

# Excess Risk Bounds for Multi-Task Learning

Ambuj Tewari

August 21, 2006

## 1 Introduction

The idea that it should be easier to learn several tasks if they are related in some way is quite intuitive and has been found to work in many practical settings. There has been some interest in obtaining theoretical results to better understand this phenomenon (e.g. [3, 4]). Maurer [4] considers the case when the “relatedness” of the tasks is captured by requiring that all tasks share a common “preprocessor”. Different linear classifiers are learned for the tasks where these classifiers all operate on the “preprocessed” input. Maurer obtains dimension-free and data-dependent bounds in this setting. He bounds the average error over tasks in terms of the margins of the classifiers and a complexity term involving the Hilbert-Schmidt norm of the selected preprocessor and the Frobenius norm of the Gram matrix for all tasks.

We work in the same setting as Maurer’s. However, we introduce a loss function to measure the performance of the selected classifiers. Our aim is to obtain bounds for the difference between the average risk per task of the classifiers learned from the data and the least possible value of the average risk per task.

Suppose we have  $m$  binary classification tasks with a common input space  $\mathcal{X}$  which is a unit ball  $\{x : \|x\| \leq 1\}$  in some Hilbert space  $H$ . Since we deal with binary classification, the output space is  $\mathcal{Y} = \{+1, -1\}$ . Let  $\mathbf{v}$  denote a tuple of classifiers  $(v_1, \dots, v_m)$  with  $v_l \in H$  for all  $l \in \{1, \dots, m\}$ . Let  $\mathcal{A}$  be a set of symmetric Hilbert-Schmidt operators with  $\|T\|_{HS} \leq t$  for all  $T \in \mathcal{A}$ . Denote the input distribution for task  $l$  by  $P^l$  and let

$$P(\mathbf{x}) = P_1(x^1) \otimes \dots \otimes P_l(x^l)$$

Given a multi-classifier  $(\mathbf{v}, T)$  and inputs  $\mathbf{x} = (x^1, \dots, x^m)$ , the  $m$  labels for these inputs are given by the signs in the  $m$ -tuple

$$(\langle Tx^1, v_1 \rangle, \dots, \langle Tx^m, v_m \rangle).$$

We assume that a multi-classifier  $\mathbf{v}, T$  is chosen from a space  $\mathcal{V} \subseteq \{(\mathbf{v}, T) : \max_l \|v_l\| \leq 1, \|T\|_{HS} \in \mathcal{A}\}$  with the following property. Whenever  $(\mathbf{v}, T)$  and  $(\mathbf{u}, S)$  belong to  $\mathcal{V}$  and  $\alpha \in [0, 1]$ , we have some  $\mathbf{v}', T' \in \mathcal{V}$  such that

$$T'v'_i = \alpha T v_i + (1 - \alpha) S u_i \tag{1}$$

for all  $l \in \{1, \dots, m\}$ . In other words, the set

$$\{(Tv_1, \dots, Tv_m) : (\mathbf{v}, T) \in \mathcal{V}\} \subseteq H^m$$

is convex. We further assume that if  $(\mathbf{v}, T) \in \mathcal{V}$  then  $(-\mathbf{v}, T) \in \mathcal{V}$ . We assume that the inputs for the  $m$  tasks are drawn independently. The inputs  $x^l, x_1^l, \dots, x_n^l$  are all drawn i.i.d. from task number  $l$ . The same is true for the labels  $y^l, y_1^l, \dots, y_n^l$ . The data set consists of

$$((x_i^l)_{(i,l)=(1,1)}^{(n,m)}, (y_i^l)_{(i,l)=(1,1)}^{(n,m)}) .$$

Suppose the loss function  $\phi : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$  is convex (as a function of a single variable) for each  $y \in \mathcal{Y}$ . We make the following two assumptions about the loss function.

**Assumption 1** For all  $y \in \mathcal{Y}$ ,  $x, x' \in \mathbb{R}$ , we have

$$\frac{\phi(x', y) + \phi(x, y)}{2} - \phi\left(\frac{x' + x}{2}, y\right) \geq C|x' - x|^2 \quad (2)$$

for some constant  $C > 0$ .

