JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–22

Excess risk bounds for multitask learning with trace norm regularization

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

Massimiliano Pontil M.PONTIL@CS.UCL.AC.UK
Department of Computer Science and Centre for Computational Statistics and Machine Learning University College London, Malet Place London WC1E, UK

Abstract

Trace norm regularization is a popular method of multitask learning. We give excess risk bounds with explicit dependence on the number of tasks, the number of examples per task and properties of the data distribution. The bounds are independent of the dimension of the input space, which may be infinite as in the case of reproducing kernel Hilbert spaces. A byproduct of the proof are bounds on the expected norm of sums of random positive semidefinite matrices with subexponential moments.

Keywords: Multitask learning, random matrices, risk bounds, trace norm regularization.

1. Introduction

A fundamental limitation of supervised learning is the cost incurred by the preparation of the large training samples required for good generalization. A potential remedy is offered by multi-task learning: in many cases, while individual sample sizes are rather small, there are samples to represent a large number of learning tasks, which share some constraining or generative property. This common property can be estimated using the entire collection of training samples, and if this property is sufficiently simple it should allow better estimation of the individual tasks despite their small individual sample sizes.

The machine learning community has tried multi-task learning for many years (see Ando and Zhang, 2005; Argyriou et al., 2008; Baxter, 2000; Caruana, 1998; Cavallanti et al., 2010; Evgeniou et al., 2005; Thrun and Pratt, 1998, contributions and references therein), but there are few theoretical investigations which clearly expose the conditions under which multi-task learning is preferable to independent learning. Following the seminal work of Baxter (2000) several authors have given performance bounds under different assumptions of task-relatedness. In this paper we consider multi-task learning with trace-norm regularization (TNML), a technique for which efficient algorithms exist and which has been successfully applied many times (see e.g. Amit et al., 2007; Argyriou et al., 2008; Harchaoui et al., 2012).

In the learning framework considered here the inputs lie in a separable Hilbert space $\mathbb{H}$, which may be finite or infinite dimensional, and the outputs are real numbers. For each of T tasks an unknown input-output relationship is modeled by a distribution μ_t on $\mathbb{H} \times \mathbb{R}$, with

$\mu_t(X,Y)$ being interpreted as the probability of observing the input-output pair (X,Y). We assume bounded inputs, for simplicity $\|X\| \leq 1$, where we use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote euclidean norm and inner product in $\mathbb{H}$, respectively.

A predictor is specified by a weight vector $w \in \mathbb{H}$ which predicts the output $\langle w, x\rangle$ for an observed input $x \in \mathbb{H}$. If the observed output is y, a loss $\ell(\langle w, x\rangle, y)$ is incurred, where ℓ is a fixed loss function on $\mathbb{R}^2$, assumed to have values in $[0,1]$, with $\ell(\cdot, y)$ being Lipschitz with constant L for each $y \in \mathbb{R}$. The expected loss, or risk, of weight vector w in the context of task t is thus $\mathbb{E}_{(X,Y)\sim\mu_t}\ell(\langle w, X\rangle, Y)$. The choice of a weight vector w_t for each task t is equivalent to the choice of a linear map $W : \mathbb{H} \to \mathbb{R}^T$, with $(Wx)_t = \langle x, w_t\rangle$. We seek to choose W so as to (nearly) minimize the total average risk $R(W)$ defined by

$$R(W) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(X,Y)\sim\mu_t}\left[\ell(\langle w_t, X\rangle, Y)\right]. \tag{1}$$

Since the distributions μ_t are unknown, the minimization is based on a finite sample of observations, which for each task t is modelled by a vector $\mathbf{Z}^t$ of n independent random variables $\mathbf{Z}^t = (Z_1^t, \ldots, Z_n^t)$, where each $Z_i^t = (X_i^t, Y_i^t)$ is distributed according to μ_t. The entire multi-sample $(\mathbf{Z}^1, \ldots, \mathbf{Z}^T)$ is denoted by $\bar{\mathbf{Z}}$.

A classical and intuitive learning strategy is empirical risk minimization. One decides on a hypothesis space $\mathcal{W}$ of candidate maps and solves the problem

$$\hat{W}(\bar{\mathbf{Z}}) = \arg\min_{W\in\mathcal{W}} \hat{R}(W, \bar{\mathbf{Z}}),$$

where the average empirical risk $\hat{R}(W, \bar{\mathbf{Z}})$ is defined as

$$\hat{R}(W, \bar{\mathbf{Z}}) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n} \ell(\langle w_t, X_i^t\rangle, Y_i^t).$$

If $\mathcal{W}$ has the form $\mathcal{W} = \{x \mapsto Wx : (Wx)_t = \langle x, w_t\rangle, w_t \in \mathcal{B}\}$ where $\mathcal{B} \subseteq \mathbb{H}$ is some hypothesis space of vectors, then this is equivalent to single task learning, solving for each task independently the problem

$$w_t(\mathbf{Z}_t) = \arg\min_{w\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^{n} \ell(\langle w, X_i^t\rangle, Y_i^t).$$

For proper multi-task learning, however, membership in $\mathcal{W}$ should imply some mutual dependence between the vectors w_t. Here we would like to require the w_t to lie near a common subspace, unknown but assumed to be of low dimension, corresponding to an approximate rank constraint on W. The convex envelope of the rank within the spectral unit ball is given by the trace-norm (see e.g. Fazel et al., 2001) $\|W\|_1 := \mathrm{tr}\big((W^*W)^{1/2}\big)$, and to obtain a tractable optimization problem we define the hypothesis space $\mathcal{W}$ of TNML as

$$\mathcal{W} = \left\{W \in \mathcal{L}(\mathbb{H}, \mathbb{R}^T) : \|W\|_1 \leq B\sqrt{T}\right\}, \tag{2}$$

where $B > 0$ is a regularization constant. The factor $\sqrt{T}$ is an important normalization which we explain below.

A good hypothesis space $\mathcal{W}$ must fulfill two requirements: for one the risk of the best map W^0 in the set,

$$W^0 = \arg\min_{W \in \mathcal{W}} R(W),$$

should be small. This depends on the set of tasks at hand and is largely a matter of domain knowledge. The second requirement is that the risk of the map returned by empirical risk minimization, $\hat{W}(\bar{\mathbf{Z}})$, is not too different from the risk of W^0, so that the excess risk, $R(\hat{W}(\bar{\mathbf{Z}})) - R(W^0)$, is small with high probability in the sample $\bar{\mathbf{Z}}$. The following result, which is the principal contribution of the paper, gives both a distribution dependent and a data dependent bound on the excess risk.

Theorem 1 *(i) For $\delta > 0$ with probability at least $1 - \delta$ in $\bar{\mathbf{Z}}$*

$$R(\hat{W}) - R(W^0) \leq 2LB\left(\sqrt{\frac{\|C\|_\infty}{n}} + 5\sqrt{\frac{\ln(nT)+1}{nT}}\right) + \sqrt{\frac{2\ln(2/\delta)}{nT}},$$

where $\|\cdot\|_\infty$ is the operator norm, and C is the task averaged, uncentered input covariance operator

$$\langle Cv, w\rangle = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(X,Y)\sim\mu_t} \langle v, X\rangle \langle X, w\rangle, \text{ for } w, v \in \mathbb{H}.$$

(ii) With probability $1 - \delta$ in $\bar{\mathbf{Z}}$

$$R(\hat{W}) - R(W^0) \leq 2LB\left(\sqrt{\frac{\left\|\hat{C}\right\|_\infty}{n}} + \sqrt{\frac{2(\ln(nT)+1)}{nT}}\right) + \sqrt{\frac{8\ln(3/\delta)}{nT}},$$

where $\hat{C}$ is the task averaged, uncentered empirical input covariance operator

$$\left\langle \hat{C}v, w\right\rangle = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \langle v, X_i^t\rangle \langle X_i^t, w\rangle, \text{ for } w, v \in \mathbb{H}.$$

Remarks:

1. Suppose that for an operator W all T column vectors w_t are equal to a common vector w, as might be the case if all the tasks T are equivalent. In this case increasing the number of tasks should not increase the regularizer. Since then $\|W\|_1 = \sqrt{T}\|w\|$ we have chosen the factor $\sqrt{T}$ in (2). It allows us to consider the limit $T \to \infty$ for a fixed value of B.
2. A simple computation shows $\|C\|_1 = \mathbb{E}\|X\|^2 \leq 1$. If there are M nonzero eigenvectors of C and the corresponding eigenvalues all happen to be equal because of some symmetry, then $\|C\|_\infty \leq 1/M$. This simple estimate will be used frequently in the interpretation of our results.
3. If the mixture of data distributions is supported on a one dimensional subspace then $\|C\|_\infty = \mathbb{E}\|X\|^2$, and the bound is always worse than standard bounds for single task learning as in (Bartlett and Mendelson, 2002). The situation is similar if the distribution is supported on a very low dimensional subspace. Thus, if learning is already easy, our bounds for TNML show no benefit.

