|T^2\right\|_{HS}^{1/2}}{\gamma\sqrt{n}}\left(\|C\|_{HS}^2 + \frac{3}{m}\right)^{1/4}$$

for some small ϵ. If we let T be an orthogonal projection with d-dimensional range then $\left\|T^2\right\|_{HS}^{1/2} = d^{1/4}$, so for large m the above expression 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 [2] or Lemma 1 below) would give $2/\left(\gamma\sqrt{n}\right)$ for this term, if all data is constrained to unit vectors. We therefore suspect an improved estimation for multi-task learning with large m whenever $d^{1/4}\left\|C\right\|_{HS}^{1/2} \ll 1$. If we 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 theoretical 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 3 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*

$$\frac{1}{m}\sum_{l=1}^{m} er\left(h_{v_l} \circ T\right) \leq \frac{1}{m}\sum_{l=1}^{m} e\hat{r}_\gamma\left(v_l \circ T\right) + \frac{8\left\|T\right\|_{HS}}{\gamma\sqrt{n}}\sqrt{\frac{1}{mn}\left\|\hat{C}\left(\mathbf{X}\right)\right\|_{Fr} + \frac{1}{m}} + \sqrt{\frac{9\ln\frac{8}{\delta\gamma}}{2nm}}.$$

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 the covariance operator of random elements in a Hilbert space. Section 3 gives a formal definition of multi-task systems and a general PAC bound in terms of Rademacher complexities. Here we follow the path prepared by Kolchinskii and Panchenko ([9]) and Bartlett and Mendelsson ([2]). Our result is taken from [2] and adapted to vector valued function classes. In section 4 we study the Rademacher complexities of linear multi-task systems. In section 5 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.

2 Hilbert-Schmidt and covariance operators

For a fixed real, separable Hilbert space H we define a second Hilbert space consisting of *Hilbert-Schmidt operators*. With HS we denote the real vector space of operators on H satisfying $\sum_{i=1}^{\infty} \|Te_i\|^2 \leq \infty$ for every orthonormal basis $(e_i)_{i=1}^{\infty}$ of H. 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 $\|.\|_{HS}$ (see Reed and Simon [13] 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.

For every $v \in H$ we define an operator Q_v by $Q_v w = \langle w, v \rangle v$. For $v \neq 0$ chose an orthonormal basis $(e_i)_1^{\infty}$, so that $e_1 = v/\|v\|$. Then

$$\|Q_v\|_{HS}^2 = \sum_i \|Q_v e_i\|^2 = \|Q_v v\|^2 / \|v\|^2 = \|v\|^4 ,$$

so $Q_v \in HS_0$ and $\|Q_v\|_{HS} = \|v\|^2$. With the same basis we have for any $T \in HS$

$$\langle T, Q_v \rangle_{HS} = \sum_i \langle Te_i, Q_v e_i \rangle = \langle Tv, Q_v v \rangle / \|v\|^2 = \langle Tv, v \rangle .$$

This gives the simple, but useful formulas

$$\langle Q_v, Q_w \rangle_{HS} = \langle v, w \rangle^2 , \tag{3}$$

and, for $T \in HS^*$ we then have

$$\left\langle T^2, Q_v \right\rangle_{HS} = \|Tv\|^2 . \tag{4}$$

For $x, y \in H$ define an operator $G_{x,y}$ by $G_{x,y} z = \langle x, z \rangle y$. Let $(e_k)_{k=1}^{\infty}$ be an orthonormal basis. We have

$$\langle Tx, y \rangle = \left\langle T \sum_k \langle x, e_k \rangle e_k, y \right\rangle = \sum_k \langle Te_k, \langle x, e_k \rangle y \rangle = \sum_k \langle Te_k, G_{x,y} e_k \rangle = \langle T, G_{x,y} \rangle_{HS} . \tag{5}$$

Also note that

$$\langle G_{x,y}, G_{x',y'} \rangle_{HS} = \sum_k \langle \langle x, e_k \rangle y, \langle x', e_k \rangle y' \rangle = \langle y, y' \rangle \sum_k \langle x, e_k \rangle \langle x', e_k \rangle = \langle x, x' \rangle \langle y, y' \rangle , \tag{6}$$

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 $P^{1/2} = P$.