**Assumption 2** For all  $y \in \mathcal{Y}$ ,  $x, x' \in \mathbb{R}$ , we have

$$|\phi(x', y) - \phi(x, y)| \leq L|x' - x|$$

for some constant  $L > 0$ .

Let  $\mathbf{v} \circ T$  denote the function

$$(x^1, \dots, x^m) \mapsto \frac{1}{m} \sum_{l=1}^m \langle Tx^l, v_l \rangle .$$

Define a pseudometric  $\mathcal{D}$  on  $\mathcal{V}$  by

$$\mathcal{D}(\mathbf{v}, T; \mathbf{u}, S) = \frac{1}{m} \sum_{l=1}^m \mathbb{E}(\langle Tx^l, v_l \rangle - \langle Sx^l, u_l \rangle)^2 .$$

Define the average per-task risk of  $\mathbf{v}, T$

$$R_\phi(\mathbf{v}, T) = \frac{1}{m} \sum_{l=1}^m \mathbb{E} \phi(\langle Tx^l, v_l \rangle, y^l) ,$$

and its empirical version,

$$\hat{R}_\phi(\mathbf{v}, T) = \frac{1}{mn} \sum_{i=1}^n \sum_{l=1}^m \phi(\langle Tx_i^l, v_l \rangle, y_i^l) .$$

We assume that there exist  $(\mathbf{v}^*, T^*)$  and  $(\hat{\mathbf{v}}, \hat{T})$  belonging to  $\mathcal{V}$  such that

$$R_\phi(\mathbf{v}^*, T^*) = \inf_{(\mathbf{v}, T) \in \mathcal{V}} R_\phi(\mathbf{v}, T) ,$$

$$\hat{R}_\phi(\hat{\mathbf{v}}, \hat{T}) = \inf_{(\mathbf{v}, T) \in \mathcal{V}} \hat{R}_\phi(\mathbf{v}, T) .$$

For a subset  $\mathcal{U}$  of  $\mathcal{V}$ , we define

$$R_n\{\mathbf{v} \circ T : (\mathbf{v}, T) \in \mathcal{U}\} = \sup_{(\mathbf{v}, T) \in \mathcal{U}} \frac{1}{mn} \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle T x_i^l, v_l \rangle ,$$

where  $\sigma_1, \dots, \sigma_n$  are independent Rademacher random variables, that is, each  $\sigma_i$  takes value  $+1$  or  $-1$  with probability  $1/2$ .

## 2 Results for Excess Risk

**Lemma 1.** *If  $(\mathbf{v}, T) \in \mathcal{V}$ ,  $\|x\| \leq 1$  then, for all  $l$ ,  $\langle T x^l, v_l \rangle \leq t$ .*

*Proof.* Follows from Lemma 2 in [4]. □

**Lemma 2.**

$$\frac{R_\phi(\mathbf{v}, T) - R_\phi(\mathbf{v}^*, T^*)}{2} \geq CD(\mathbf{v}, T; \mathbf{v}^*, T^*) .$$

*Proof.* For  $\alpha = 1/2$ ,  $\mathbf{u} = \mathbf{v}^*$ ,  $S = T^*$ , get  $\mathbf{v}', T'$  such that (1) holds. Then we have, for any  $x_l$  and for all  $l \in \{1, \dots, m\}$ ,

$$\begin{aligned} \langle T' x^l, v'_l \rangle &= \langle x^l, T' v'_l \rangle \\ &= \frac{1}{2} \langle x^l, T v_l \rangle + \frac{1}{2} \langle x^l, T^* v_l^* \rangle \\ &= \frac{1}{2} \langle T x^l, v_l \rangle + \frac{1}{2} \langle T^* x^l, v_l^* \rangle , \end{aligned} \tag{3}$$

