Transfer bounds for linear feature learning

Andreas Maurer

Adalbertstr. 55
D-80799 München
am@andreas-maurer.eu

Abstract. If regression tasks are sampled from a distribution, then the expected error for a future task can be estimated by the average empirical errors on the data of a finite sample of tasks, uniformly over a class of regularizing or pre-processing transformations. The bound is dimension free, justifies optimization of the pre-processing feature-map and explains the circumstances under which learning-to-learn is preferable to single task learning.[1]

1 Introduction

Suppose that an agency offers to train predictors from the data-sets supplied by its customers. The training data are previously sampled from whichever prediction problems the customers happen to be interested in. The agency creates the predictor and delivers it to the customer, who then tests it and rewards the agency with a fixed amount minus a quantity in proportion to the predictors loss observed in the test.

To compute the predictors from the data-sets the agency uses a fixed base algorithm composed with a linear feature map which may be updated from customer to customer.

For optimal rewards the agency should select a feature map which minimizes the expected loss incurred in its future use. As the true distributions are unknown, the simplest strategy is to minimize an empirical analogue of this loss on the data-sets already supplied by previous customers. In this way the feature-map should improve with experience, and the agency is *learning to learn* in a very literal sense, as it uses past experience to improve learning performance on future, yet unseen data. A justification of this method would be a high probability bound on the expected future loss in terms of the empirical loss, the bound being uniform over the space of feature-maps parametrizing the algorithms.

Such bounds are the subject of this paper. They must not be confused with similar results for *multi-task learning*, which bound the average expected loss for a *fixed* set of distributions in terms of the corresponding empirical average. Here we bound the expected loss for an unknown future distribution in terms of the average empirical loss observed from the realizations of past distributions.

[1] AMS subject classification: primary 68T05, secondary 62J12.
Keywords: Learning to learn, transfer learning, kernel methods, generalization.

Such results require that all distributions in question are drawn from a common distribution of distributions, a construction which will be explained in the sequel.

For a more formal perspective on the problem assume that all the inputs x lie in (or are mapped to) a Hilbert space H, all the outputs y are members of the interval $[0,1]$, and that all predictors are given by bounded linear forms on H, so they are of the type $x \in H \mapsto \langle w, x\rangle$ for some weight vector w. The restriction to linear predictors is partly compensated by allowing H to be infinite dimensional, so that kernel methods become applicable. The loss incurred by a predictor w on an input-output pair $(x,y) \in H \times [0,1]$ is assumed to be the square loss $(\langle w, x\rangle - y)^2$.

A customer's problem is described by a probability measure μ on the set of input-output pairs $H \times [0,1]$, and we will assume that all inputs lie in the unit ball of H almost surely. The corresponding training data $(\mathbf{x}, \mathbf{y})$ are sampled in m independent trials from this distribution, that is $(\mathbf{x}, \mathbf{y}) \sim \mu^m$. For simplicity we assume m to be the same for all customers and fixed throughout this paper. We will imagine m to be a rather small number, which will make the potential advantages of the proposed method more pronounced. The training data $(\mathbf{x}, \mathbf{y})$ are then brought to the machine learning shop.

Assume that the agency's base algorithm is regularized least squares regression with a fixed regularization parameter $\lambda > 0$ (see sections 2.2 and 5), and that the set of potential feature-maps is the set $\mathcal{P}_d$ of orthogonal projections P with d-dimensional range in H. The choice of this set of feature-maps expresses the belief that some initially unknown d-dimensional subspace of H contains the prediction-relevant features for all customers. This belief can originate in the fact that all the agency's customers come from the same region or share some interests, such as weather prediction or image processing. If P is the chosen pre-processing projection, the agency computes and delivers the corresponding weight vector $\omega_{\lambda^{-1}P}(\mathbf{x}, \mathbf{y})$ (the odd-looking notation will become clear later).

The customer then draws a test pair (x,y) from the distribution μ and computes the loss $(\langle \omega_{\lambda^{-1}P}(\mathbf{x}, \mathbf{y}), x\rangle - y)^2$ which is subtracted from the agencies reward.

To proceed we make the key assumptions that the encounter with a new customer is itself a random event, governed by a probability distribution ρ, and that the customers, as they appear in the agencies business history, correspond to independent realizations of such events. Since the only statistically relevant property of a customer is the distribution μ, the encounter with a customer corresponds to a draw $\mu \sim \rho$, so that ρ is a distribution on the set of input-output distributions, a construction which has been termed an *environment* by J. Baxter [6].

To compute the expected loss incurred from future use of the feature map $P \in \mathcal{P}_d$ we have to take the expectation of $(\langle \omega_{\lambda^{-1}P}(\mathbf{x}, \mathbf{y}), x\rangle - y)^2$ as

- a new customer appears ($\mathbb{E}_{\mu\sim\rho}$),

– and prepares a training-set ($\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}$),
– and a test pair ($\mathbb{E}_{(x,y)\sim\mu}$),

so that an optimal feature map P would minimize the *transfer risk*

$$R_\rho\left(\omega_{\lambda^{-1}P}\right) = \mathbb{E}_{\mu\sim\rho}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\mathbb{E}_{(x,y)\sim\mu}\left[\left(\left\langle\omega_{\lambda^{-1}P}\left(\mathbf{x},\mathbf{y}\right),x\right\rangle - y\right)^2\right].$$

This quantity depends on the unknown distribution ρ, so an empirically accessible estimator has to be used.

Suppose that there were n customers in the agency's past business history. This means that the data-sets $\left(\mathbf{x}^1,\mathbf{y}^1\right),...,\left(\mathbf{x}^n,\mathbf{y}^n\right)$ were observed through n independent draws of input-output distributions μ_k from ρ and subsequent iid draws of training sets $\left(\mathbf{x}^k,\mathbf{y}^k\right)$ from the μ_k. On each data-set $\left(\mathbf{x}^k,\mathbf{y}^k\right)$ the use of the projection P incurs the empirical loss

$$\hat{\ell}_{\omega_{\lambda^{-1}P}}\left(\mathbf{x}^k,\mathbf{y}^k\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\left\langle\omega_{\lambda^{-1}P}\left(\mathbf{x}^k,\mathbf{y}^k\right),x_i^k\right\rangle - y_i^k\right)^2,$$

where $\left(x_i^k,y_i^k\right)$ is the i-th pair in $\left(\mathbf{x}^k,\mathbf{y}^k\right)$. A conceptually simple algorithm to select P then just minimizes the total empirical loss incurred by P over all past data-sets:

$$P\left(\left(\mathbf{x}^1,\mathbf{y}^1\right),...,\left(\mathbf{x}^n,\mathbf{y}^n\right)\right) = \arg\min_{P\in\mathcal{P}_d}\frac{1}{n}\sum_{k=1}^{n}\hat{\ell}_{\omega_{\lambda^{-1}P}}\left(\mathbf{x}^k,\mathbf{y}^k\right).$$

We can give a generalization guarantee for this algorithm in terms of a uniform bound valid for all $P\in\mathcal{P}_d$.