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

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

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

Suppose that $E\left[\|X\|^2\right] \leq \infty$. Passing to the space of Hilbert-Schmidt operators the same construction can be carried out again: We then have $E[\|Q_X\|_{HS}] = E\left[\|X\|^2\right] \leq \infty$, so there is a unique operator $E[Q_X] \in HS_0$ such that $E[\langle Q_X, T \rangle_{HS}] = \langle E[Q_X], T \rangle_{hs}, \forall T \in HS$, with $\|E[Q_X]\|_{HS} \leq E[\|Q_X\|_{HS}]$. Let $y, z \in H$. Then using (5) above we get

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

The operator $E[Q_X]$ is called the *covariance operator* of X. It is positive and one easily verifies

$$tr(E[Q_X]) = E\left[\|X\|^2\right].$$

In particular $\|E[Q_X]\|_{HS} \leq E[\|Q_X\|_{HS}] = tr(E[Q_X])$. Notice that for any two X_1 and X_2 we have by (3) that

$$\langle E[Q_{X_1}], E[Q_{X_1}] \rangle_{HS} = E[\langle Q_{X_1}, Q_{X_1} \rangle_{HS}] = E\left[\langle X_1, X_2 \rangle^2\right] \geq 0.$$

Theorem 3. *Suppose that $X_1, ..., X_m$ are independent random variables in H with $\|X_i\| \leq 1$. Then for all $\delta > 0$ with probability greater than δ we have*

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

This result is proved by applying the corresponding result for vector valued random variables (see Shawe-Taylor [12]) to the HS-valued random variables Q_{X_i}. If the X_i are identically distributed, it shows that the covariance operator can be estimated with high probability by the empirical approximations $(1/m)\sum_i Q_{X_i}$. 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 matrix $G_{ij} = \langle X_i, X_j\rangle$.

3 Multi-task systems and general bounds

For our discussion of multi-task learning we concentrate on binary labeled data. We assume that there are m independent random variables $Z^l = \left(X^l, Y^l\right)$, where

- $l \in \{1, ..., m\}$ identifies one of the m learning tasks,
- $X^l \in H$ models the input data of the l-th task, distributed in a real, separable Hilbert-space H with inner product $\langle ., .\rangle$ and norm $\|.\|$, satisfying $\left\|X^l\right\| \leq 1$. We write C_l for the covariance operator $E\left[Q_{X^l}\right]$ and C for the total covariance operator $C = (1/m)\sum_l C_l$.
- $Y^l \in \{-1, 1\}$ 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 $\mathbf{Z} = \left(\left(Z_i^l\right)_{i=1}^n\right)_{l=1}^m$ is called the *training sample* or training data. We also write $\mathbf{X} = \left(\left(X_i^l\right)_{i=1}^n\right)_{l=1}^m$.

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

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

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 $\mathbf{f} : H \to \mathbb{R}^m$ with $\mathbf{f} = (f_1, ..., f_m)$. A function $\mathbf{f} \in \mathcal{F}$ defines a multi-classifier $\mathbf{h_f}$ through $\left(\mathbf{h_f}\left(x\right)\right)_l = \mathrm{sign}(f_l\left(x\right))$. To give uniform error bounds for such classifiers in terms of empirical estimates, we define for $\gamma > 0$ the margin functions

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

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

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

called the *empirical γ-margin error* of $\mathbf{f}$. The following Theorem (taken from [2] with minor modifications and combined with the model-selection lemma 15.5 in Anthony/Bartlett [1]) gives a bound on $\mathrm{er}(\mathbf{h_f})$ in terms of $\hat{\mathrm{er}}_\gamma(\mathbf{f})$, valid with high probability uniformly in $\mathbf{f}$ and γ.

Theorem 4. *Let $\epsilon, \delta \in (0,1)$*

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

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

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

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

Here $\hat{\mathcal{R}}_n^m(\mathcal{F})$ and $\mathcal{R}_n^m(\mathcal{F})$ are the *empirical and expected Rademacher averages* in the sense of the following

Definition 1. *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 average is the random variable*

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

The expected Rademacher average is

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

The bounds in Theorem 4 each involve three terms. The last one expresses the dependence of the estimation error on the confidence parameter δ and a model-selection penalty $(\ln(1/(\gamma\epsilon))$ 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 [6] 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 4 can be chosen very small, so that the factor $1/(1-\epsilon)$ 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 multiclassifier $\mathbf{h}_{\mathbf{f}^*}$, 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 are the empirical and expected Rademacher complexities $\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X})$ and $\mathcal{R}_n^m(\mathcal{F})$, measuring the complexity of the function class $\mathcal{F}$, and bounds on these quantities can be substituted in Theorem 4.