4. If the mixture of data distributions is uniform on an M-dimensional unit sphere in $\mathbb{H}$ then $\|C\|_\infty = 1/M$ and the corresponding term in the bound becomes small. Suppose now that for $W = [w_1, \ldots, w_T]^*$ the w_t all are constrained to be unit vectors lying in a common K-dimensional subspace of $\mathbb{H}$, as might be the solution returned by a method of subspace learning (Ando and Zhang, 2005). If we choose $B = K^{1/2}$ then $W \in \mathcal{W}$, and our bound also applies. The subspace then corresponds to the property shared among the tasks. The cost of its estimation vanishes in the limit $T \to \infty$ as the bound approaches the limiting value

$$2L\sqrt{\frac{K}{nM}},$$

at a rate of $\sqrt{\ln(T)/T}$. If T and M are large and K is small, the excess risk will be very small even for small sample sizes n.

The case of noiseless half-space learning illustrates this remark, allows comparison to a lower bound for single task learning, and to our knowledge provides the first theoretical proof of the superiority of multi-task learning under specific conditions. See Appendix A.

The proof of Theorem 1 is based on the well established method of Rademacher averages (Bartlett and Mendelson, 2002) and more recent advances on tail bounds for sums of random matrices, drawing upon the work of Ahlswede and Winter (2002), Oliveira (2010) and Tropp (2010). In this context two auxiliary results are established (Theorems 4 and 7 below), which may be of independent interest.

We list several important related results. Their proofs add little novelty to the proof of Theorem 1 and are deferred to an appendix.

1. The assumption of equal sample sizes for all tasks is often violated in practice. If n_t is the number of examples available for the t-th task the resulting imbalance can be compensated by a modification of the regularizer, replacing $\|W\|_1$ by a weighted trace norm $\|SW\|_1$, where the diagonal matrix $S = \mathrm{diag}(s_1, \ldots, s_T)$ weights the t-th task with $s_t = \sqrt{\bar{n}/n_t}$, with $\bar{n} = (1/T)\sum_t n_t$ being the average sample size. With this modification Theorem 1 holds with $\bar{n}$ in place of n. See Appendix C.
2. Instead of pre-assigned sample sizes n_t for each task t, one could generate an iid sample by choosing at random a task t and then an example (X, Y) from the task-specific distribution μ_t and repeating these two steps N times independently. For the risk as in (1) we can obtain a bound similar to Theorem 1 (i). See Appendix D.
3. The result mentioned in the previous remark can be specialized to matrix completion, where a matrix is only partially observed and estimated by a matrix of small trace norm (see e.g. Srebro and Shraibman, 2005; Recht, 2009; Candès and Tao, 2009; Shamir et al., 2011; Foygel et al., 2011, and references therein). If we take $\mathbb{H} = \mathbb{R}^d$, the input marginal as the uniform distribution supported on the basis vectors of $\mathbb{R}^d$, and the outputs as defined by the matrix values themselves, without or with the addition of noise, then the bound applies. This result, which can be generalized to certain inhomogeneous sampling distributions, is presented in Appendix E and matches the

bounds given by Foygel et al. (2011). This is important, because *tightness of the latter results thus implies tightness of our bounds.*

In addition to the proofs of these related results the appendix contains a few auxiliary lemmata and a short section on operator theory.

2. Earlier work.

The foundations to a theoretical understanding of multi-task learning were laid by Baxter (2000), where covering numbers are used to expose the potential benefits of multi-task and transfer learning. In (Ando and Zhang, 2005) Rademacher averages are used to give excess risk bounds for a method of multi-task subspace learning. Similar results are obtained in (Maurer, 2006a). Ben-David and Schuller (2003) use a special assumption of task-relatedness to give interesting bounds, not on the average but the maximal risk over the tasks.

A lot of important work on trace norm regularization concerns matrix completion. The connection of our results to matrix completion is discussed in Section E.

Multitask learning is considered in (Lounici et al., 2011), where special assumptions (coordinate-sparsity of the solution, restricted eigenvalues) are used to derive fast rates and the recovery of shared features. Such assumptions are absent in this paper, and (Lounici et al., 2011) also considers a different regularizer. Consistency of trace norm regularization has been studied by Bach (2008). Here we consider finite sample size guarantees.

(Maurer, 2006b) and (Kakade et al., 2012) seem to be most closely related to the present work. In (Maurer, 2006b) the general form of the bound is very similar to Theorem 1. The result is dimension independent, but it falls short of giving the rate of $\sqrt{\ln(T)/T}$ in the number of tasks. Instead it gives $T^{-1/4}$.

Kakade et al. (2012) introduce a general and elegant method to derive bounds for learning techniques which employ matrix norms as regularizers. For $\mathbb{H} = \mathbb{R}^d$, and applied to multi task learning and the trace-norm, a data-dependent bound is given whose dominant term reads as (omitting constants and observing that $\|W\|_1 \leq B\sqrt{T}$)

$$LB\sqrt{\max_i \left\|\hat{C}_i\right\|_\infty \frac{\ln \min\{T,d\}}{n}}, \tag{3}$$

where the matrix $\hat{C}_i$ is the empirical covariance of the data for all tasks observed in the i-th observation. The bound (3) does not paint a clear picture of the role of the number of tasks T. Using our methods we can estimate its expectation and convert it into a distribution dependent bound which resembles the bound in Theorem 1 (i). This is done in Appendix F. The principal disadvantage of (3) however is that it diverges in the simultaneous limit $d, T \to \infty$.

3. Notation and tools

The letter $\mathbb{H}$ will denote a finite or infinite dimensional separable real Hilbert space. For standard notation and results in operator theory see Appendix G. The reader who prefers to ignore the complications of infinite dimensions may take $\mathbb{H} = \mathbb{R}^d$ and regard all operators as

matrices. The main notational difference is that operator composition is left multiplication, thus for example if $A \in \mathcal{L}(\mathbb{H}, \mathbb{H}')$ then $A^*A \in \mathcal{L}(\mathbb{H})$ and not $A^*A \in \mathcal{L}(\mathbb{H}')$.

If $A \in \mathcal{L}(\mathbb{H})$ is self-adjoint, we denote by $\lambda_{\max}(A)$ its largest eigenvalue of A. If M is a closed subspace invariant under A, that is $AM \subseteq M$, we define: $\operatorname{tr}_M(A) = \sum_i \langle Ae_i, e_i \rangle$, where $\{e_i\}$ is a orthonormal basis of M. For $\operatorname{tr}_{\mathbb{H}}(A)$ we simply write $\operatorname{tr}(A)$.

Finally, for $w \in \mathbb{H}$ we define an operator[1] $Q_w \in \mathcal{L}(\mathbb{H})$ by

$$Q_w v = \langle v, w \rangle w, \text{ for } v \in \mathbb{H}.$$

In this notation the covariance operators in Theorem 1 are given by $C = \frac{1}{T}\sum_t \mathbb{E}_{(X,Y)\sim\mu_t} Q_X$ and $\hat{C} = \frac{1}{nT}\sum_{t,i} Q_{X_i^t}$, respectively.

The symbols σ_i or σ_i^t will always stand for Rademacher variables which are uniformly distributed on $\{-1, 1\}$, mutually independent and independent of all other random variables, and $\mathbb{E}_\sigma$ is the expectation conditional on all other random variables present. Two numbers $p, q > 1$ are called conjugate exponents if $1/p + 1/q = 1$.

We will use the following important result of Tropp (Tropp, 2010, Lemma 3.4), derived from Lieb's concavity theorem (Bhatia, 1997, Section IX.6):

Theorem 2 *Consider a finite sequence A_k of independent, random, self-adjoint operators and a finite dimensional subspace $M \subseteq \mathbb{H}$ such that $A_k M \subseteq M$. Then for $\theta \in \mathbb{R}$*

$$\mathbb{E}\ \operatorname{tr}_M\ \exp\left(\theta \sum_k A_k\right) \leq \operatorname{tr}_M\ \exp\left(\sum_k \ln \mathbb{E} e^{\theta A_k}\right).$$

A corollary suited to our applications is the following

Theorem 3 *Let $A_1, \ldots, A_N$ be independent, random, self-adjoint operators on $\mathbb{H}$ and let $M \subseteq \mathbb{H}$ be a nontrivial, finite dimensional subspace such that $\operatorname{Ran}(A_k) \subseteq M$ a.s. for all k.*

(i) If $A_k \succeq 0$ a.s then

$$\mathbb{E}\ \exp\left(\left\|\sum_k A_k\right\|\right) \leq \dim(M) \exp\left(\lambda_{\max}\left(\sum_k \ln \mathbb{E} e^{A_k}\right)\right).$$

(ii) If the A_k are symmetrically distributed then

$$\mathbb{E}\ \exp\left(\left\|\sum_k A_k\right\|\right) \leq 2\dim(M) \exp\left(\lambda_{\max}\left(\sum_k \ln \mathbb{E} e^{A_k}\right)\right).$$

Proof Let $A = \sum_k A_k$. Observe that $M^\perp \subseteq \operatorname{Ker}(A) \cap (\cup_k \operatorname{Ker}(A_k))$, and that M is a nontrivial invariant subspace for A as well as for all the A_k.