Theorem 1. *For all $\delta>0$, we have with probability greater $1-\delta$ in the data $\left(\mathbf{x}^1,\mathbf{y}^1\right),...,\left(\mathbf{x}^n,\mathbf{y}^n\right)$ that for all feature maps $P\in\mathcal{P}_d$*

$$R\left(\omega_{\lambda^{-1}P}\right) \leq \frac{1}{n}\sum_{k=1}^{n}\hat{\ell}_{\omega_{\lambda^{-1}P}}\left(\mathbf{x}^k,\mathbf{y}^k\right) + \frac{\sqrt{8\pi d}}{\lambda}\left(2\sqrt{\frac{\left\|C\right\|_\infty}{m}}+\sqrt{\frac{1}{n}}\right)+\sqrt{\frac{\ln\left(1/\delta\right)}{2n}}.$$

The quantity $\left\|C\right\|_\infty$ appearing above is the largest eigenvalue of the covariance operator C for the total input distribution. This distribution is obtained by averaging all input marginals over ρ, and the operator C describes its moments of inertia. The largest eigenvalue $\left\|C\right\|_\infty$ can be interpreted geometrically as the square of the length of the largest principal axis of the corresponding ellipsoid.

To make a case for learning-to-learn we have to show that there are reasonable conditions under which the agency using this algorithm outperforms a competitor who just uses conventional regularized least squares regression for all customers, with the same regularization parameter and without pre-processing projections. Without the projection the competitor will always achieve a lower empirical error and conventional upper bounds on the competitors estimation

error have the order of $\lambda^{-1}/\sqrt{m}$ (see e.g. Bousquet and Elisseeff [8]). For a fair comparison we assume that the competitor achieves an empirical error of zero and a fast decay of the competitors estimation error as $1/(\lambda m)$. These assumptions appear to be rather optimal for the competitor and a maximal handicap for learning to learn. So, if we ignore the dependence on the confidence parameter δ on both sides, for learning to learn to still be preferable we should have

$$\frac{1}{n}\sum_{k=1}^{n}\hat{\ell}_{\omega_{\lambda^{-1}P}}\left(\mathbf{x}^k,\mathbf{y}^k\right)+\frac{2}{\lambda}\sqrt{\frac{8\pi d\left\|C\right\|_\infty}{m}}+\frac{1}{\lambda}\sqrt{\frac{8\pi d}{n}}<\frac{1}{\lambda m}.$$

This happens if each of the three terms on the l.h.s. is substantially smaller than the r.h.s., that is

1. The tasks which were observed in the past must be *empirically related*, that is $(1/n)\sum_{k=1}^{n}\hat{\ell}_{\omega_{\lambda^{-1}P}}\left(\mathbf{x}^k,\mathbf{y}^k\right)\approx 0$ for some choice of a d-dimensional subspace with projection P. This corresponds to the essential "prior belief" mentioned above. Note however, that with standard model selection techniques (e.g. Lemma 15.5 in [2]) our bound can be modified to permit the choice of d *after* seeing the data $\left(\mathbf{x}^1,\mathbf{y}^1\right),...,\left(\mathbf{x}^n,\mathbf{y}^n\right)$ at only a logarithmic penalty in d.
2. The input distribution must be high-dimensional so that $d\left\|C\right\|_\infty \ll m^{-1}$. To make this explicit suppose that the input distribution is concentrated and uniform on a k-dimensional unit-sphere in H. Then, as the eigenvalues of C must add up to one and are equal by symmetry, we have $\left\|C\right\|_\infty = k^{-1}$, so that $d\left\|C\right\|_\infty$ decreases with the dimensionality k of the distribution. The quantity $d\left\|C\right\|_\infty = d/k$ can be interpreted as the ratio of utilizable to totally present information.
3. The number n of past customers must be large in relation to d and m, that is $dm^2 \ll n$. This is always satisfied for a large enough history. The practically interesting regime is for very small m, say from $m=3$ to $m=20$.

If these conditions are met, learning-to-learn is the method of choice. They coincide with the conditions under which multi-task learning is preferable to single task learning. The differences and similarities between multi-task learning and learning-to-learn are discussed in more detail in section 6.1.

An analogue of Theorem 1 holds under much more general circumstances with a class of base algorithms and loss functions satisfying a certain Lipschitz condition to be expected for many kernel algorithms, and for essentially arbitrary classes of compact linear feature maps, with possibly infinite-dimensional range, where the complexity penalty $\sqrt{d}$ for projections is replaced by Schatten norms. This is the content of Theorem 2, which is the principal result of this paper.

The next section introduces the necessary material to state Theorem 2. Section 3 contains a summary of notation and auxiliary results needed for the proof of Theorem 2 which is given in section 4. In section 5 it is shown that regularized least squares regression satisfies the hypotheses of Theorem 2 and Theorem 1 is derived. Finally we compare Theorem 2 to some related results in the literature. An appendix summarizes the most frequently used notation in tabular form.

2 Overview

We study regression where input data x lies in a Hilbert space H with inner product $\langle .,. \rangle$ and norm $\|.\|$ and unit ball $B_1(H)$. Output data y lies in the interval $[0,1]$. Weight vectors are denoted w and make the prediction $\langle w, x \rangle$ for an input point $x \in H$. The loss of a weight vector $w \in H$ on a pair $(x,y) \in H \times [0,1]$ is given by $\ell(\langle w, x \rangle, y)$ where the loss function $\ell : \mathbb{R} \times [0,1] \to \mathbb{R}_+$ measures the discrepancy between the prediction in its first, and the true output in its second argument. An important example is given by the square loss function $\ell_{\text{sqr}}(y', y) = (y' - y)^2$.

2.1 Tasks and samples

A task is a probability measure $\mu \in M_1(B_1(H) \times [0,1])$, where $M_1(\mathcal{X})$ generally denotes the set of probability measures on a space $\mathcal{X}$. Questions of measurability are ignored throughout, and the reader who feels uneasy about this is asked to consider only measures with a fixed finite support of very large cardinality.