and so we have

$$\begin{aligned} & \frac{R_\phi(\mathbf{v}, T) - R_\phi(\mathbf{v}^*, T^*)}{2} \\ &= \frac{R_\phi(\mathbf{v}, T) + R_\phi(\mathbf{v}^*, T^*)}{2} - R_\phi(\mathbf{v}^*, T^*) \\ &\geq \frac{R_\phi(\mathbf{v}, T) + R_\phi(\mathbf{v}^*, T^*)}{2} - R_\phi(\mathbf{v}', T') \\ &= \frac{1}{m} \sum_{l=1}^m \mathbb{E} \left[ \frac{\phi(\langle T x^l, v_l \rangle, y^l) + \phi(\langle T^* x^l, v_l^* \rangle, y^l)}{2} - \phi(\langle T' x^l, v'_l \rangle, y^l) \right] \\ &\geq \frac{C}{m} \sum_{l=1}^m \mathbb{E}(\langle T x^l, v^l \rangle - \langle T^* x^l, v_l^* \rangle)^2 \\ &= CD(\mathbf{v}, T; \mathbf{v}^*, T^*) , \end{aligned}$$

where the first inequality follows from the definition of  $(\mathbf{v}^*, T^*)$  and the second one follows from (2) and (3). □

By a (non-trivial) *sub-root* function we mean a function  $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$  which is non-negative, non-decreasing, not identically zero and is such that  $r \mapsto \psi(r)/r$  is non-increasing. Sub-root functions have unique fixed points (see, for example, [2]) on  $\mathbb{R}_+$ . If  $r^*$  is the fixed point of  $\psi$  then  $r \geq \psi(r)$  iff  $r \geq r^*$ .

**Theorem 3.** *If  $\psi$  is a sub-root function satisfying*

$$\psi(r) \geq \frac{L^3}{2Cm} \mathbb{E}R_n\{\mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r\} \quad (4)$$

then for any  $x > 0$  and any  $r \geq \psi(r)$ , with probability  $\geq 1 - e^{-x}$ ,

$$R_\phi(\hat{\mathbf{v}}, \hat{T}) - R_\phi(\mathbf{v}^*, T^*) \leq \frac{c_1 C}{L^2} mr + (c_2 Lt + c_3 L^2/Cm) \frac{x}{n}$$

for some universal constants  $c_1, c_2, c_3$ .

*Proof.* Let  $\phi_{\mathbf{v}, T}$  denote the function

$$(x^1, \dots, x^m) \mapsto \frac{1}{m} \sum_{l=1}^m \phi(\langle Tx^l, v_l \rangle, y^l) .$$

Note that  $\mathbb{E}(\phi_{\mathbf{v}, T}) = R_\phi(\mathbf{v}, T)$ . Let  $\mathcal{F}$  be the class

$$\{\phi_{\mathbf{v}, T} - \phi_{\mathbf{v}^*, T^*} : (\mathbf{v}, T) \in \mathcal{V}\} .$$

Let us bound the variance of the functions in the class in terms of their expectations. Since the  $m$  tasks are independent, we have

$$\begin{aligned} \text{var}(\phi_{\mathbf{v}, T} - \phi_{\mathbf{v}^*, T^*}) &= \frac{1}{m^2} \sum_{l=1}^m \text{var}(\phi(\langle Tx^l, v_l \rangle, y^l) - \phi(\langle T^* x^l, v_l^* \rangle, y^l)) \\ &\leq \frac{1}{m^2} \sum_{l=1}^m \mathbb{E}(\phi(\langle Tx^l, v_l \rangle, y^l) - \phi(\langle T^* x^l, v_l^* \rangle, y^l))^2 \\ &\leq \frac{L^2}{m^2} \sum_{l=1}^m \mathbb{E}(\langle Tx^l, v_l \rangle - \langle T^* x^l, v_l^* \rangle)^2 \\ &= L^2 \mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \\ &\leq L^2 (R_\phi(\mathbf{v}, T) - R_\phi(\mathbf{v}^*, T^*)) / 2Cm . \end{aligned}$$