(i) Assume $A_k \succeq 0$. Then also $A \succeq 0$. Since $M^\perp \subseteq \operatorname{Ker}(A)$ there is $x_1 \in M$ with $\|x_1\| = 1$ and $Ax_1 = \|A\| x_1$ (this also holds if $A = 0$, since M is nontrivial). Thus $e^A x_1 = e^{\|A\|} x_1$. Extending x_1 to a basis $\{x_i\}$ of M we get

$$e^{\|A\|} = \langle e^A x_1, x_1 \rangle \leq \sum_i \langle e^A x_i, x_i \rangle = \operatorname{tr}_M\ e^A.$$

1. In matrix notation this would be the matrix $ww^\top$. It can also be written as the tensor product $w \otimes w$. We opted for the less usual notation Q_w as it will save space in many of the formulas below.

Theorem 2 applied to the matrices which represent A_k restricted to the finite dimensional invariant subspace M then gives

$$\begin{aligned}\mathbb{E}\exp\left(\|A\|\right) &\leq \mathbb{E}\mathrm{tr}_M \exp(A) \\ &\leq \mathrm{tr}_M \exp\left(\sum_k \ln\left(\mathbb{E}e^{A_k}\right)\right) \leq \dim(M)\exp\left(\lambda_{\max}\Big(\sum_k \ln\left(\mathbb{E}e^{A_k}\right)\Big)\right),\end{aligned}$$

where the last inequality results from bounding tr_M by $\dim(M)\,\lambda_{\max}$ and $\lambda_{\max}(\exp(\cdot)) = \exp(\lambda_{\max}(\cdot))$.

(ii) Our hypotheses imply that A is symmetrically distributed and $M^\perp \subseteq \mathrm{Ker}(A)$. Hence, there is $x_1 \in M$ with $\|x_1\| = 1$ and either $Ax_1 = \|A\|\,x_1$ or $-Ax_1 = \|A\|\,x_1$, so that either $e^A x_1 = e^{\|A\|}x_1$ or $e^{-A}x_1 = e^{\|A\|}x_1$. Extending to a basis again gives

$$e^{\|A\|} \leq \left\langle e^A x_1, x_1\right\rangle + \left\langle e^{-A}x_1, x_1\right\rangle \leq \mathrm{tr}_M e^A + \mathrm{tr}_M e^{-A}.$$

Taking the expectation we conclude that $\mathbb{E}e^{\|A\|} \leq 2\mathbb{E}\mathrm{tr}_M\, e^A$. Then continue as in case (ii). ■

4. Sums of random operators

In this section we prove two concentration results for sums of nonnegative operators with finite dimensional ranges. The first (Theorem 4) assumes only a weak form of boundedness, but it is strongly dimension dependent. Similar results appear in Tropp (2010), but they do not quite apply to our case. The second result (Theorem 7) requires strong boundedness, but is independent of the ambient dimension.

Theorem 4 *Let $M \subseteq \mathbb{H}$ be a subspace of dimension $d < \infty$ and suppose that $A_1, \ldots, A_N$ are independent random operators satisfying $A_k \succeq 0$, $\mathrm{Ran}(A_k) \subseteq M$ a.s. and*

$$\mathbb{E}A_k^m \preceq m! R^{m-1}\mathbb{E}A_k \tag{4}$$

for some $R \geq 0$, all $m \in \mathbb{N}$ and all $k \in \{1, \ldots, N\}$. Then for $s \geq 0$ and conjugate exponents p and q

$$\Pr\left\{\left\|\sum_k A_k\right\|_\infty > p\left\|\mathbb{E}\sum_k A_k\right\|_\infty + s\right\} \leq de^{-s/(qR)}.$$

Also

$$\sqrt{\mathbb{E}\left\|\sum_k A_k\right\|_\infty} \leq \sqrt{\left\|\mathbb{E}\sum_k A_k\right\|_\infty} + \sqrt{R(\ln d + 1)}.$$

Proof Let θ be any number satisfying $0 \leq \theta < \frac{1}{R}$. From (4) we get for any $k \in \{1, \ldots, N\}$

$$\begin{aligned}\mathbb{E}e^{\theta A_k} &= I + \sum_{m=1}^\infty \frac{\theta^m}{m!}\mathbb{E}A_k^m \preceq I + \sum_{m=1}^\infty (\theta R)^m\left(R^{-1}\mathbb{E}A_k\right) \\ &= I + \frac{\theta}{1-R\theta}\mathbb{E}A_k \preceq \exp\left(\frac{\theta}{1-R\theta}\mathbb{E}A_k\right).\end{aligned}$$

Abbreviate $\mu = \|\mathbb{E}\sum_k A_k\|_\infty$ and let $r = s + p\mu$ and set $\theta = (1/R)\left(1 - \sqrt{\mu/r}\right)$, so that $0 \leq \theta < 1/R$. Applying the above operator inequalities and the operator monotonicity of the logarithm (see e.g. Bhatia, 1997) we get for all k that $\ln \mathbb{E}\exp(\theta A_k) \preceq \theta/(1 - R\theta)\,\mathbb{E}A_k$. Summing this relation over k and passing to the largest eigenvalue yields

$$\lambda_{\max}\left(\sum_k \ln \mathbb{E}e^{\theta A_k}\right) \leq \frac{\theta\mu}{1 - R\theta}.$$

Combining Markov's inequality, Theorem 3 (i) and the last inequality gives

$$\begin{aligned}
\Pr\left\{\left\|\sum A_k\right\|_\infty \geq r\right\} &\leq e^{-\theta r}\mathbb{E}\,\exp\left(\theta\left\|\sum_k A_k\right\|\right) \\
&\leq de^{-\theta r}\exp\left(\lambda_{\max}\left(\sum_k \ln \mathbb{E}e^{\theta A_k}\right)\right) \\
&\leq d\exp\left(-\theta r + \frac{\theta\mu}{1 - R\theta}\right) \\
&= d\exp\left(\frac{-1}{R}\left(\sqrt{r} - \sqrt{\mu}\right)^2\right).
\end{aligned}$$

By Lemma 8 (i) $\left(\sqrt{r} - \sqrt{\mu}\right)^2 = \left(\sqrt{s + p\mu} - \sqrt{\mu}\right)^2 \geq s/q$, so this proves the first conclusion. The second result follows from the first one and Lemma 9. ■

We will use this result to prove Theorem 1 (ii) by applying it to sums of rank-one operators of the form Q_V where $V = \sum_i \sigma_i x_i$, the σ_i are Rademacher variables and the $x_i \in \mathbb{H}$ are bounded. To pave the way we show that the Q_V do indeed satisfy the subexponential condition (4).

Lemma 5 *Let $x_1, \ldots, x_n$ be in $\mathbb{H}$ and satisfy $\|x_i\| \leq b$. Define a random vector by $V = \sum_i \sigma_i x_i$. Then for every $m \geq 1$, it holds that*

$$\mathbb{E}\left[(Q_V)^m\right] \preceq m!\left(2nb^2\right)^{m-1}\mathbb{E}\left[Q_V\right].$$

Proof Let $K_{m,n}$ be the set of all sequences $\mathbf{j} := (j_1, \ldots, j_{2m})$ with $j_k \in \{1, \ldots, n\}$, such that each integer in $= \{1, \ldots, n\}$ occurs an even number of times. It is easily shown by induction (Lemma 10) that the number of sequences in $K_{m,n}$ is bounded by $(2m-1)!!n^m$, where $(2m-1)!! = \prod_{i=1}^m (2i-1) \leq m!2^{m-1}$.