Relative to a task μ the expected loss of a weight vector is $\mathbb{E}_{(x,y)\sim\mu} \ell(\langle w, x \rangle, y)$ and our main goal is to find weight vectors w to make this quantity small.

Information on the task μ is obtained by independently drawing a finite number m of training examples $(x_i, y_i) \sim \mu$. Such an m-tuple $((x_1, y_1), ..., (x_m, y_m)) \sim \mu^m$ is called a *sample*, which we also denote $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), ..., (x_m, y_m))$ with the understanding that $\mathbf{x} = (x_1, ..., x_m)$ and $\mathbf{y} = (y_1, ..., y_m)$.

If $\mathbf{x} = (x_1, ..., x_m) \in H^m$ the Gramian (or kernel-matrix) of $\mathbf{x}$ is the $m \times m$-matrix $\mathbf{G}(\mathbf{x})$ given by

$$\mathbf{G}(\mathbf{x})_{ij} = \langle x_i, x_j \rangle .$$

The Frobenius norm of any $m \times m$-matrix $\mathbf{A}$ is denoted $\|\mathbf{A}\|_{Fr}$. If T is a linear operator on H we write $(T\mathbf{x}, \mathbf{y}) = ((Tx_1, y_1), ..., (Tx_m, y_m))$. The sample size m will be fixed throughout.

2.2 Algorithms

A learning algorithm w is a function $w : (H \times [0,1])^m \to H$. From the training sample $(\mathbf{x}, \mathbf{y}) \in (H \times [0,1])^m$ it computes a weight vector $w(\mathbf{x}, \mathbf{y}) \in H$, which for an input-output pair (x, y) predicts $\langle w(\mathbf{x}, \mathbf{y}), x \rangle$ and incurs the loss $\ell(\langle w(\mathbf{x}, \mathbf{y}), x \rangle, y)$.

A real valued function associated with any algorithm w is its empirical loss $\hat{\ell}_w : (H \times [0,1])^m \to \mathbb{R}^+$, defined by

$$\hat{\ell}_w(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\langle w(\mathbf{x}, \mathbf{y}), x_i \rangle, y_i),$$

which is just the average loss of the weight returned by the algorithm on its own training data.

Definition 1. *Relative to a fixed loss function ℓ, an algorithm $w : (H \times [0,1])^m \to H$ is said to*

(i) be 1-bounded if $\|w(\mathbf{x},\mathbf{y})\| \leq 1$ and $\hat{\ell}_w(\mathbf{x},\mathbf{y}) \leq 1$.

(ii) have kernel stability L if

$$\hat{\ell}_w(\mathbf{x}_1,\mathbf{y}) - \hat{\ell}_w(\mathbf{x}_2,\mathbf{y}) \leq \frac{L}{m}\|\mathbf{G}(\mathbf{x}_1) - \mathbf{G}(\mathbf{x}_2)\|_{Fr}.$$

for any $\mathbf{x}_1, \mathbf{x}_2 \in H^m$ and $\mathbf{y} \in [0,1]^m$.

An example of a learning algorithm is regularized least squares regression ω with regularization parameter 1

$$\omega(\mathbf{x},\mathbf{y}) = \arg\min_{w\in H}\left(\frac{1}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2 + \|w\|^2\right).$$

It will be shown in section 5 that the algorithm ω is 1-bounded with kernel stability 2 relative to the square loss ℓ_{sqr}.

Kernel stability of regularized least squares regression is not an exception. It is easy to see that the empirical loss of any kernelizable algorithm w on the data $(\mathbf{x},\mathbf{y})$ is a function only of $\mathbf{G}(\mathbf{x})$ and $\mathbf{y}$. If this function has Lipschitz constant L for all $\mathbf{y}$, then w has kernel stability mL.

2.3 Linear feature maps and modified algorithms

Suppose w is an algorithm and $D \in \mathcal{L}^+(H)$, which means that D is a symmetric, positive semidefinite bounded linear operator. We can construct a new algorithm w_D with the formula

$$w_D(\mathbf{x},\mathbf{y}) = D^{1/2} w\left(D^{1/2}\mathbf{x},\mathbf{y}\right).$$

For an input $x \in H$ the weight $w_D(\mathbf{x},\mathbf{y})$ makes the prediction $\langle w_D(\mathbf{x},\mathbf{y}), x\rangle = \left\langle w\left(D^{1/2}\mathbf{x},\mathbf{y}\right), D^{1/2}x\right\rangle$, so that $D^{1/2}$ can be interpreted as a linear feature map in a layered model: The behavior of w_D is the same as that of w with all inputs, both for training and testing, being pre-processed through the transformation $D^{1/2}$.

The empirical loss of the modified algorithm w_D is

$$\hat{\ell}_{w_D}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m}\ell\left(\left\langle w\left(D^{1/2}\mathbf{x},\mathbf{y}\right), D^{1/2}x_i\right\rangle, y_i\right) = \hat{\ell}_w\left(D^{1/2}\mathbf{x},\mathbf{y}\right).$$

For a simple example take regularized least squares regression $w = \omega$ and $D = \lambda^{-1}I$, where $\lambda > 0$ and I is the identity on H. Then by a change of variables

$$\begin{aligned}\omega_{\lambda^{-1}I}(\mathbf{x},\mathbf{y}) &= \lambda^{-1/2}\arg\min_{w\in H}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\left\langle w, \lambda^{-1/2}x_i\right\rangle - y_i\right)^2 + \|w\|^2\right)\\ &= \arg\min_{w\in H}\left(\frac{1}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2 + \lambda\|w\|^2\right),\end{aligned}$$

so that $\omega_{\lambda^{-1}I}$ is just the familiar regularized least squares algorithm with regularization parameter λ. More generally, if $D \in \mathcal{L}^+(H)$ is nonsingular, a similar change of variables gives

$$\omega_D(\mathbf{x}, \mathbf{y}) = \arg\min_{w \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \left(\langle w, x_i \rangle - y_i \right)^2 + \left\langle D^{-1} w, w \right\rangle \right).$$

This can be extended to general D by using pseudo-inverses and defining the regularizer to be infinite if w is not in the range of D (as in [3]), so for regularized algorithms the inverse square of the feature map becomes a modified regularizer.