4 The Rademacher complexity of linear multi-task learning

We will now concentrate on multi-task learning in the linear case, when the data lives in a Hilbert space, by means of some kernel induced-embedding (see [8]), the details of which will not concern us at this point. As stated above we assume that $X^l \in H$ for all $l \in \{1, ..., m\}$ with $\left\|X^l\right\| \leq 1$. The case $\left\|X^l\right\| = 1$ where the data is constrained to the unit sphere in H is of particular interest, corresponding to a class of radial basis function kernels.

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 chose 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 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 (possible based on experience with other data) was lucky enough to allow for a sufficiently low empirical error.

The noninteracting 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 involves only a minimal modification of the proof of Lemma 22 in [2].

Lemma 1. *We have*

$$\hat{\mathcal{R}}_n^m\left(\mathcal{F}_B\right)(\mathbf{X}) \leq \frac{2B}{nm}\sum_{l=1}^{m}\left(\sum_{i=1}^{n}\left\|X_i^l\right\|^2\right)^{1/2}$$

$$\mathcal{R}_n^m\left(\mathcal{F}_B\right) \leq \frac{2B}{m}\sum_{l=1}^{m}\left(\frac{E\left[\left\|X^l\right\|^2\right]}{n}\right)^{1/2} = \frac{2B}{m}\sum_{l=1}^{m}\left(\frac{tr\left(C_l\right)}{n}\right)^{1/2}$$

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

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

Jensen's inequality gives the second conclusion □

This is just the average of the bounds given by Bartlett and Mendelson in [2] 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 becomes $2B/\sqrt{n}$, 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

$$\hat{\mathcal{R}}_n^m\left(\mathcal{F}_B \circ T\right)(\mathbf{X}) = \hat{\mathcal{R}}_n^m\left(\mathcal{F}_B\right)(T\mathbf{X}) \leq \frac{2B}{nm}\sum_{l=1}^{m}\left(\sum_{i=1}^{n}\left\|TX_i^l\right\|^2\right)^{1/2}, \qquad (7)$$

which can be significantly smaller than $B/\sqrt{n}$, because $\|Tx\| \leq \|x\|, \forall x$, the extent depending on how strongly T annihilates the vectors in the sample. There are reasonable limits to this annihilation: Suppose that for some (small) margin γ a multiclassifier $(v_1, ..., v_m)$ gives an empirical margin error of zero on the transformed data, i.e. $y_i^l\left\langle TX_i^l, v_l\right\rangle \geq \gamma$ for all l and i. Then by Schwartz' inequality

we must have $\left\|TX_i^l\right\| \geq \gamma/B$. If we are lucky however, or by divine inspiration, we may chose T such that $\left\|TX_i^l\right\|$ is much smaller than unity on average. The art of the engineer is to select a preprocessor T which annihilates the irrelevant components to improve estimation and furthers the relevant ones to improve the result of empirical optimization.

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 $v_1, ..., v_l$. We give two bounds, one which is somewhat similar to other bounds for multi-task learning (e.g. (2)) and another one which is tighter in the limit when the number of tasks m goes to infinity.

Theorem 5. *We have*

$$\hat{\mathcal{R}}_n^m\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{X}) \leq \frac{2B\left\|\mathcal{A}\right\|_{HS}}{nm}\left(\sum_{l=1}^m\sum_{i=1}^n\left\|X_i^l\right\|^2 + \sum_{l\neq r}\left(\sum_{i,j=1}^n\left\langle X_i^l, X_j^r\right\rangle^2\right)^{1/2}\right)^{1/2} \tag{8}$$

and

$$\hat{\mathcal{R}}_n^m\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{X}) \leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}}\left(2\sum_{l=1}^m\left(\sum_{i=1}^n\left\|X_i^l\right\|^2\right)^2 + \sum_{l,r=1}^m\sum_{i,j=1}^n\left\langle X_i^l, X_j^r\right\rangle^2\right)^{1/4}. \tag{9}$$

We also have

$$\mathcal{R}_n^m\left(\mathcal{F}_B \circ \mathcal{A}\right) \leq \frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\left(\frac{1}{m} + \frac{1}{m^2}\sum_{l\neq r}\left\langle C_l, C_r\right\rangle_{HS}^{1/2}\right)^{1/2}. \tag{10}$$

and

$$\mathcal{R}_n^m\left(\mathcal{F}_B \circ \mathcal{A}\right) \leq \frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{\sqrt{n}}\left(\frac{3}{m} + \left\|C\right\|_{HS}^2\right)^{1/4} \tag{11}$$

The proof requires a simple, but technical looking lemma, bounding a sum of inner products involving a common operator T.