Now let $v \in \mathbb{H}$ be arbitrary. By the definition of V and Q_V we have for any $v \in \mathbb{H}$ that

$$\left\langle \mathbb{E}\left[(Q_V)^m\right]v, v\right\rangle = \sum_{j_1,\ldots,j_{2m}=1}^{n} \mathbb{E}\left[\sigma_{j_1}\sigma_{j_2}\cdots\sigma_{j_{2m}}\right]\left\langle v, x_{j_1}\right\rangle\left\langle x_{j_2}, x_{j_3}\right\rangle\ldots\left\langle x_{j_{2m}}, v\right\rangle.$$

The properties of independent Rademacher variables imply that $\mathbb{E}\left[\sigma_{j_1}\sigma_{j_2}\cdots\sigma_{j_{2m}}\right] = 1$ if $j \in K_{m,n}$ and zero otherwise. For $m = 1$ this shows that $\left\langle \mathbb{E}\left[Q_V\right]v, v\right\rangle = \sum_{j=1}^n \left\langle v, x_j\right\rangle^2$. For

$m > 1$, since $\|x_i\| \leq b$ and by two applications of the Cauchy-Schwarz inequality

$$\begin{aligned}
\langle \mathbb{E}\left[(Q_V)^m\right] v, v\rangle &= \sum_{\mathbf{j}\in K_{m,n}} \langle v, x_{j_1}\rangle \langle x_{j_2}, x_{j_3}\rangle \cdots \langle x_{j_{2m}}, v\rangle \\
&\leq b^{2(m-1)} \sum_{\mathbf{j}\in K_{m,n}} |\langle v, x_{j_1}\rangle| \, |\langle x_{j_{2m}}, v\rangle| \\
&\leq b^{2(m-1)} \left[\sum_{\mathbf{j}\in K_{m,n}} \langle v, x_{j_1}\rangle^2\right]^{1/2} \left[\sum_{\mathbf{j}\in K_{m,n}} \langle v, x_{j_{2m}}\rangle^2\right]^{1/2} \\
&= b^{2(m-1)} \sum_{j=1}^{n} \langle v, x_j\rangle^2 \sum_{\mathbf{j}\in K_{m,n} \text{ s.t. } j_1=j} 1 \\
&= (2m-1)!! \left(nb^2\right)^{m-1} \langle \mathbb{E}\left[Q_V\right] v, v\rangle \;\leq\; m! \left(2nb^2\right)^{m-1} \langle \mathbb{E}\left[Q_V\right] v, v\rangle .
\end{aligned}$$

The result follows since for self-adjoint operators $A \preceq B \iff \langle Av, v\rangle \leq \langle Bv, v\rangle, \forall \in \mathbb{H}$. ■

The following is the key to the proof of Theorem 1 and related results.

Proposition 6 *Let $n_1, \ldots, n_T \in \mathbb{N}$, $N = \min\{\sum_t n_t, \dim \mathbb{H}\}$ and let $x_i^t \in \mathbb{H}$, $t = 1, \ldots, T$ $i = 1, \ldots, n_t$ such that $\left\|x_i^t\right\| \leq b_t$, for some constants $b_t \geq 0$. Define the random operator $D : \mathbb{H} \to \mathbb{R}^T$ by*

$$(Dy)_t = \left\langle y, \sum_{i=1}^{n_t} \sigma_i^t x_i^t \right\rangle .$$

Then

$$\mathbb{E}\,\|D\|_\infty \leq \sqrt{\left\|\sum_t \sum_{i=1}^{n_t} Q_{x_i^t}\right\|_\infty} + \sqrt{2 \max_t \left\{n_t b_t^2\right\} (\ln N + 1)}.$$

Proof Let V_t be the random vector $V_t = \sum_{i=1}^{n_t} \sigma_i^t x_i^t$ and recall that the corresponding rank-one operator Q_{V_t} is defined, for every $w \in \mathbb{H}$, as $Q_{V_t} w = \langle w, V_t\rangle V_t = \left\langle w, \sum_{i=1}^{n_t} \sigma_i^t x_i^t\right\rangle \sum_{i=1}^{n_t} \sigma_i^t x_i^t$. Then $D^*D = \sum_{t=1}^{T} Q_{V_t}$, so by Jensen's inequality

$$\mathbb{E}\,\|D\|_\infty = \mathbb{E}\sqrt{\|D^*D\|_\infty} \leq \sqrt{\left\|\sum_t Q_{V_t}\right\|_\infty} .$$

Since $\mathrm{Ran}(Q_{V_t}) \subseteq \mathrm{Span}(\{x_i^t, i = 1, \ldots, n_t,\ t = 1, \ldots, T\})$, Lemma 5 yields that

$$\mathbb{E}\left[(Q_{Vt})^m\right] \preceq m! \left(2n_t b_t^2\right)^{m-1} \mathbb{E}\left[Q_{V_t}\right] \preceq m! \left(2 \max_t \left\{n_t b_t^2\right\}\right)^{m-1} \mathbb{E}\left[Q_{V_t}\right] .$$

We can then apply Theorem 4 with $R = 2 \max_t \left\{n_t b_t^2\right\}$ and $d = N$ to conclude that

$$\sqrt{\mathbb{E}\left\|\sum_t Q_{V_t}\right\|_\infty} \leq \sqrt{\left\|\mathbb{E}\sum_t Q_{V_t}\right\|_\infty} + \sqrt{2 \max_t \left\{n_t b_t^2\right\} (\ln N + 1)},$$

which implies the result, since $\mathbb{E}Q_{V_t} = \sum_{i=1}^{n_t} Q_{x_i^t}$. ■

To pass from the data-dependent bound Theorem 1 (ii) to the distribution dependent bound (i) wee need the next result. Its proof builds upon (Oliveira, 2010, Lemma 1), but see also Mendelson and Pajor (2006). We give a slightly more general version which eliminates the assumption of identical distribution and has smaller constants.

Theorem 7 *Let $A_1, \ldots, A_N$ be independent random operators satisfying $0 \preceq A_k \preceq I$ and suppose that for some $d \in \mathbb{N}$*

$$\dim \operatorname{Span}\left(\operatorname{Ran}\left(A_1\right), \ldots, \operatorname{Ran}\left(A_N\right)\right) \leq d \tag{5}$$

almost surely. Then

$$(i) \qquad \Pr\left\{\left\|\sum_k \left(A_k - \mathbb{E}A_k\right)\right\|_\infty > s\right\} \leq 4d^2 \exp\left(\frac{-s^2}{9\left\|\sum_k \mathbb{E}A_k\right\|_\infty + 6s}\right);$$

$$(ii) \qquad \Pr\left\{\left\|\sum_k A_k\right\|_\infty > p\left\|\mathbb{E}\sum_k A_k\right\|_\infty + s\right\} \leq 4d^2 e^{-s/(6q)};$$

$$(iii) \qquad \sqrt{\mathbb{E}\left\|\sum_k A_k\right\|_\infty} \leq \sqrt{\left\|\mathbb{E}\sum_k A_k\right\|_\infty} + \sqrt{6\left(\ln\left(4d^2\right)+1\right)}.$$

In the previous theorem the subspace M was deterministic and had to contain the ranges of *all* possible random realizations of the A_k. By contrast the span appearing in (5) is the random subspace spanned by a single random realization of the A_k. If all the A_k have rank one, for example, we can take $d = N$ and apply the present theorem even if each $\mathbb{E}A_k$ has infinite rank. This allows to estimate the empirical covariance in terms of the true covariance for a bounded data distribution in an infinite dimensional space.

Proof Let $0 \leq \theta < 1/4$ and abbreviate $A = \sum_k A_k$. A standard symmetrization argument (see Ledoux and Talagrand, 1991, Lemma 6.3) shows that

$$\mathbb{E}e^{\theta\|A-\mathbb{E}A\|} \leq \mathbb{E}\mathbb{E}_\sigma \exp\left(2\theta\left\|\sum_k \sigma_k A_k\right\|\right),$$

where the σ_k are Rademacher variables and $\mathbb{E}_\sigma$ is the expectation conditional on the $A_1, \ldots, A_N$. For fixed $A_1, \ldots, A_N$ let M be the linear span of their ranges, which has dimension at most d and also contains the ranges of the symmetrically distributed operators $2\theta\sigma_k A_k$. Invoking Theorem 3 (ii) we get

$$\begin{aligned}
\mathbb{E}_\sigma \exp\left(2\theta\left\|\sum_k \sigma_k A_k\right\|\right) &\leq 2d \exp\left(\lambda_{\max}\left(\sum_k \ln \mathbb{E}_\sigma e^{2\theta\sigma_k A_k}\right)\right) \\
&\leq 2d \exp\left(2\theta^2\left\|\sum_k A_k^2\right\|\right) \leq 2d \exp\left(2\theta^2\|A\|\right).
\end{aligned}$$

The second inequality comes from $\mathbb{E}_\sigma e^{2\theta\sigma_k A_k} = \cosh(2\theta A_k) \preceq e^{2\theta^2 A_k^2}$, the operator monotonicity of the logarithm and the fact that for positive operators $\lambda_{\max}$ and the norm coincide. The last inequality follows from the implications

$$0 \preceq A_k \preceq I \implies A_k^2 \preceq A_k \implies \sum_k A_k^2 \preceq \sum_k A_k \implies \left\| \sum_k A_k^2 \right\| \leq \|A\|.$$

Now we take the expectation in $A_1, \ldots, A_N$. Together with the previous inequalities we obtain

$$\mathbb{E}e^{\theta\|A-\mathbb{E}A\|} \leq 2d\mathbb{E}e^{2\theta^2\|A\|} \leq 2d\mathbb{E}e^{2\theta^2\|A-\mathbb{E}A\|}e^{2\theta^2\|\mathbb{E}A\|} \leq 2d\left(\mathbb{E}e^{\theta\|A-\mathbb{E}A\|}\right)^{2\theta} e^{2\theta^2\|\mathbb{E}A\|}.$$

The last inequality holds by Jensen's inequality since $\theta < 1/4 < 1/2$. Dividing both sides of the above series of inequalities by $\left(\mathbb{E}\exp\left(\theta\|A-\mathbb{E}A\|\right)\right)^{2\theta}$, taking the power of $1/(1-2\theta)$ and multiplying with $e^{\theta s}$ gives

$$\Pr\left\{\|A-\mathbb{E}A\| > s\right\} \leq e^{-\theta s}\mathbb{E}e^{\theta\|A-\mathbb{E}A\|} \leq (2d)^{1/(1-2\theta)} \exp\left(\frac{2\theta^2}{1-2\theta}\|\mathbb{E}A\| - \theta s\right).$$

Since $\theta < 1/4$, we have $(2d)^{1/(1-2\theta)} < (2d)^2$. Substitution of $\theta = s/\left(6\|\mathbb{E}A\| + 4s\right) < 1/4$ together with some simplifications gives (i).