Given a fixed base-algorithm w and a family of positive operators $\mathcal{D} \subseteq \mathcal{L}^+(H)$ we obtain a family of algorithms $\{w_D : D \in \mathcal{D}\}$. The principal goal is the sensible choice of an algorithm from this family on the basis of experience obtained from an environment of tasks.

2.4 Environments

The encounter with a task μ is itself a random event, corresponding to a draw $\mu \sim \rho$ where ρ is a probability measure on the set of tasks, that is $\rho \in M_1(M_1(B_1(H) \times [0,1]))$. Baxter [6] calls such probability measures *environments*.

To test the performance of a given learning algorithm w in such an environment we should

- make a random choice of a task $\mu \sim \rho$
- draw a training sample $(\mathbf{x}, \mathbf{y}) \sim \mu^m$
- draw a test pair $(x, y) \sim \mu$
- run the algorithm to obtain $w(\mathbf{x}, \mathbf{y})$
- return the loss $\ell(\langle w(\mathbf{x}, \mathbf{y}), x \rangle, y)$.

The expected output of this procedure should be a good (negative) measure of the performance of the algorithm in the given environment. This motivates the definition of the transfer risk associated with the algorithm w

$$R_\rho(w) = \mathbb{E}_{\mu \sim \rho} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \mu^m} \mathbb{E}_{(x,y) \sim \mu} \ell(\langle w(\mathbf{x}, \mathbf{y}), x \rangle, y). \tag{1}$$

Information on the environment ρ is obtained by independently drawing a finite number n of tasks $\mu_l \sim \rho$ and representing each task μ_l by a sample $(\mathbf{x}^l, \mathbf{y}^l) \sim (\mu_l)^m$, $(\mathbf{x}^l, \mathbf{y}^l) = \left((x_1^l, y_1^l), ..., (x_m^l, y_m^l) \right)$, with the understanding that $\mathbf{x}^l = (x_1^l, ..., x_m^l)$ and $\mathbf{y}^l = (y_1^l, ..., y_m^l)$. Write $(\mathbf{X}, \mathbf{Y}) = \left((\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^n, \mathbf{y}^n) \right)$ for the training data generated in this manner.

Define a probability measure $\hat{\rho}$ on the set of samples $(B_1(H) \times [0,1])^m$ by $\mathbb{E}_{\hat{\rho}}(f) = \mathbb{E}_{\mu \sim \rho} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \mu^m} f(\mathbf{x}, \mathbf{y})$ for every Borel function f on $(B_1(H) \times [0,1])^m$. The measure $\hat{\rho}$ models the draw of a training sample by

- making a random choice of a task $\mu \sim \rho$
- drawing the sample $(\mathbf{x}, \mathbf{y}) \sim \mu^m$.

The entire training data $(\mathbf{X}, \mathbf{Y})$ above is therefore generated in n independent draws from $\hat{\rho}$, that is $(\mathbf{X}, \mathbf{Y}) = \left((\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^n, \mathbf{y}^n) \right) \sim \hat{\rho}^n$.

2.5 A bound for learning to learn

Suppose the given data $(\mathbf{X}, \mathbf{Y}) \sim \hat{\rho}^n$ constitutes all our experience with the environment ρ, and that we work with a fixed base algorithm w and a fixed class $\mathcal{D} \subseteq \mathcal{L}^+(H)$ of positive semidefinite operators. We want to use $(\mathbf{X}, \mathbf{Y})$ to select some $D(\mathbf{X}, \mathbf{Y}) \in \mathcal{D}$ such that the expected future performance of the modified algorithm $w_{D(\mathbf{X},\mathbf{Y})}$ is optimal, which is to say that $R\left(w_{D(\mathbf{X},\mathbf{Y})}\right)$ should be minimal or near minimal. The conceptually simplest way to do this is empirical risk minimization,

$$D(\mathbf{X}, \mathbf{Y}) = \arg\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{l=1}^{n} \hat{\ell}_{w_D}\left(\mathbf{x}^l, \mathbf{y}^l\right), \tag{2}$$

which just selects the algorithm with the best average performance on the available data. To justify this procedure we need a high probability bound of $R\left(w_{D(\mathbf{X},\mathbf{Y})}\right)$ in terms of $(1/n)\sum_l \hat{\ell}_{w_{D(\mathbf{X},\mathbf{Y})}}\left(\mathbf{x}^l, \mathbf{y}^l\right)$ and the simplest form for such a bound is uniform over $\mathcal{D}$, that is it holds for all $D \in \mathcal{D}$, not just for $D(\mathbf{X}, \mathbf{Y})$.

Theorem 2. *Take conjugate exponents p and q with $1 = 1/p + 1/q$. Suppose the algorithm w is 1-bounded and has kernel stability L relative to the loss function ℓ and that for every $K > 0$ there exists $M(K)$ such that for all $y \in [0,1]$ and for all $s, t \in [-K, K]$ we have*

$$\ell(s, y) - \ell(t, y) \leq M(K)\,|s - t|.$$

Then for any environment $\rho \in M_1(M_1(B_1(H)) \times [0,1]))$ and for any $\delta > 0$, with probability greater $1 - \delta$ in the data $(\mathbf{X}, \mathbf{Y}) \sim \hat{\rho}^n$ we have for all $D \in \mathcal{D}$

$$R_\rho(w_D) \leq \frac{1}{n} \sum_{l=1}^{n} \hat{\ell}_{w_D}\left(\mathbf{x}^l, \mathbf{y}^l\right) + \\ + \sqrt{\frac{2\pi}{m}} M\left(\|D\|_\infty^{1/2}\right) \|D\|_q^{1/2} \|C\|_p^{1/2} + \sqrt{\frac{2\pi}{n}} L \|\mathcal{D}\|_2 + \sqrt{\frac{\ln(1/\delta)}{2n}},$$

where $\|\mathcal{D}\|_p = \sup_{D \in \mathcal{D}} \|D\|_p$ and $\|D\|_p = tr\left(|D|^p\right)^{1/p}$ ($\|D\|_\infty$ being the absolute value of the largest singular value of D) and $C = \mathbb{E}_{\mu\sim\rho}\mathbb{E}_{(x,y)\sim\mu} Q_x$ is the covariance operator (see section 3.1).

The three terms in the second line above bound the estimation error and correspond respectively to a bound on the within-task estimation error, a bound on the inter-task estimation error and a term expressing the dependence on the confidence parameter δ. For $p \geq 2$, which is the interesting regime, the norms of the covariance operator, although in principle unknown, can be well estimated from the normalized Gramian, thus converting the theorem to a data dependent bound (see Theorem 3 below).