The first inequality above is straightforward. The second one is due to our assumption about Lipschitz continuity of  $\phi(\cdot, y)$ . Lemma 2 gives the third inequality. The range of functions in  $\mathcal{F}$  is  $[-2Lt, 2Lt]$  due to Lemma 1 and Lipschitz continuity of  $\phi(\cdot, y)$ . We now use Theorem 3.3, part 1 from [2] with  $b - a \leftarrow 4Lt$ ,  $T(f) \leftarrow L^2 \mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m$  (for  $f = \phi_{\mathbf{v}, T} - \phi_{\mathbf{v}^*, T^*}$ ),  $B \leftarrow L^2/2Cm$ ,  $K \leftarrow 2$ . Note that (4) implies

$$\begin{aligned} \psi(r) &\geq \frac{L^2}{2Cm} \mathbb{E}R_n\{\phi_{\mathbf{v}, T} : L^2 \mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r\} \\ &= \frac{L^2}{2Cm} \mathbb{E}R_n\{\phi_{\mathbf{v}, T} - \phi_{\mathbf{v}^*, T^*} : L^2 \mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r\} \end{aligned}$$

by the contraction property of Rademacher averages [2, Thm A.6]. So for any  $r$  with  $r \geq \psi(r)$ , we have, with probability  $\geq 1 - e^{-x}$ ,

$$R_\phi(\hat{\mathbf{v}}, \hat{T}) - R_\phi(\mathbf{v}^*, T^*) \leq 2(\hat{R}_\phi(\hat{\mathbf{v}}, \hat{T}) - \hat{R}_\phi(\mathbf{v}^*, T^*)) \\ + \frac{2816Cm}{L^2}r + \frac{x(44Lt + 26L^2/2Cm)}{n}.$$

Observing that the first term on the right hand side above is non-positive completes the proof.  $\square$

**Lemma 4.** *The function  $\psi$  given by*

$$r \mapsto \mathbb{E}R_n\{\mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r\}$$

*is sub-root.*

*Proof.* Positivity follows from Jensen's inequality in the form

$$\mathbb{E} \sup(\cdot) \geq \sup \mathbb{E}(\cdot),$$

and the non-decreasing property is obvious. So, we only need to ensure that if  $0 < r_1 < r_2$  then  $\psi(r_1) \geq \sqrt{r_1/r_2}\psi(r_2)$ . Fix the data set and a realization of the Rademacher averages. Let  $(\mathbf{u}, S)$  achieve the supremum in

$$\sup_{L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r_2} \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle T x_i^l, v_l \rangle.$$

For  $\alpha = 1 - \sqrt{r_1/r_2}$ ,  $\mathbf{v} = \mathbf{v}^*$ ,  $T = T^*$ , get  $\mathbf{v}', T'$  such that (1) holds. Thus we have, for all  $l \in \{1, \dots, m\}$ ,

$$T'v'_l = T^*v_l^* + \sqrt{\frac{r_1}{r_2}}(Su_l - T^*v_l^*).$$

This gives

$$\begin{aligned} \mathcal{D}(\mathbf{v}', T'; \mathbf{v}^*, T^*) &= \frac{1}{m} \sum_{l=1}^m \mathbb{E} (\langle T' x^l, v'_l \rangle - \langle T^* x^l, v_l^* \rangle)^2 \\ &= \frac{1}{m} \sum_{l=1}^m \mathbb{E} (\langle x^l, T'v'_l - T^*v_l^* \rangle)^2 \\ &= \frac{1}{m} \sum_{l=1}^m \mathbb{E} \left( \sqrt{\frac{r_1}{r_2}} \langle x^l, Su_l - T^*v_l^* \rangle \right)^2 \\ &= \frac{r_1}{r_2} \frac{1}{m} \sum_{l=1}^m \mathbb{E} (\langle Sx^l, u_l \rangle - \langle T^* x^l, v_l^* \rangle)^2 \\ &= \frac{r_1}{r_2} \mathcal{D}(\mathbf{u}, S; \mathbf{v}^*, T^*) \leq r_1 m / L^2. \end{aligned}$$