It follows from elementary algebra that for $\delta > 0$ with probability at least $1-\delta$ we have

$$\|A\| \leq \|\mathbb{E}A\| + 2\sqrt{\|\mathbb{E}A\|}\sqrt{\frac{9}{4}\ln\left(4d^2/\delta\right)} + 6\ln\left(4d^2/\delta\right) \leq p\|\mathbb{E}A\| + 6q\ln\left(4d^2/\delta\right),$$

where the last line follows from $(9/4) < 6$ and Lemma 8 (iii). Equating the second term in the last line to s and solving for the probability δ we obtain (ii), and (iii) follows from Lemma 9. ■

5. Proof of Theorem 1

The first steps in the proof follow a standard pattern. We write

$$R(\hat{W}) - R(W^0) = \left[R(\hat{W}) - \hat{R}(\hat{W}, \bar{\mathbf{Z}})\right] + \left[\hat{R}(\hat{W}, \bar{\mathbf{Z}}) - \hat{R}(W^0, \bar{\mathbf{Z}})\right] + \left[\hat{R}(W^0, \bar{\mathbf{Z}}) - R(W^0)\right].$$

The second term is always negative by the definition of $\hat{W}$. The third term depends only on W^0. Using Hoeffding's inequality (Hoeffding, 1963) it can be bounded with probability at least $1-\delta$ by $\sqrt{\ln(1/\delta)/(2nT)}$. There remains the first term which we bound by $\sup_{W\in\mathcal{W}} R(W) - \hat{R}(W)$.

It has by now become a standard technique (see e.g. Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002) to show that this quantity is with probability at least $1-\delta$ bounded in terms of the Rademacher complexity by

$$\mathbb{E}_{\bar{\mathbf{Z}}}\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) + \sqrt{\frac{\ln(1/\delta)}{2nT}} \tag{6}$$

or, in terms of the empirical Rademacher complexity, by

$$\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) + \sqrt{\frac{9\ln\left(2/\delta\right)}{2nT}}, \tag{7}$$

where $\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right)$ is defined for a multisample $\bar{\mathbf{Z}}$ with values in $(\mathbb{H}\times\mathbb{R})^{nT}$ by

$$\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) = \frac{2}{nT}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{W\in\mathcal{W}} \sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_i^t \ell\left(\left\langle w_t, X_i^t\right\rangle, Y_i^t\right).$$

Standard results on Rademacher averages allow us to eliminate the Lipschitz loss functions and give us

$$\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{z}}\right) \leq \frac{2L}{nT}\mathbb{E}_{\sigma} \sup_{W\in\mathcal{W}} \sum_{t,i} \sigma_i^t \left\langle w_t, X_i^t\right\rangle = \frac{2L}{nT}\mathbb{E}_{\sigma} \sup_{W\in\mathcal{W}} \operatorname{tr}\left(W^* D\right),$$

where the random operator $D : \mathbb{H} \to \mathbb{R}^T$ is defined for $v \in \mathbb{H}$ by $(Dv)_t = \left\langle v, \sum_{i=1}^n \sigma_i^t X_i^t\right\rangle$. By Hölder's inequality (see Theorem 12) we have that

$$\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) \leq \frac{2L}{nT} \sup_{W\in\mathcal{W}} \|W\|_1 \mathbb{E}_{\sigma} \|D\|_{\infty} = \frac{2LB}{n\sqrt{T}}\mathbb{E}_{\sigma} \|D\|_{\infty}.$$

Now applying Proposition 6 with $n_t = n$, $X_i^t = x_i^t$ and $b_t = 1$ and using $\sum_{t,i} Q_{X_i^t} = nT\hat{C}$, we get

$$\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) \leq 2LB\left(\sqrt{\frac{\left\|\hat{C}\right\|_{\infty}}{n}} + \sqrt{\frac{2\left(\ln\left(nT\right)+1\right)}{nT}}\right), \tag{8}$$

which, together with (7), gives the second assertion of Theorem 1.

To obtain the first assertion we take the expectation of (8), which confronts us with the problem of bounding $\mathbb{E}\left\|\hat{C}\right\|_{\infty}$ in terms of $\|C\|_{\infty} = \left\|\mathbb{E}\hat{C}\right\|_{\infty}$. Note that $nT\hat{C} = \sum_{t,i} Q_{X_i^t}$. Here Theorem 4 doesn't help because the covariance may have infinite rank, so that we cannot find a finite dimensional subspace containing the ranges of all the $Q_{X_i^t}$. But since $\left\|X_i^t\right\| \leq 1$ all the $Q_{X_i^t}$ satisfy $0 \preceq Q_{X_i^t} \preceq I$ and are rank-one operators, we can invoke Theorem 7 with $d = nT$. Taking the expectation of (8), using Jensen's inequality, substitution of the conclusion of Theorem 7 with $d = nT$ and some simplifications give

$$\mathbb{E}\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) \leq 2LB\left(\sqrt{\frac{\|C\|_{\infty}}{n}} + 5\sqrt{\frac{\ln\left(nT\right)+1}{nT}}\right),$$

which, together with (6), gives the first assertion of Theorem 1.

Acknowledgments

This work was supported in part by EPSRC Grant EP/H027203/1 and Royal Society International Joint Project Grant 2012/R2.

References

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

Y. Amit, M. Fink, N. Srebro, S. Ullman. Uncovering shared structures in multiclass classification. *Proc. 24th International Conference on Machine Learning (ICML)*, pages 17–24, 2007.

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.

A. Argyriou, T. Evgeniou, M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

F.R. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.

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

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. *Proc. 16th Annual Conference on Computational Learning Theory (COLT)*, pages 567–580, 2003.

R. Bhatia. *Matrix Analysis*. Springer, 1997.

E. Candès and T. Tao. The power of convex relaxation: Near optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2009.

R. Caruana. Multi-task learning. *Machine Learning*, 28(1):41–75, 1997.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2597–2630, 2010.

T. Evgeniou, C. Micchelli and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. *Proc. American Control Conference*, Vol. 6, pages 4734–4739, 2001.

R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in Neural Information Processing Systems*, 24, pages 2133–2141, 2011.

Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, J. Malick. Large-scale image classification with trace-norm regularization. *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3386–3393, 2012.

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

S. M. Kakade, S. Shalev-Shwartz, A. Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research* 13:1865–1890, 2012.

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

M. Ledoux, M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, Berlin, 1991.

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

A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.

A. Maurer. The Rademacher complexity of linear transformation classes. *Proc. 19th Annual Conference on Learning Theory (COLT)*, pages 65–78, 2006.

A. Maurer and M. Pontil. A uniform lower error bound for half-space learning. *Proc. 19th International Conference on Algorithmic Learning Theory*, pages 70–78, 2008.

S. Mendelson and A. Pajor. On singular values of matrices with independent rows. *Bernoulli* 12(5):761–773, 2006.

R.I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson, *Electronic Communications in Probability,* 15:203–212, 2010.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2009.

O. Shamir and S. Shalev-Shwartz. Collaborative filtering with the trace norm: Learning, bounding and transducing. *Proc. 24th Annual Conference on Learning Theory (COLT)*, pages 661–678, 2011.

N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *Proc. 18th Annual Conference on Learning Theory (COLT)*, pages 545–560, 2005.

S. Thrun and L. Pratt. *Learning to Learn*. Springer, 1998.

J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

Appendix A. Half-space learning

In this section, we assume that all input marginals are given by the uniform distribution on the unit sphere $\mathcal{S}^{M-1}$ in $\mathbb{R}^M$ and the objective is for each task to classify membership in the half-space $\{x : \langle x, u_t \rangle > 0\}$ defined by a task-specific (unknown) unit vector u_t. All the u_t are constrained to lie in an (unknown) K-dimensional subspace of $\mathbb{R}^M$, the relevant common constraint in this case. We are interested in the regime

$$K \ll n \ll M \ll T.$$

An algorithm is orthogonally equivariant if for data transformed by an orthogonal transformation it produces a correpondingly transformed hypothesis. This class of algorithms includes all kernel methods, but it excludes the lasso and other algorithms which depend on a specific coordinate system. In (Maurer and Pontil, 2008) it is shown that for any orthogonally equivariant single-task algorithm the error is bounded below by $(1/\pi)\sqrt{(M-n)/M}$ with overwhelming probability, so in our regime single task learning provably fails with an error at best about $1/\pi$.