2.6 Proof strategy

The idea of the proof is to write

$$R_\rho(w_D) - \frac{1}{n}\sum_{l=1}^{n} \hat{\ell}_{w_D}\left(\mathbf{x}^l, \mathbf{y}^l\right)$$
$$= \left(R_\rho(w_D) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\hat{\ell}_{w_D}(\mathbf{x},\mathbf{y})\right) + \left(\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\hat{\ell}_{w_D}(\mathbf{x},\mathbf{y}) - \frac{1}{n}\sum_{l=1}^{n} \hat{\ell}_{w_D}\left(\mathbf{x}^l, \mathbf{y}^l\right)\right)$$

and to bound the two terms separately.

The first term can be rewritten as

$$\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\left(\mathbb{E}_{(x,y)\sim\mu}\ell\left(\langle w_D(\mathbf{x},\mathbf{y}), x\rangle, y\right) - \hat{\ell}_{w_D}(\mathbf{x},\mathbf{y})\right),$$

which is the estimation difference expected for the future task. This can be bounded using the 1-boundedness of w and the Lipschitz properties of the loss function ℓ. Such is done in Section 4.1 (see Theorem 6) and gives rise to the m-dependent term in the bound of Theorem 2.

The other term in the decomposition is the estimation difference between the expected empirical error of the algorithms output on a new task and the corresponding average on the known empirical errors of the output on the tasks of the past. Bounding it involves complexity estimates of the function class

$$\mathcal{F} = \left\{(\mathbf{x},\mathbf{y}) \mapsto \hat{\ell}_D(\mathbf{x},\mathbf{y}) : D \in \mathcal{D}\right\},$$

which is accomplished using Slepian's lemma and the boundedness and Lipschitz (kernel stability) properties of w. See section 4.2.

3 Notation, definitions and auxiliary results

There is also an appendix containing a tabular summary of the notation used in the paper.

3.1 Operators and matrices

For any Hilbert space H we use $\mathcal{L}(H)$ to denote the set of bounded operators on H, and $\mathcal{L}^+(H)$ for the positive (semidefinite) members of $\mathcal{L}(H)$. $\mathcal{L}_2(H)$ is the set of Hilbert-Schmidt operators, which becomes itself a Hilbert space with the inner product $\langle T, S\rangle_2 = tr(T^*S)$ and the corresponding (Frobenius-) norm $\|.\|_2$. The set of positive (semidefinite) members of $\mathcal{L}_2(H)$ is denoted $\mathcal{L}_2^+(H)$. For operators $T \in \mathcal{L}(H)$ the Schatten norms are denoted $\|T\|_p = tr\left(|T|^p\right)^{1/p}$, where $|T| = \left(T^*T\right)^{1/2}$, and Hölder's inequality asserts that

$$|\langle T, S\rangle_2| \leq \|T\|_p \|S\|_q, \tag{3}$$

if $1/p+1/q=1$, $p,q\in[1,\infty]$ and both norms on the right are finite (Reed Simon [19] and [18]).

For $x,y\in H$ the operators Q_x and $J_{x,y}$ are defined respectively by $Q_x z=\langle z,x\rangle\, x$ and $J_{xy}z=\langle z,x\rangle\, y$. For properties of Q_x and $J_{x,y}$ see [16]. Here we will only use the following facts, which are easily verified:

Lemma 1. *Let $x,y,x',y'\in H$ and $T\in\mathcal{L}_2(H)$. Then*
(i) $\langle Q_x,Q_y\rangle_2=\langle x,y\rangle^2$.
(ii) $\langle T^*T,Q_x\rangle_2=\|Tx\|^2$.
(iii) $\langle J_{x,y},J_{x',y'}\rangle_2=\langle x,x'\rangle\langle y,y'\rangle$
(iv) $\langle T,J_{x,y}\rangle_2=\langle Tx,y\rangle$.

In the finite dimensional case and matrix notation we could write $Q_x=xx^T$.

If μ is a probability measure on H, the operator valued expectation $\mathbb{E}_{x\sim\mu}[Q_x]$ is called the *covariance operator* of μ. If μ has support in $B_1(H)$ we have $\|\mathbb{E}[Q_x]\|_1\leq 1$. From a theorem in [20] on the concentration of means for vector-valued variables we obtain, upon transferring to $\mathcal{L}_2(H)$

Theorem 3. *Suppose that μ is a probability measure on $B_1(H)$. Then for all $\delta>0$ with probability greater than $1-\delta$ in $\mathbf{x}=(x_1,...,x_m)\sim\mu^m$ we have*

$$\left\|\mathbb{E}_{x\sim\mu}[Q_x]-\frac{1}{m}\sum_{i=1}^{m}Q_{x_i}\right\|_2\leq\frac{2}{\sqrt{m}}\left(1+\sqrt{\frac{\ln(1/\delta)}{2}}\right).$$

The quantity

$$\left\|\sum_i Q_{x_i}\right\|_2=\left(\sum_{i,j}\langle x_i,x_j\rangle^2\right)^{1/2}$$

is the Frobenius norm of the Gramian (or kernel-) matrix $\mathbf{G}(\mathbf{x})_{ij}=\langle x_i,x_j\rangle$, denoted $\|\mathbf{G}(\mathbf{x})\|_{Fr}$. An immediate corollary to the above is, that $(1/m)\|\mathbf{G}(\mathbf{x})\|_{Fr}$ is with high probability a good approximation of $\|\mathbb{E}[Q_x]\|_2$. It follows from the triangle inequality and the decreasing property of the Schatten norms $\|.\|_p$ that for $p\geq 2$ the norm $\|\mathbb{E}[Q_x]\|_p$ may be equally well estimated by $(1/m)\|\sum_i Q_{x_i}\|_p$ on a finite sample. See also [21] and [24].

With $\langle .,.\rangle_{\mathbb{R}^m}$ and $\|.\|_{\mathbb{R}^m}$ we denote the canonical inner product and norm in $\mathbb{R}^m$. We sometimes write $\mathbb{N}_k=\{1,...,k\}$.

3.2 Rademacher and Gaussian averages

We will use $(\sigma_i:i\in\mathbb{N})$ to denote a sequence of independent random variables, uniformly distributed on $\{-1,1\}$ and $(\gamma_i:i\in\mathbb{N})$ for a sequence of independent, $N(0,1)$-distributed (standardized normal) variables, independent also of the σ's.