Lemma 2. *Let T be a bounded operator on H, $w_1, ..., w_m$ and $v_1, ..., v_m$ vectors in H with $\|v_i\| \leq B$. Then*

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

and

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

Proof. Without loss of generality assume $B = 1$. Using the formulas (5) and (6) we have

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

This proves the first inequality. Also, using Schwartz' inequality in both H and HS

$$\begin{aligned}\sum_{l=1}^{m}\langle Tw_l, v_l\rangle &\leq \left(\sum_{l=1}^{m}\|v_l\|^2\right)^{1/2}\left(\sum_{l=1}^{m}\|Tw_l\|^2\right)^{1/2} \leq \sqrt{m}\left\langle T^2, \sum_{l=1}^{m} Q_{w_l}\right\rangle_{HS}^{1/2} \\ &\leq \sqrt{m}\left\|T^2\right\|_{HS}^{1/2}\left\|\sum_{l=1}^{m} Q_{w_l}\right\|_{HS}^{1/2} = \sqrt{m}\left\|T^2\right\|_2^{1/2}\left(\sum_{l,r}\langle w_l, w_r\rangle^2\right)^{1/4} \quad \square\end{aligned}$$

Proof (of Theorem 5). Define the random variables $w_l = w_l(\sigma) = \sum_{i=1}^{n}\sigma_i^l X_i^l$ depending on the Rademacher variables σ_i^l. Then by Lemma 2 and Jensen's inequality

$$\begin{aligned}\hat{\mathcal{R}}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) &= E_\sigma\left[\sup_{T\in\mathcal{A}}\sup_{v_1,\ldots,v_m,\|v_1\|\leq B}\frac{2}{nm}\sum_{l=1}^{m}\langle Tw_l, v_l\rangle\right] \qquad (12)\\ &\leq \frac{2B}{nm}\|\mathcal{A}\|_{HS}\, E_\sigma\left[\left(\sum_{l,r}|\langle w_l, w_r\rangle|\right)^{1/2}\right] \\ &\leq \frac{2B}{nm}\|\mathcal{A}\|_{HS}\left(\sum_{l,r}E_\sigma\left[|\langle w_l, w_r\rangle|\right]\right)^{1/2}.\end{aligned}$$

Now we have

$$E_\sigma\left[\|w_l\|^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}E_\sigma\left[\sigma_i^l\sigma_j^l\right]\langle x_i^l, x_j^l\rangle = \sum_{i=1}^{n}\left\|x_i^l\right\|^2. \qquad (13)$$

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

$$\begin{aligned}\left(E_\sigma\left[|\langle w_l, w_r\rangle|\right]\right)^2 &\leq E_\sigma\left[\langle w_l, w_r\rangle^2\right] \\ &= \sum_{i=1}^n\sum_{j=1}^n\sum_{i'=1}^n\sum_{j'=1}^n E_\sigma\left[\sigma_i^l\sigma_j^r\sigma_{i'}^l\sigma_{j'}^r\right]\left\langle X_i^l, X_j^r\right\rangle\left\langle X_{i'}^l, X_{j'}^r\right\rangle \\ &= \sum_{i.j=1}^n \left\langle X_i^l, X_j^r\right\rangle^2.\end{aligned} \tag{14}$$

Taking the square-root and inserting it together with (13) in (12) we get (8).

For (9) we first use the second part of Lemma 2 and Jensen's inequality

$$\hat{\mathcal{R}}\left(\mathcal{F}_B \circ \mathcal{A}\right)(\mathbf{x}) \leq \frac{2B}{n\sqrt{m}}\left\|\mathcal{A}^2\right\|_{HS}^{1/2}\left(\sum_{l,r} E_\sigma\left[\langle w_l, w_r\rangle^2\right]\right)^{1/4}. \tag{15}$$

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

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

Inserting this together with (14) in (15) gives (9).