The 0-1-loss is unsuited for our bounds on multitask learning because it is not Lipschitz. Instead we will use the truncated hinge loss with margin $\epsilon/\sqrt{M}$, given by $\ell(y', y) = h(y'y)$, where h is the real function

$$h(t) = \begin{cases} 1 & \text{if} \quad t \leq 0, \\ 1 - t\sqrt{M}/\epsilon & \text{if} \quad 0 < t \leq \epsilon/\sqrt{M}, \\ 0 & \text{if} \quad \epsilon/\sqrt{M} < t. \end{cases}$$

Here ϵ is a parameter to be optimized later. This loss is an upper bound of the 0-1-loss. We use it to perform TNML and then threshold the linear functionals returned for each task at zero. The matrix $[u_1, \ldots, u_T]$ satisfies $\|[u_1, \ldots, u_T]\|_1 \leq \sqrt{TK}$, since the task specific target vectors $u_1, \ldots, u_T$ are unit vectors and all lie within a K-dimensional subspace of $\mathbb{R}^M$. Thus, if we set $B = \sqrt{K}$, the matrix $W^0 := [u_1, \ldots, u_T]^*$ is in the feasible set. The corresponding average risk is

$$R\left(W^0\right) \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X\sim\mu_M}\, h\left(|\langle u_t, X\rangle|\right) = \mathbb{E}_{X\sim\mu_M}\left[h\left(|\langle u_1, X\rangle|\right)\right],$$

where μ_M is the uniform distribution on the unit sphere $\mathcal{S}^{M-1}$ in $\mathbb{R}^M$, which is the input marginal for all tasks. This follows from invariance of μ_M. But the density of the distribution of $|\langle u_1, X\rangle|$ has maximum A_{M-1}/A_M, where A_M is the volume of $\mathcal{S}^{M-1}$ in the metric inherited from $\mathbb{R}^M$, which can be bounded by $\sqrt{M}$. Thus

$$R\left(W^0\right) \leq \sqrt{M}\int_{-\infty}^{\infty} h\left(|s|\right) ds = \epsilon,$$

and together with Theorem 1 the bound on the average risk becomes (omitting the confidence dependent term)

$$\epsilon + \frac{2}{\epsilon}\sqrt{\frac{K}{n}} + \frac{10}{\epsilon}\sqrt{\frac{KM\left(\ln\left(nT\right)+1\right)}{nT}}.$$

The burden of the high dimension M is carried exclusively by the last term, which is small in our regime because of the large number of tasks, regardless of the individual sample sizes n. The individual samples must only well outnumber the dimension K, roughly the number of shared features.

Letting $T \to \infty$ and optimizing in ϵ gives an upper bound for the average error of order $(K/n)^{1/4}$, which is small in the regime we consider, in contrast to the lower bound for single task learning.

Appendix B. Auxiliary results

Recall that $p, q > 0$ are called conjugate exponents if $1/p + 1/q = 1$.

Lemma 8 *Let p, q be conjugate exponents and $s, a, b \geq 0$, Then*

$$\begin{aligned} &(i) && \left(\sqrt{s+pa} - \sqrt{a}\right)^2 \geq s/q; \\ &(ii) && \min\left\{\sqrt{pa+qb} : p, q > 1,\ \frac{1}{p} + \frac{1}{q} = 1\right\} = \sqrt{a} + \sqrt{b}; \\ &(iii) && 2\sqrt{ab} \leq (p-1)\, a + (q-1)\, b. \end{aligned}$$

Proof For conjugate exponents p and q we have $p - 1 = p/q$ and $q - 1 = q/p$. Therefore $pa + qb - \left(\sqrt{a} + \sqrt{b}\right)^2 = \left(\sqrt{pa/q} - \sqrt{qb/p}\right)^2 \geq 0$, which proves (iii) and gives

$$\sqrt{pa+qb} \geq \sqrt{a} + \sqrt{b}. \tag{9}$$

Take $s = qb$, subtract $\sqrt{a}$ and square to get (i). Set $p = 1 + \sqrt{b/a}$ and $q = 1 + \sqrt{a/b}$ in (9) to get (ii). ■

Lemma 9 *Let $a, c > 0, b \geq 1$ and suppose the real random variable $X \geq 0$ satisfies $\Pr\{X > pa + s\} \leq b \exp(-s/(qc))$ for all $s \geq 0$ and all conjugate exponents p and q. Then*

$$\sqrt{\mathbb{E}X} \leq \sqrt{a} + \sqrt{c\,(\ln b + 1)}.$$

Proof We use partial integration.

$$\begin{aligned} \mathbb{E}X &\leq pa + qc \ln b + \int_{qc \ln b}^{\infty} \Pr\{X > pa + s\}\, ds \\ &\leq pa + qc \ln b + b \int_{qc \ln b}^{\infty} e^{-s/(qc)} ds \\ &= pa + qc\,(\ln b + 1)\,. \end{aligned}$$

Take the square root of both sides and use Lemma 8 (ii) to optimize in p and q to obtain the conclusion. ■

Lemma 10 *Let $m, n \in \mathbb{N}$ and let $K_{m,n}$ be the set of all sequences $\mathbf{j} = (j_1, \ldots, j_{2m})$ with $j_k \in \{1, \ldots, n\}$, in which each integer in $\{1, \ldots, n\}$ occurs an even number of times. Then*

$$|K_{m,n}| \leq (2m-1)!! n^m,$$

where $(2m-1)!! = \prod_{i=1}^{m} (2i-1)$.

Proof By induction on m. The case $m = 1$ is obvious. Assume it true for $m-1$, $m > 1$. For $i \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, 2m-1\}$ let $K_{m,n}(i,l)$ be the set of those sequences $\mathbf{j} \in K_{m,n}$ such that $j_l = j_{2m} = i$. Now for every $\mathbf{j} \in K_{m,n}$ the index j_{2m} must have some value i which occurs at least twice. Thus

$$K_{m,n} \subseteq \bigcup_{l=1}^{2m-1} \bigcup_{i=1}^{n} K_{m,n}(i,l). \tag{10}$$

For $l \in \{1, \ldots, 2m-1\}$ let $\pi_l : K_{m,n} \to K_{m,n}$ be the map which exchanges j_l and j_{2m-1}. Then

$$\pi_l(K_{m,n}(i,l)) = \left\{ \left(j_1, \ldots, j_{2(m-1)}, i, i\right) : \left(j_1, \ldots, j_{2(m-1)}\right) \in K_{m-1,n} \right\}.$$

Since π_l is a bijection $|K_{m,n}(i,l)| = |\pi_l(K_{m,n}(i,l))| = |K_{m-1,n}| \leq (2(m-1)-1)!! n^{m-1}$, by induction hypothesis. The result thus follows from (10). ∎

Appendix C. Unequal sample sizes and weighted trace norm

We prove the excess risk bound for heterogeneous sample sizes with a weighted trace norm. The sample size for the n-th task is n_t and we abbreviate $\bar{n}$ for the average sample size, $\bar{n} = (1/T)\sum_t n_t$, so that $\bar{n}T$ is the total number of examples. The class of linear maps W considered is

$$\mathcal{W} = \left\{ W \in \mathcal{L}(\mathbb{H}, \mathbb{R}^T) : \|SW\|_1 \leq B\sqrt{T} \right\},$$

with $S = \operatorname{diag}(s_1, \ldots, s_T)$ and $s_t = \sqrt{\bar{n}/n_t}$. With $\mathcal{W}$ so defined we will prove the inequalities in Theorem 1 with n replaced by $\bar{n}$. It should not be surprising that the tasks are regularized proportional to the inverse square root of their sample sizes.

The first steps in the proof are the same as in Section 5. The empirical Rademacher average which we now have to bound is

$$\begin{aligned} \mathcal{R}(\mathcal{W}, \bar{\mathbf{Z}}) &= \frac{2}{T}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{W \in \mathcal{W}} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \sigma_i^t \ell\left(\langle w_t, X_i^t \rangle, Y_i^t\right) \\ &\leq \frac{2L}{T}\mathbb{E}_{\sigma} \sup_{W \in \mathcal{W}} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sigma_i^t \langle w_t, X_i^t/n_t \rangle = \frac{2L}{T}\mathbb{E}_{\sigma} \sup_{W \in \mathcal{W}} \operatorname{tr}(W^* S D). \end{aligned}$$

where the random operator $D : \mathbb{H} \to \mathbb{R}^T$ is now defined for $v \in \mathbb{H}$ by

$$(Dv)_t = \left\langle v, \sum_{i=1}^{n_t} \frac{\sigma_i^t X_i^t}{s_t n_t} \right\rangle$$

and the diagonal matrix S is as above. From Hölder's inequality we get

$$\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) \leq \frac{2L}{T} \sup_{W \in \mathcal{W}} \|SW\|_1 \, \mathbb{E}_\sigma \|D\|_\infty = \frac{2LB}{\sqrt{T}} \mathbb{E}_\sigma \|D\|_\infty$$

and applying Proposition 6 with $X_i^t/\left(s_t n_t\right)$ in place of x_i^t and $b_t = 1/\left(s_t n_t\right) = 1/\sqrt{\bar{n} n_t}$ yields

$$\mathbb{E}_\sigma \|D\|_\infty \leq \sqrt{\frac{1}{\bar{n}} \left\| \sum_t \frac{1}{n_t} \sum_{i=1}^{n_t} Q_{X_i^t} \right\|_\infty} + \sqrt{\frac{2\left(\ln\left(\bar{n}T\right)+1\right)}{\bar{n}}}.$$

Since $\sum_t \left(1/\bar{n}\right) \sum_i \left(1/n_t\right) Q_{X_i^t} = T\hat{C}/\bar{n}$, we obtain

$$\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) \leq 2LB \left(\sqrt{\frac{\left\|\hat{C}\right\|_\infty}{\bar{n}}} + \sqrt{\frac{2\left(\ln\left(\bar{n}T\right)+1\right)}{\bar{n}T}} \right),$$

which gives the second assertion of Theorem 1 with n replaced by $\bar{n}$. The first assertion follows exactly as before.