Thus we have

$$\begin{aligned}
& \sup_{L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r_1} \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle T x_i^l, v_l \rangle \\
& \geq \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle T' x_i^l, v_l' \rangle \\
& = \sqrt{\frac{r_1}{r_2}} \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle S x_i^l, u_l \rangle + \left(1 - \sqrt{\frac{r_1}{r_2}}\right) \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle T^* x_i^l, v_l^* \rangle \\
& = \sqrt{\frac{r_1}{r_2}} \sup_{L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r_2} \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle T x_i^l, v_l \rangle \\
& \quad + \left(1 - \sqrt{\frac{r_1}{r_2}}\right) \sum_{i=1}^n \sum_{l=1}^m \sigma_i \langle T^* x_i^l, v_l^* \rangle .
\end{aligned}$$

Dividing by  $n$  and taking expectations (w.r.t. the data set and the Rademacher variables) makes the last term vanish and we get

$$\psi(r_1) \geq \sqrt{\frac{r_1}{r_2}} \psi(r_2) .$$

□

The following corollary is an almost immediate consequence of Theorem 3 and Lemma 4.

**Corollary 5.** *Define*

$$\psi(r) = \frac{L^3}{2C} \mathbb{E} R_n \{ \mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*) \leq r \} \quad (5)$$

Then, with probability  $\geq 1 - e^{-x}$ ,

$$R_\phi(\hat{\mathbf{v}}, \hat{T}) - R_\phi(\mathbf{v}^*, T^*) \leq \frac{c_1 C}{L^2} r^* + (c_2 L t + c_3 L^2 / C m) \frac{x}{n}$$

where  $r^*$  is the fixed point of  $\psi$ .

*Proof.* Combine Theorem 3 and Lemma 4. Note that  $r^*$  is the fixed point of

$$r \mapsto \frac{L^3}{2Cm} \mathbb{E} R_n \{ \mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*)/m \leq r \}$$

iff  $mr^*$  is the fixed point of

$$r \mapsto \frac{L^3}{2C} \mathbb{E} R_n \{ \mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*) \leq r \} .$$

□

Define the following inner product on  $H^m$ ,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{m} \sum_{l=1}^m \langle x^l, y^l \rangle$$

where  $\mathbf{x} = (x^1, \dots, x^m)$ ,  $\mathbf{v} = (y^1, \dots, y^m)$ .

**Lemma 6.** *Let  $(\lambda_j)$  be the eigenvalues of the operator  $\mathbb{T}$  defined by*

$$(\mathbb{T}f)(\mathbf{x}) = \int_{H^m} \langle \mathbf{x}, \mathbf{y} \rangle f(\mathbf{y}) dP(\mathbf{y})$$

*arranged in decreasing order. Then we have*

$$\mathbb{E}R_n\{\mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*) \leq r\} \leq 2 \sqrt{\frac{c}{n} \sum_{j=1}^{\infty} \min(\lambda_j, r)}$$

where  $c = \mathcal{A}^2 + 1/2L^2$ .

*Proof.* We have

$$\begin{aligned} & \mathbb{E}R_n\{\mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*) \leq r\} \\ &= \mathbb{E}R_n\{\mathbf{v} \circ T - \mathbf{v}^* \circ T^* : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{v}^*, T^*) \leq r\} \\ &\leq \mathbb{E}R_n\{\mathbf{v} \circ T - \mathbf{u} \circ S : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{u}, S) \leq r\} \\ &= 2 \mathbb{E}R_n\{\mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{0}, 0) \leq r/4\} . \end{aligned}$$

The last equality follows because if  $(\mathbf{v}, T), (\mathbf{u}, S) \in \mathcal{V}$  then  $(\mathbf{v}, T), (-\mathbf{u}, S) \in \mathcal{V}$  and so, by (1), we get  $(\mathbf{v}', T') \in \mathcal{V}$  such that, for all  $l \in \{1, \dots, m\}$ ,

$$T'v'_l = \frac{1}{2}(Tv_l - Su_l) .$$

For such a  $(\mathbf{v}', T')$ , we have

$$2\mathbf{v}' \circ T' = \mathbf{v} \circ T - \mathbf{u} \circ S ,$$

$$\mathcal{D}(\mathbf{v}', T'; \mathbf{0}, 0) = \mathcal{D}(\mathbf{v}, T; \mathbf{u}, S)/4 .$$