For $A \subseteq \mathbb{R}^k$ we define the Rademacher and Gaussian averages of A ([14],[4]) as

$$\mathcal{R}(A) = \mathbb{E}_\sigma \sup_{(x_1,...,x_k)\in A} \frac{2}{k}\sum_{i=1}^{k} \sigma_i x_i,$$

$$\Gamma(A) = \mathbb{E}_\gamma \sup_{(x_1,...,x_k)\in A} \frac{2}{k}\sum_{i=1}^{k} \gamma_i x_i.$$

We will use the following inequality (see [14], 4.2, p 97):

$$\mathcal{R}(A) \leq \sqrt{\pi/2}\ \Gamma(A). \tag{4}$$

If $\mathcal{F}$ is a class of real functions on a space $\mathcal{X}$ and $\mathbf{x} = (x_1, ..., x_k) \in \mathcal{X}^k$ we write

$$\mathcal{F}(\mathbf{x}) = \mathcal{F}(x_1, ..., x_k) = \{(f(x_1), ..., f(x_k)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^k.$$

The empirical Rademacher and Gaussian complexities of $\mathcal{F}$ on $\mathbf{x}$ are respectively $\mathcal{R}(\mathcal{F}(\mathbf{x}))$ and $\Gamma(\mathcal{F}(\mathbf{x}))$. If $\mu \in M_1(\mathcal{X})$ is a probability measure on $\mathcal{X}$ then the corresponding expected complexities are $\mathbb{E}_{\mathbf{x}\sim\mu^m}\mathcal{R}(\mathcal{F}(\mathbf{x}))$ and $\mathbb{E}_{\mathbf{x}\sim\mu^m}\Gamma(\mathcal{F}(\mathbf{x}))$.

The following Theorem is fundamental for our results. For the readers benefit we sketch a proof (for detailed proofs see van der Vaart and Wellner [23] for (i) and Koltchinskii and Panchenko [13], Bartlett and Mendelson [4] for (ii)).

Theorem 4. *Let $\mathcal{F}$ be a real-valued function class on a space $\mathcal{X}$ and $\mu \in M_1(\mathcal{X})$. For $\mathbf{x} = (x_1, ..., x_m) \in \mathcal{X}^m$ define*

$$\Phi(\mathbf{x}) = \sup_{f\in\mathcal{F}} \left(\mathbb{E}_{x\sim\mu}[f(x)] - \frac{1}{m}\sum_{i=1}^{m} f(x_i) \right).$$

(i) $\mathbb{E}_{\mathbf{x}\sim\mu^m}[\Phi(\mathbf{x})] \leq \mathbb{E}_{\mathbf{x}\sim\mu^m}\mathcal{R}(\mathcal{F}(\mathbf{x}))$.
(ii) If $\mathcal{F}$ is $[0,1]$-valued then $\forall\delta > 0$ we have with probability greater than $1-\delta$ in $\mathbf{x} \sim \mu^m$ that

$$\Phi(\mathbf{x}) \leq \mathbb{E}_{\mathbf{x}\sim\mu^m}\mathcal{R}(\mathcal{F}(\mathbf{x})) + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

(iii) $\mathcal{R}(\mathcal{F}(\mathbf{x}))$ may be replaced by $\sqrt{\pi/2}\ \Gamma(\mathcal{F}(\mathbf{x}))$ in (i) and (ii).

Proof. For any realization $\sigma = \sigma_1, ..., \sigma_m$ of the Rademacher variables

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mu^m}[\Phi(\mathbf{x})] &= \mathbb{E}_{\mathbf{x}\sim\mu^m} \sup_{f\in\mathcal{F}} \frac{1}{m}\mathbb{E}_{\mathbf{x}'\sim\mu^m} \sum_{i=1}^{m} (f(x_i') - f(x_i)) \\ &\leq \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mu^m\times\mu^m} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i (f(x_i') - f(x_i)),\end{aligned}$$

because of the symmetry of the measure $\mu^m \times \mu^m$ under the interchange $x_i \leftrightarrow x_i'$. Taking the expectation in σ and applying the triangle inequality gives (i). To prove (ii) apply the Hoeffding-Azuma-McDiarmid concentration inequality [17] to Φ and solve for the deviation. (iii) follows from (4) above.

We will only use Gaussian averages in the sequel. In doing so we loose little ($\sqrt{\pi/2} \approx 1.25$), but we gain the comparison properties of Gaussian averages, which can be derived from the following Theorem, known as Slepian's Lemma (Ledoux and Talagrand 1991 [14]):

Theorem 5. *Let Ω and Ξ be mean zero, separable Gaussian processes indexed by a common set A, such that*

$$\mathbb{E}\left(\Omega_a - \Omega_b\right)^2 \leq \mathbb{E}\left(\Xi_a - \Xi_b\right)^2 \text{ for all } a, b \in A.$$

Then $\mathbb{E}\sup_{a\in A}\Omega_a \leq \mathbb{E}\sup_{a\in A}\Xi_a$.

Corollary 1. *Let $A \subseteq \mathbb{R}^k$ and $\phi_1, ..., \phi_k$ be real functions, each with Lipschitz constant L. Denote $\boldsymbol{\phi}\circ A = \{(\phi_1(x_1), ..., \phi_k(x_k)) : (x_1, ..., x_k) \in A\}$. Then $\Gamma(\boldsymbol{\phi} \circ A) \leq L\ \Gamma(A)$.*

Proof. Define Gaussian processes Ω and Ξ, indexed by A as $\Omega_{\mathbf{x}} = \sum_{i=1}^k \gamma_i \phi_i(x_i)$ and $\Xi_{\mathbf{x}} = L\ \sum_{i=1}^k \gamma_i x_i$. Then by the orthonormality of the γ_i

$$\mathbb{E}\left(\Omega_{\mathbf{x}} - \Omega_{\mathbf{x}'}\right)^2 = \sum_{i=1}^k \left(\phi_i(x_i) - \phi_i(x_i')\right)^2 \leq L^2 \sum_{i=1}^k \left(x_i - x_i'\right)^2 = \mathbb{E}\left(\Xi_{\mathbf{x}} - \Xi_{\mathbf{x}'}\right)^2,$$

whence it follows from Slepians Lemma that

$$\Gamma(\boldsymbol{\phi} \circ A) = \frac{2}{k}\mathbb{E}\sup_{\mathbf{x}\in A}\Omega_{\mathbf{x}} \leq \frac{2}{k}\mathbb{E}\sup_{\mathbf{x}\in A}\Xi_{\mathbf{x}} = L\ \Gamma(A).$$

In this case there is an analogous version for Rademacher averages (Ledoux and Talagrand [14], Bartlett, Bousquet et al [5]), but in the proof of Theorem 7 below we will encounter a situation where the comparison Theorems for Rademacher averages are insufficient and Theorem 5 is needed.