Appendix D. IID sampling

We sketch the application to iid sampled multi-task learning. Now the sample $\bar{\mathbf{Z}}$ is generated by selecting a task t_j at random, sampling a single example pair $\left(X_{i_j}^{t_j}, Y_{i_j}^{t_j}\right)$ from μ_{t_j} and repeating this process N times independently. The numbers n_t (size of sample available for task t) now become random variables. The idea is that for sufficiently large N with overwhelming probability all tasks will have sample sizes bounded by $2N/T$. When conditioning on the sample sizes n_t we can apply our method in the "normal case" and bound the expectation for the "pathological case" using its small probability. We give a corresponding distribution-dependent bound.

Theorem 11 *Assume $N \ln N \geq (8/3)\, T$ and let $N_0 = \min\left\{N, \dim\left(\mathbb{H}\right)\right\}$. For $\delta > 0$ with probability at least $1-\delta$ in $\bar{\mathbf{Z}}$ generated as described above*

$$R(\hat{W}) - R(W^0) \leq 2LB \left(\sqrt{\frac{T\|C\|_\infty}{N}} + 6\sqrt{\frac{\ln N_0 + 2}{N}} \right) + \frac{4T}{N} + \sqrt{\frac{2\ln\left(2/\delta\right)}{N}},$$

Proof For any task t we have $n_t = \sum_{i=1}^N \beta_i$, where the β_i are independent Bernoulli variables with $\mathbb{E}\beta_i = 1/T$. From Bernstein's inequality we find

$$\Pr\left\{ n_t > \frac{2N}{T} \right\} \leq \exp\left(\frac{-3N}{8T}\right).$$

Let *Bad* be the event that there exists a task t with $n_t > 2N/T$ and *Good* its complement. A union bound gives

$$\Pr\left(Bad\right) \leq T \exp\left(\frac{-3N}{8T}\right) \leq \frac{T}{N},$$

where the second inequality follows from $N \ln N \geq (8/3)\, T$. Let Σ be the σ-algebra generated by the variables $\{n_t : t \in \{1, \ldots, T\}\}$. Observe that Bad, $Good \in \Sigma$. Define a Σ-measurable random variable by

$$F = \mathbb{E}\left[\sup_{W \in \mathcal{W}} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sigma_i^t \ell\left(\left\langle w_t, X_i^t \right\rangle, Y_i^t\right) | \Sigma\right].$$

The distribution dependent Rademacher average we wish to bound is

$$\mathbb{E}\mathcal{R}\left(\mathcal{W}, \bar{\mathbf{Z}}\right) \quad = \quad \frac{2}{N}\mathbb{E}F = \frac{2}{N}\mathbb{E}\left[1_{Good}F\right] + \frac{2}{N}\mathbb{E}\left[1_{Bad}F\right] \leq \frac{2}{N}\mathbb{E}\left[1_{Good}F\right] + \frac{2T}{N},$$

where the inequality follows from the fact that F is bounded by N and the bound on $\Pr(Bad)$. In bounding $\mathbb{E}1_{Good}F$ we can assume that all the sample sizes are bounded by $2N/T$. The bound then follows the same steps as the proof of Theorem 1. Note that

$$1_{Good}F \leq L\mathbb{E}\left[\sup_{W \in \mathcal{W}} \operatorname{tr}(WD) | \Sigma\right],$$

where $D : \mathbb{H} \to \mathbb{R}^T$ is defined as $(Dv)_t = \left\langle v, \sum_{i=1}^{n_t} \sigma_i^t X_i^t \right\rangle$. With Hölder's inequality we get, as in the proof of Theorem 1,

$$1_{Good}F \;\leq\; LB\sqrt{T}\mathbb{E}\left[\|D\|_\infty | \Sigma\right] \;\leq\; LB\sqrt{T}\mathbb{E}\left[\sqrt{\left\|\sum_{t=1}^{T}\sum_{i=1}^{n_t} Q_{X_i^t}\right\|_\infty} + 2\sqrt{\frac{N}{T}(\ln N_0 + 1)} | \Sigma\right].$$

In the second inequality we used Proposition 6 with $b_t = 1$ and the fact that $n_t \leq 2N/T$ on the event $Good$. With Jensen's inequality and Theorem 7 (iii) we get

$$\mathbb{E}\left[\sqrt{\left\|\sum_{t=1}^{T}\sum_{i=1}^{n_t} Q_{X_i^t}\right\|_\infty} | \Sigma\right] \leq \sqrt{N\|C\|_\infty} + \sqrt{6\left(\ln\left(4N_0^2\right) + 1\right)}.$$

It follows that

$$1_{Good}F \leq LB\left(\sqrt{NT\|C\|_\infty} + 6\sqrt{N(\ln N_0 + 2)}\right).$$

Observe that this bound is now a constant, independent of the $n_1, \ldots, n_t$ and is the same for $\mathbb{E}1_{Good}F$, whence

$$\mathbb{E}\mathcal{R}_N(\mathcal{W}, \mu) \leq 2LB\left(\sqrt{\frac{T\|C\|_\infty}{N}} + 6\sqrt{\frac{\ln N_0 + 2}{N}}\right) + \frac{2T}{N}.$$

■

Appendix E. Matrix completion

We specialize the result of the previous section to matrix completion, assuming at first a uniform sampling distribution for the entries of a $T \times d$ matrix. We replace $\mathbb{H}$ by $\mathbb{R}^d$ and since the inputs are sampled uniformly from the basis vectors in $\mathbb{R}^d$ we can replace $\|C\|_\infty$ by $1/d$. To obtain a bound comparable to those in (Foygel et al., 2011) we replace the condition $\|W\|_1 \leq B\sqrt{T}$ by $\|W\|_1 \leq B\sqrt{dT}$. With these values the dominant term in the bound of the previous section becomes

$$2B\left(\sqrt{\frac{T}{N}} + 6\sqrt{\frac{d\left(\ln \min\{d, N\} + 2\right)}{N}}\right) + \frac{2T}{N}.$$

With $T \leq d \leq N$ this is $O(d \ln d/N)$, in agreement with other performance guarantees for matrix completion (Srebro and Shraibman, 2005; Recht, 2009).

The present argument addresses only the case of a uniform sampling distribution. If the sampling distribution is the product of row and column distributions, we can just as in (Foygel et al., 2011) employ a weighted trace norm $\|S_1 W S_2\|_1$, where S_1 compensates the inhomogeneities in the tasks (T-axis), much as in Section C above, and S_2 compensates inhomogeneities in the coordinates (d-axis), essentially ensuring that the covariance of $S_2 X$ is $1/d$ times the identity. In this way the above bound is reproduced. The corresponding proof consists largely of repetitions of arguments already presented in this paper and is omitted.