Define  $\mathcal{V}_r = \{(\mathbf{v}, T) \in \mathcal{V} : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{0}, 0) \leq r/4\}$ . Then we have

$$\begin{aligned} & \mathbb{E}R_n\{\mathbf{v} \circ T : L^2\mathcal{D}(\mathbf{v}, T; \mathbf{0}, 0) \leq r/4\} \\ &= \mathbb{E} \sup_{(\mathbf{v}, T) \in \mathcal{V}_r} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \frac{1}{m} \sum_{l=1}^m \langle Tx_i^l, v_l \rangle \right) \quad (6) \end{aligned}$$

If  $(\mathbf{v}, T) \in \mathcal{V}_r$  then

$$\begin{aligned} & \mathbb{E} \left( \frac{1}{m} \sum_{l=1}^m \langle x^l, T v_l \rangle^2 \right) \leq r/4L^2 \\ \Rightarrow & \mathbb{E} \left( \frac{1}{m} \sum_{l=1}^m \langle x^l, T v_l \rangle \right)^2 \leq r/4L^2 \\ \Rightarrow & \mathbb{E} \langle \mathbf{x}, \mathbf{u} \rangle^2 \leq r/4L^2 \end{aligned}$$

where  $\mathbf{u} = (T v_1, \dots, T v_m)$ . The first implication holds because for any real numbers  $r_1, \dots, r_m$ ,  $(\sum_l r_l/m)^2 \leq (\sum_l r_l^2)/m$ .  $(\mathbf{v}, T) \in \mathcal{V}_r$  also implies that

$$\langle \mathbf{u}, \mathbf{u} \rangle = \frac{1}{m} \sum_{l=1}^m \|T v_l\|^2 \leq \mathcal{A},$$

because  $\|T v_l\| \leq \mathcal{A} \|v_l\| \leq \mathcal{A}$ . If we define the set

$$\mathcal{V}'_r = \{ \mathbf{u} \in H^m : \|\mathbf{u}\| \leq \mathcal{A}, \mathbb{E} \langle \mathbf{x}, \mathbf{u} \rangle^2 \leq r/4L^2 \},$$

we thus see that (6) becomes

$$\mathbb{E} \sup_{\mathbf{u} \in \mathcal{V}'_r} \frac{1}{n} \sum_{l=1}^m \sigma_l \langle \mathbf{x}_i, \mathbf{u} \rangle \quad (7)$$

where  $\mathbf{x}_i = (x_i^1, \dots, x_i^m)$ . Appendix A shows that (7) is no more than

$$\sqrt{\frac{c}{n} \sum_{j=1}^{\infty} \min(\lambda_j, r)},$$

where  $c = \mathcal{A}^2 + 1/4L^2$  and  $\lambda_j$ 's are the eigenvalues of the operator defined by

$$(\mathbb{T}f)(\mathbf{x}) = \int_{H^m} \langle \mathbf{x}, \mathbf{y} \rangle f(\mathbf{y}) dP(\mathbf{y}) \quad (8)$$

arranged in decreasing order.

Suppose  $\lambda$  is a eigenvalue associated with an eigenvector  $g$  of the operator

$$(\mathbb{T}^l g)(x^l) \mapsto \int_H \langle x^l, y^l \rangle g(y^l) dP_l(y^l).$$

Then we have

$$\lambda g(x^l) = \int_H \langle x^l, y^l \rangle g(y^l) dP_l(x^l)$$

Let us see if

$$\tilde{g}(\mathbf{y}) = (0, \dots, g(y^l) + \alpha, \dots, 0)$$

can work as an eigenvector of (8) for some choice of the constant  $c$ .