4 Proof of Theorem 2

We proceed as outlined in section 2.6 by separately bounding the expected estimation difference for the future task and the estimation error between the expected empirical loss in the future and the empirical loss observed in the past.

4.1 Bounding the estimation error for the future task

Here we bound the expected estimation error when a fixed symmetric operator $D^{1/2}$ is used as a preprocessor. This will give the m-dependent term in the bound of Theorem 1. We first give a bound on the Gaussian complexity of uniformly bounded linear functionals acting on pre-processed inputs.

Lemma 2. *Let p and q be conjugate exponents, let $D \in \mathcal{L}^+(H)$ and let $\mathcal{G}_D$ be the class of real functions on $B_1(H)\times[0,1]$ defined by $\mathcal{G}_D = \left\{(x,y) \mapsto \left\langle w, D^{1/2}x\right\rangle : \|w\| \leq 1\right\}$.*

(i) For any $\mu \in M_1(B_1(H)\times[0,1])$

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m} \Gamma(\mathcal{G}_D(\mathbf{x},\mathbf{y})) \leq \frac{2}{\sqrt{m}} \|C_\mu\|_p^{1/2} \|D\|_q^{1/2},$$

where C_μ is the covariance operator of μ defined by $C_\mu = \mathbb{E}_{x\sim\mu} Q_x$.

(ii) For any environment $\rho \in M_1(M_1(H\times[0,1]))$

$$\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m} \Gamma(\mathcal{G}_D(\mathbf{x},\mathbf{y})) \leq \frac{2}{\sqrt{m}} \|C\|_p^{1/2} \|D\|_q^{1/2},$$

where C is the total covariance operator of μ defined by $C = \mathbb{E}_{\mu\sim\rho}\mathbb{E}_{x\sim\mu} Q_x$.

Proof. The inequalities below are Cauchy-Schwarz', Jensen's and Hölder's (3) inequalities.

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m} \Gamma(\mathcal{G}_D(\mathbf{x},\mathbf{y})) &= \frac{2}{m}\mathbb{E}_{\mathbf{x}\sim\mu^m}\mathbb{E}_\gamma \sup_{\|w\|\leq 1} \sum_{i=1}^m \gamma_i \left\langle w, D^{1/2}x_i\right\rangle \\
&= \frac{2}{m}\mathbb{E}_{\mathbf{x}\sim\mu^m}\mathbb{E}_\gamma \sup_{\|w\|\leq 1} \left\langle w, \sum_{i=1}^m \gamma_i D^{1/2}x_i\right\rangle \\
&\leq \frac{2}{m}\mathbb{E}_{\mathbf{x}\sim\mu^m}\mathbb{E}_\gamma \left\|\sum_{i=1}^m \gamma_i D^{1/2}x_i\right\| \text{ (Cauchy-Schwarz)} \\
&\leq \frac{2}{m}\left(\mathbb{E}_{\mathbf{x}\sim\mu^m}\mathbb{E}_\gamma \left\|\sum_{i=1}^m \gamma_i D^{1/2}x_i\right\|^2\right)^{1/2} \text{ (Jensen)} \\
&= \frac{2}{m}\left(\mathbb{E}_{\mathbf{x}\sim\mu^m} \sum_{i=1}^m \left\|D^{1/2}x_i\right\|^2\right)^{1/2} \text{ (independence of } \gamma_i) \\
&= \frac{2}{\sqrt{m}} \left\langle \mathbb{E}_{x\sim\mu}[Q_x], D\right\rangle_2^{1/2} \text{ (Lemma 1, (ii))} \\
&\leq \frac{2}{\sqrt{m}} \|C_\mu\|_p^{1/2} \|D\|_q^{1/2} \text{ (Hölder).}
\end{aligned}$$

This proves (i). To prove (ii) go through the same steps with $\mathbb{E}_{\mathbf{x}\sim\mu^m}$ replaced by $\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{\mathbf{x}\sim\mu^m}$, and $\mathbb{E}_{x\sim\mu}$ replaced by $\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{x\sim\mu}$ in the last two lines.

Theorem 6. *Suppose that the learning algorithm w satisfies $\|w(\mathbf{x},\mathbf{y})\| \leq 1$ for all $(\mathbf{x},\mathbf{y}) \in (H\times[0,1])^m$, and that for every $y \in [0,1]$ and $K < \infty$ the loss function $\ell(.,y)$ has Lipschitz constant $M(K)$ on the interval $[-K,K]$. Then for every environment $\rho \in M_1(M_1(B_1(H)\times[0,1]))$ and every $D \in \mathcal{L}^+(H)$ we have*

$$R(w_D) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\left[\hat{\ell}_{w_D}(\mathbf{x},\mathbf{y})\right] \leq \sqrt{\frac{2\pi}{m}} M\left(\|D\|_\infty^{1/2}\right) \|D\|_q^{1/2} \|C\|_p^{1/2}.$$

Proof. By assumption we have $\left\|w\left(D^{1/2}\mathbf{x},\mathbf{y}\right)\right\| \leq 1, \forall\left(\mathbf{x},\mathbf{y}\right)$, so we can bound any quantity depending on $w\left(D^{1/2}\mathbf{x},\mathbf{y}\right)$ by the corresponding supremum over all w with $\|w\| \leq 1$. Using this and Theorem 4 (i) and (iii) we get for any $\mu \in M_1\left(B_1\left(H\right)\times[0,1]\right)$