Appendix F. An alternative bound

Finally we consider the bound on the empirical Rademacher complexity proposed by Kakade et al. (2012). In our notation it reads (omitting constants)

$$\mathcal{R}\left(\mathcal{M}, \bar{\mathbf{Z}}\right) \leq LB\sqrt{\max_i \left\|\hat{C}_i\right\|_\infty \frac{\ln \min\{T, d\}}{n}},$$

where the matrix $\hat{C}_i$ is the empirical covariance of the data for all tasks observed, restricted to the i-th example, that is

$$\hat{C}_i = \frac{1}{T}\sum_t Q_{X_i^t}.$$

While the bound does not clearly spell out the role of the number T of tasks, it can be used to obtain a bound similar to Theorem 1 by passage to the expected Rademacher complexity. This involves the expectation $\mathbb{E}\max_i \left\|\sum_t Q_{X_i^t}\right\|_\infty$. Note that $\sum_t \mathbb{E} Q_{X_i^t} = TC$. Just as at the end of Section 5 we can apply Theorem 7 with $d = T$. We get for some parameter η

and conjugate exponents p and q

$$\begin{aligned}\mathbb{E}\max_i \left\|\sum_t Q_{X_i^t}\right\|_\infty &\leq pT\left\|C\right\|_\infty + \eta + \\ &\quad + \int_\eta^\infty \Pr\left\{\max_{1\leq i\leq n}\left\|\sum_t Q_{X_i^t}\right\|_\infty > pT\left\|C\right\|_\infty + s\right\} ds \\ &\leq pT\left\|C\right\|_\infty + \eta + 4nT^2\int_\eta^\infty e^{-s/(6q)} ds \\ &\leq pT\left\|C\right\|_\infty + q\left(6\ln\left(24nT^2\right)+1\right),\end{aligned}$$

if we choose $\eta = 6q\ln\left(24nT^2\right)$. With Lemma 8 (ii) we get

$$\sqrt{\mathbb{E}\max_i \left\|\sum_t \hat{C}_i\right\|_\infty} \leq \sqrt{\left\|C\right\|_\infty} + \sqrt{\frac{6\ln\left(24nT^2\right)+1}{T}}.$$

Substitution then gives (up to a constant)

$$\mathbb{E}\mathcal{R}\left(\mathcal{M}, \bar{\mathbf{Z}}\right) \leq LB\sqrt{\ln\min\left\{T,d\right\}}\left(\sqrt{\frac{\left\|C\right\|_\infty}{n}} + \sqrt{\frac{6\ln\left(24nT^2\right)+1}{nT}}\right),$$

which resembles the bound in Theorem 1 (i). Note however that the bound diverges in the simultaneous limits $d\to\infty$ and $T\to\infty$.

Appendix G. Some elements of operator theory

The letters $\mathbb{H}$, $\mathbb{H}'$, $\mathbb{H}''$ will denote finite or infinite dimensional separable real Hilbert spaces. With $\mathcal{L}\left(\mathbb{H},\mathbb{H}'\right)$ we denote the set of linear transformations $A:\mathbb{H}\to\mathbb{H}'$ satisfying $\left\|A\right\|_\infty := \sup_{\left\|x\right\|\leq 1}\left\|Ax\right\| < \infty$. $\mathcal{L}\left(\mathbb{H},\mathbb{H}\right)$ is abbreviated as $\mathcal{L}\left(\mathbb{H}\right)$ and its members are called operators. For $A\in\mathcal{L}\left(\mathbb{H},\mathbb{H}'\right)$ and $B\in\mathcal{L}\left(\mathbb{H}',\mathbb{H}''\right)$ the product $BA\in\mathcal{L}\left(\mathbb{H},\mathbb{H}''\right)$ is defined by $\left(BA\right)x = B\left(Ax\right)$. With this product operation and pointwise addition $\mathcal{L}\left(\mathbb{H}\right)$ is an algebra whose identity element is the identity map $I\in\mathcal{L}\left(\mathbb{H}\right)$. For $A\in\mathcal{L}\left(\mathbb{H},\mathbb{H}'\right)$ we denote by A^* the unique member of $\mathcal{L}\left(\mathbb{H}',\mathbb{H}\right)$ satisfying $\left\langle Ax,y\right\rangle = \left\langle x,A^*y\right\rangle$ for all $x\in\mathbb{H}$ and $y\in\mathbb{H}'$. An operator $A\in\mathcal{L}\left(\mathbb{H}\right)$ is called self-adjoint if $A=A^*$ and nonnegative (or positive) if it is self-adjoint and $\left\langle Ax,x\right\rangle\geq 0$ (or $\left\langle Ax,x\right\rangle > 0$) for all $x\in\mathbb{H}$, $x\neq 0$, in which case we write $A\succeq 0$ (or $A\succ 0$). We use "$\preceq$" to denote the order induced by the cone of nonnegative operators. For $A\in\mathcal{L}\left(\mathbb{H},\mathbb{H}'\right)$ we denote the range by $\mathrm{Ran}\left(A\right)$ and the null space by $\mathrm{Ker}\left(A\right)$. The spectrum of $A\in\mathcal{L}\left(\mathbb{H}\right)$, denoted $\mathrm{Spec}(A)$, is the set of complex numbers λ such that $A+\lambda I$ is not invertible. $\mathrm{Spec}(A)$ is always compact and a subset of the real line if A is self-adjoint, in which case we write $\lambda_{\max}\left(A\right)$ for its supremum. If $A\succeq 0$ then $\mathrm{Spec}(A)$ consists only of nonnegative numbers and $\lambda_{\max}\left(A\right) = \left\|A\right\|_\infty$.

An $A\in\mathcal{L}\left(\mathbb{H},\mathbb{H}'\right)$ is called compact if the image of the open unit ball of $\mathbb{H}$ under A is pre-compact (totally bounded) in $\mathbb{H}'$. If $\mathrm{Ran}\left(\mathbb{H}\right)$ is finite dimensional then A is compact, finite linear combinations of compact linear maps and products with bounded linear maps are compact.

If $A \in \mathcal{L}(\mathbb{H})$ is compact and self-adjoint then there exists an orthonormal basis e_i of $\mathbb{H}$ and a sequence of real numbers λ_i satisfying $|\lambda_i| \to 0$ such that $A = \sum_i \lambda_i Q_{e_i}$, where Q_{e_i} is the operator defined by $Q_{e_i} x = \langle x, e_i \rangle e_i$. The e_i are eigenvectors and the λ_i eigenvalues of A. In this case $\mathrm{Spec}(A)$ is the closure of the set of eigenvalues $\{\lambda_i\}$. If f is a continuous real function defined on the spectrum a self-adjoint $f(A) \in \mathcal{L}(\mathbb{H})$ is defined by

$$f(A) = \sum_i f(\lambda_i) Q_{e_i}.$$

$f(A)$ has the same eigenvectors as A and eigenvalues $f(A)$. In this paper all members of $\mathcal{L}(\mathbb{H})$ are either compact or of the form $f(A)$ with A compact, so that there always exists a basis of eigenvectors. In fact with the exception of the covariance operator all operators in this paper have either finite dimensional range or are equal to the identity on the complement of a finite dimensional subspace.

A compact $A \in \mathcal{L}(\mathbb{H})$ is nonnegative (positive) if all its eigenvalues are nonnegative (positive). If A is positive and its spectrum bounded away from zero, then $\ln(A)$ exists and has the property $\ln(A) \preceq \ln(B)$ whenever B is positive and $A \preceq B$. This property of operator monotonicity has been used only for operators which are equal to the identity on the complement of a finite dimensional subspace, so it derives just from the corresponding property of matrices (Bhatia, 1997).

A linear subspace $M \subseteq \mathbb{H}$ is called invariant under $A \in \mathcal{L}(\mathbb{H})$ if $AM \subseteq M$. For a linear subspace $M \subseteq \mathbb{H}$ we use $M^\perp$ to denote the orthogonal complement $M^\perp = \{x \in \mathbb{H} : \langle x, y \rangle = 0, \forall y \in M\}$. For selfadjoint $A \in \mathcal{L}(\mathbb{H})$ we have $\mathrm{Ran}(A)^\perp = \mathrm{Ker}(A)$.

If $M \subseteq \mathbb{H}$ is a closed subspace and $A \in \mathcal{L}(\mathbb{H})$ then the trace of A relative to M is defined

$$\mathrm{tr}_M A = \sum_i \langle A e_i, e_i \rangle,$$

where $\{e_i\}$ is a orthonormal basis of M. If M is invariant under A then the choice of basis does not affect the value of tr_M. For $M = \mathbb{H}$ we just write tr without subscript.

If $A \in \mathcal{L}(\mathbb{H}, \mathbb{H}')$ then $A^* A \in \mathcal{L}(\mathbb{H})$ and $A^* A \succeq 0$. The trace-norm of A is defined as

$$\|A\|_1 = \mathrm{tr}\left((A^* A)^{1/2}\right).$$

If $\|A\|_1 < \infty$ then A is compact. If $A \in \mathcal{L}(\mathbb{H})$ and $A \succeq 0$ then $\|A\|_1$ is simply the sum of eigenvalues of A.

For $w \in \mathbb{H}$ we define an operator $Q_w \in \mathcal{L}(\mathbb{H})$ by $Q_w v = \langle v, w \rangle w$, for $v \in \mathbb{H}$. Then Q_w has one dimensional range, satisfies $Q_w \succeq 0$ and $\mathrm{tr}(Q_w) = \|Q_w\|_1 = \|Q_w\|_\infty = \|w\|^2$. The covariance operator $C = \mathbb{E}_X Q_X$ is the only truly infinite dimensional operator in this paper. Since the trace is linear and commutes with expectation we have $\|C\|_1 = \mathbb{E}_X \|X\|^2$, so C is compact whenever the distribution of X is bounded.

Finally, we recall Hoelder's inequality (Bhatia, 1997) for linear maps in the following form.

Theorem 12 *Let A and B be two linear maps $\mathbb{H} \to \mathbb{R}^T$. Then $\mathrm{tr}(A^* B) \leq \|A\|_1 \|B\|_\infty$.*