Taking the expectation of (8), using Jensen's inequality, $\left\|X^l\right\| \leq 1$ and independence of X^l and X^r for $l \neq r$, we get

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

which gives (10). In a similar way we obtain from (9)

$$\mathcal{R}_n^m\left(\mathcal{F}_B\circ\mathcal{A}\right)\leq\frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}}\left(2mn^2+mn+\sum_{l,r}^{m}\sum_{\substack{i,j=1\\(l,i)\neq(r,j)}}^{n}\left\langle E\left[Q_{X_i^l}\right],E\left[Q_{X_j^r}\right]\right\rangle^2\right)^{1/4}$$
$$\leq\frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}}\left(3mn^2+m^2n^2\left\|E\left[\frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}Q_{X_i^l}\right]\right\|_{HS}^2\right)^{1/4},$$

which gives (11) □

5 Bounds for multi-task learning

Inserting the bounds of Theorem 5 in Theorem 4 immediately gives

Theorem 6. *Let $\mathcal{A}$ be a be set of bounded, symmetric operators in H and $\epsilon,\delta\in(0,1)$*

(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

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

where A is either

$$A=\frac{2B\left\|\mathcal{A}\right\|_{HS}}{\sqrt{n}}\left(\frac{1}{m}+\frac{1}{m^2}\sum_{l\neq r}\left\langle C^l,C^r\right\rangle_{HS}^{1/2}\right)^{1/2}$$

or

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

(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

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

where the random variable A is either

$$A\left(\mathbf{X}\right)=\frac{2B\left\|\mathcal{A}\right\|_{HS}}{nm}\left(\sum_{l=1}^{m}\sum_{i=1}^{n}\left\|X_i^l\right\|^2+\sum_{l\neq r}\left(\sum_{i,j=1}^{n}\left\langle X_i^l,X_j^r\right\rangle^2\right)^{1/2}\right)^{1/2}$$

or

$$A\left(\mathbf{X}\right)=\frac{2B\left\|\mathcal{A}^2\right\|_{HS}^{1/2}}{n\sqrt{m}}\left(2\sum_{l=1}^{m}\left(\sum_{i=1}^{n}\left\|X_i^l\right\|^2\right)^2+\sum_{l,r=1}^{m}\sum_{i,j=1}^{n}\left\langle X_i^l,X_j^r\right\rangle^2\right)^{1/4}.$$

We can invoke again the model-selection Lemma 15.5 from [1] to stratify over different operator norms. This will give a very similar looking result, which we state in abbreviated fashion.

Theorem 7. *Theorem 6 holds with the following modifications:*

- *The class* $\mathcal{F}_B \circ \mathcal{A}$ *is replaced by all of* $\mathcal{F}_B \circ HS^*$.
- $\|\mathcal{A}\|_{HS}$ *and* $\left\|\mathcal{A}^2\right\|_{HS}$ *are replaced by* $\|T\|_{HS} \vee 1$ *and* $\left\|T^2\right\|_{HS} \vee 1$ *respectively.*
- $(1-\epsilon)$ *and* $(\delta\gamma\epsilon)$ *are replaced by* $(1-\epsilon)^2$ *and* $\left(\delta\gamma\epsilon^2\right)$ *respectively.*

The passage from $\|T\|_{HS}$ to $\|T\|_{HS} \vee 1$ is an artifact introduced by the stratification. We could also require $\|T\|_{HS} \geq 1$. It is easy to derive the results stated in the introduction: Jensen's inequality gives the bounds

$$\frac{1}{m^2}\sum_{l\neq r}\langle C_l, C_r\rangle_{HS}^{1/2} \leq \left(\frac{1}{m^2}\sum_{l,r}\langle C_l, C_r\rangle_{HS}\right)^{1/2} = \|C\|_{HS}$$

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

Using these inequalities and setting $\epsilon = 1/2$ in Theorem 7 we obtain Theorem 1 and Theorem 2.

References

1. M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.
2. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 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. Nello Cristianini and John Shawe-Taylor, Support Vector Machines, *Cambridge University Press*, 2000.
9. 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.
10. Colin McDiarmid, Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195-248. Springer, Berlin, 1998.

11. S.Mika, B.Schölkopf, A.Smola, K.-R.Müller, M.Scholz and G.Rätsch. Kernel PCA and De-noising in Feature Spaces, in *Advances in Neural Information Processing Systems* 11, 1998.
12. J. Shawe-Taylor, N. Christianini, Estimating the moments of a random vector, *Proceedings of GRETSI 2003 Conference*, I: 47–52, 2003.
13. Michael Reed and Barry Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.
14. S.Thrun, Lifelong Learning Algorithms, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998
15. R. K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, to appear in *Journal of Machine Learning Research*.