$$\begin{aligned}
(\mathbb{T}\tilde{g})(\mathbf{x}) &= \int_{H^m} \langle \mathbf{x}, \mathbf{y} \rangle \tilde{g}(\mathbf{y}) dP(\mathbf{y}) \\
&= \int_{H^m} \frac{1}{m} \left( \sum_{k=1}^m \langle x^k, y^k \rangle \right) (g(y^l) + \alpha) dP(\mathbf{y}) \\
&= \frac{1}{m} \sum_{k \neq l} \langle x^k, \mathbb{E}y^k \rangle \mathbb{E}(g(y^l)) + \frac{1}{m} \left( \int_H \langle x^l, y^l \rangle g(y^l) dP_l(y^l) \right) + \frac{\alpha}{m} \sum_{k=1}^m \langle x^k, \mathbb{E}y^k \rangle \\
&= \frac{1}{m} \sum_{k \neq l} \langle x^k, \mathbb{E}y^k \rangle \mathbb{E}(g(y^l)) + \frac{1}{m} \lambda g(x^l) + \frac{\alpha}{m} \sum_{k=1}^m \langle x^k, \mathbb{E}y^k \rangle
\end{aligned}$$

If we want the above to be equal to  $\lambda/m(g(x^l) + \alpha)$  then  $\alpha$  should satisfy

$$\sum_{k \neq l} \langle x^k, \mathbb{E}y^k \rangle \mathbb{E}(g(y^l)) + \alpha \sum_{k=1}^m \langle x^k, \mathbb{E}y^k \rangle = \lambda \alpha .$$

But we have a problem here as the “constant”  $\alpha$  depends on  $\mathbf{x}$ . Note, however, that there are two special cases which make  $\alpha = 0$  work:

1.  $\forall l, \mathbb{E}y^l = 0$  (the input vectors for the various tasks are centered to have mean zero).
2.  $\mathbb{E}(g(y^l)) = 0$  (This might be the case if the constant function is eigenvector for task  $l$ . Then the rest of the eigenvectors, being orthogonal to it, would satisfy this.)

□

## Appendix A

We essentially follow the argument presented in [5]. Let

$$R = \mathbb{E} \sup_{\mathbf{u} \in \mathcal{V}'_r} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle \mathbf{x}_i, \mathbf{u} \rangle .$$

Embed  $\mathbf{u}$  in  $\ell_2$  via  $\Phi$ . Let  $(e_i)$  be the standard basis of  $\ell_2$ . Then we have

$$\Phi(\mathbf{u}) = \sum_{j=1}^{\infty} \nu_j e_j$$

For  $\mathbf{u} \in \mathcal{V}'_r$ ,  $\|\mathbf{u}\| \leq \mathcal{A}$  and so we have  $\sum_j \nu_j^2 \leq \mathcal{A}^2$ . We also have  $\mathbb{E} \langle \mathbf{x}, \mathbf{u} \rangle^2 \leq r/4L^2$ . Since  $\mathbf{x}$  has representation

$$\Phi(\mathbf{x}) = \sum_{j=1}^{\infty} \phi_j(\mathbf{x}) \sqrt{\lambda_j} e_j$$

where  $\phi_j$ 's are the orthogonal eigenvectors of the integral operator (8), we have

$$r/4L^2 \geq \mathbb{E}\langle \mathbf{x}, \mathbf{u} \rangle^2 = \mathbb{E}\langle \Phi(\mathbf{x}), \Phi(\mathbf{u}) \rangle_{\ell_2} = \sum_{j=1}^{\infty} \lambda_j \nu_j^2 \mathbb{E}(\phi_j(\mathbf{x}))^2 = \sum_{j=1}^{\infty} \lambda_j \nu_j^2 .$$

This, along with  $\sum_j \nu_j^2 \leq \mathcal{A}^2$ , gives

$$\sum_{j=1}^{\infty} \nu_j^2 (1 + \lambda_j/r) \leq \mathcal{A}^2 + 1/4L^2 \Rightarrow \sum_{j=1}^{\infty} \nu_j^2 \max(1, \lambda_j/r) \leq c ,$$

where  $c = \mathcal{A}^2 + 1/4L^2$ . Set  $\mu_j = \max(1, \lambda_j/r)$  to get

$$\begin{aligned} R^2 &= \mathbb{E} \sup \left\{ \left\langle \sum_{i=1}^n \sigma_i \mathbf{x}_i, \mathbf{u} \right\rangle^2 : \mathbf{u} \in \mathcal{V}'_r \right\} \\ &\leq \frac{1}{n^2} \mathbb{E} \sup \left\{ \left\langle \sum_{j=1}^{\infty} \sqrt{\lambda_j} \left( \sum_{i=1}^n \sigma_i \phi_j(\mathbf{x}_i) \right) e_j, \sum_{j=1}^{\infty} \nu_j e_j \right\rangle_{\ell_2}^2 : \sum_{j=1}^{\infty} \nu_j^2 \mu_j \leq c \right\} \\ &\leq \frac{c}{n^2} \mathbb{E} \sup \left\{ \left\langle \sum_{j=1}^{\infty} \sqrt{\frac{\lambda_j}{\mu_j}} \left( \sum_{i=1}^n \sigma_i \phi_j(\mathbf{x}_i) \right) e_j, \sum_{j=1}^{\infty} \nu_j \sqrt{\frac{\mu_j}{c}} e_j \right\rangle_{\ell_2}^2 : \sum_{j=1}^{\infty} \nu_j^2 \mu_j \leq c \right\} \end{aligned}$$

Now  $\sum_{j=1}^{\infty} \nu_j^2 \mu_j \leq c$  implies that  $\| \sum_{j=1}^{\infty} \nu_j \sqrt{\mu_j/c} e_j \| \leq 1$ . So Cauchy-Schwarz inequality for  $\ell_2$  gives us

$$\begin{aligned} R^2 &\leq \frac{c}{n^2} \mathbb{E} \left\| \sum_{j=1}^{\infty} \sqrt{\frac{\lambda_j}{\mu_j}} \left( \sum_{i=1}^n \sigma_i \phi_j(\mathbf{x}_i) \right) e_j \right\|_{\ell_2}^2 \\ &= \frac{c}{n^2} \sum_{j=1}^{\infty} \frac{\lambda_j}{\mu_j} \mathbb{E} \left( \sum_{i=1}^n \sigma_i \phi_j(\mathbf{x}_i) \right)^2 \\ &= \frac{c}{n^2} \sum_{j=1}^{\infty} \frac{\lambda_j}{\mu_j} \left( \sum_{i,k} \mathbb{E}(\sigma_i \sigma_k \phi_j(\mathbf{x}_i) \phi_j(\mathbf{x}_k)) \right) \\ &= \frac{c}{n^2} \sum_{j=1}^{\infty} \frac{\lambda_j}{\mu_j} \left( \sum_{i,k} \mathbb{E}(\sigma_i \sigma_k) \mathbb{E}(\phi_j(\mathbf{x}_i) \phi_j(\mathbf{x}_k)) \right) \\ &= \frac{c}{n^2} \sum_{j=1}^{\infty} \frac{\lambda_j}{\mu_j} n \end{aligned}$$

where we used the fact that  $\mathbb{E}(\sigma_i \sigma_k) = 1$  iff  $i = k$  and that  $\mathbb{E}\phi_j^2(\mathbf{x}_i) = 1$ . Observing that  $\lambda_j/\mu_j = \min(\lambda_j, r)$  finally yields

$$R^2 \leq \frac{c}{n} \sum_{j=1}^{\infty} \min(\lambda_j, r)$$

and hence

$$\mathbb{E} \sup_{\mathbf{u} \in \mathcal{V}'_r} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle \mathbf{x}_i, \mathbf{u} \rangle \leq \sqrt{\frac{c}{n} \sum_{j=1}^{\infty} \min(\lambda_j, r)}$$

## References

- [1] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results, *Journal of Machine Learning Research* 3, pp. 463–482, 2002.
- [2] Peter L. Bartlett, Olivier Bousquet and Shahar Mendelson. Local Rademacher Complexities, *Annals of Statistics* 33:4, pp. 1497–1537, 2005.
- [3] Jonathan Baxter. A Model of Inductive Bias Learning, *Journal of Artificial Intelligence Research* 12, pp. 149–198, 2000.
- [4] Andreas Maurer. Bounds for Linear Multi-Task Learning, *Journal of Machine Learning Research* 7, pp. 117–139, 2006.
- [5] Shahar Mendelson. Geometric Parameters of Kernel Machines, *Proceedings of the 15th Annual Conference on Computational Learning Theory*, pp. 29–43, 2002.