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\left(\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle w_D\left(\mathbf{x},\mathbf{y}\right),x\right\rangle,y\right)\right]-\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)\right)\\
&\leq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\left[\sup_{\|w\|\leq 1}\mathbb{E}_{(x,y)\sim\mu}\ell\left(\left\langle w,D^{1/2}x\right\rangle,y\right)-\frac{1}{m}\sum\ell\left(\left\langle w,D^{1/2}x_i\right\rangle,y_i\right)\right]\\
&\leq \sqrt{\pi/2}\,\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\Gamma\left(\mathcal{H}_D\left(\mathbf{x},\mathbf{y}\right)\right),
\end{aligned}$$

where $\mathcal{H}_D$ is the function class $\mathcal{H}_D = \left\{(x,y)\to\ell\left(\left\langle w,D^{1/2}x\right\rangle,y\right) : \|w\|\leq 1\right\}$. Let $(\mathbf{x},\mathbf{y})$ be any sample in the support of μ^m. For $\|w\|\leq 1$ we have $\left\langle w,D^{1/2}x_i\right\rangle \in \left[-\|D\|_\infty^{1/2},\|D\|_\infty^{1/2}\right]$. On this interval the function $\phi_i : t\mapsto\ell\left(t,y_i\right)$ has Lipschitz constant $M\left(\|D\|_\infty^{1/2}\right)$ by assumption. If $\mathcal{G}_D$ is the function class of the previous lemma then

$$\mathcal{H}_D\left(\mathbf{x},\mathbf{y}\right) = \left\{\left(\phi_1\circ g\left(x_1\right),...,\phi_m\circ g\left(x_m\right)\right) : g\in\mathcal{G}_D\right\},$$

so it follows from Corollary 1 that $\Gamma\left(\mathcal{H}_D\left(\mathbf{x},\mathbf{y}\right)\right) \leq M\left(\|D\|_\infty^{1/2}\right)\Gamma\left(\mathcal{G}_D\left(\mathbf{x}\right)\right)$. From the definition of the transfer risk (1) and the previous lemma we get

$$\begin{aligned}
&R\left(w_D\right)-\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\left[\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)\right]\\
&=\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\left(\mathbb{E}_{(x,y)\sim\mu}\left[\ell\left(\left\langle w_D\left(\mathbf{x},\mathbf{y}\right),x\right\rangle,y\right)\right]-\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)\right)\\
&\leq\sqrt{\pi/2}M\left(\|D\|_\infty^{1/2}\right)\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\Gamma\left(\mathcal{H}_D\left(\mathbf{x},\mathbf{y}\right)\right)\\
&\leq\sqrt{\frac{2\pi}{m}}M\left(\|D\|_\infty^{1/2}\right)\|D\|_q^{1/2}\,\|C\|_p^{1/2}.
\end{aligned}$$

4.2 Predicting the empirical error for the future task

For this section fix a class of operators $\mathcal{D}\subseteq\mathcal{L}^+\left(H\right)$. Now we want to use the average of the empirical errors $\hat{\ell}_{w_D}\left(\mathbf{x}^l,\mathbf{y}^l\right)$ incurred on the tasks of the past to predict the empirical error $\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)$ on the data $(\mathbf{x},\mathbf{y})$ drawn from a future task. We are aiming for a bound on the prediction error

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)-\frac{1}{n}\sum_{l=1}^{n}\hat{\ell}_{w_D}\left(\mathbf{x}^l,\mathbf{y}^l\right)$$

which is uniformly valid for all $D\in\mathcal{D}$. For 1-bounded w we have $\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)\in[0,1]$, so we can invoke Theorem 4 if we have a bound on the Gaussian complexity of the function class

$$\mathcal{F} = \left\{(\mathbf{x},\mathbf{y})\mapsto\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right) : D\in\mathcal{D}\right\}.$$

Since boundedness and kernel-stability are all we need, the corresponding result can be cast in a more general form which can also be used to predict the minimal value of the objective function for regularized least squares regression on future tasks.

The proof reveals the reason why Gaussian averages are used. While Rademacher averages can be compared for Lipschitz functions $\phi : \mathbb{R} \to \mathbb{R}$ (e.g. Theorem 4.12) in [14]) no such results seem available for Lipschitz functions $\phi : \mathbb{R}^d \to \mathbb{R}$, where $\mathbb{R}^d$ is given the euclidean metric.

Theorem 7. *Suppose* $f : (H \times [0,1])^m \to [0,1]$ *satisfies the Lipschitz condition*

$$f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}', \mathbf{y}) \leq \frac{L}{m} \|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}')\|_{Fr}, \tag{5}$$

for all $\mathbf{x}, \mathbf{x}' \in H^m$ *and all* $\mathbf{y} \in [0,1]^m$*. Let* $\mathcal{D}$ *be a class of nonnegative definite operators on* H*. Then with*

$$\mathcal{F} = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto f\left(D^{1/2}\mathbf{x}, \mathbf{y}\right) : D \in \mathcal{D} \right\}$$

we have for any probability measure $\hat{\rho}$ *on* $(B_1(H) \times [0,1])^m$

$$\mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\hat{\rho}^n} \Gamma(\mathcal{F}(\mathbf{X}, \mathbf{Y})) \leq \frac{2L \|\mathcal{D}\|_2}{\sqrt{n}}.$$

Proof. Fix a meta-sample $(\mathbf{X}, \mathbf{Y}) = \left(\left(\mathbf{x}^1, \mathbf{y}^1\right), ..., \left(\mathbf{x}^n, \mathbf{y}^n\right)\right)$. Define Gaussian processes Ω_D and Ξ_D indexed by $\mathcal{D}$ as follows:

$$\Omega_D = \sum_{l=1}^{n} \gamma^l f\left(D^{1/2}\mathbf{x}^l, \mathbf{y}^l\right)$$

$$\Xi_D = \frac{L}{m} \sum_{l=1}^{n} \sum_{i,j=1}^{m} \gamma_{ij}^l \left\langle x_i^l, Dx_j^l \right\rangle,$$

where all the γ^l and γ_{ij}^l are mutually independent $N(0,1)$ distributed variables. Observe that $(2/n)\,\mathbb{E}\sup_D \Omega_D = \Gamma(\mathcal{F}(\mathbf{X}, \mathbf{Y}))$. Then for D_1 and D_2 in $\mathcal{D}$ we have by orthonormality of the γ^l and the Lipschitz condition

$$\begin{aligned}
\mathbb{E}(\Omega_{D_1} - \Omega_{D_2})^2 &= \mathbb{E}_\gamma \left(\sum_{l=1}^{n} \gamma^l \left(f\left(D_1^{1/2}\mathbf{x}^l, \mathbf{y}^l\right) - f\left(D_2^{1/2}\mathbf{x}^l, \mathbf{y}^l\right) \right) \right)^2 \\
&= \sum_{l=1}^{n} \left(f\left(D_1^{1/2}\mathbf{x}^l, \mathbf{y}^l\right) - f\left(D_2^{1/2}\mathbf{x}^l, \mathbf{y}^l\right) \right)^2 \\
&\leq \left(\frac{L}{m}\right)^2 \sum_{l=1}^{n} \left\| \mathbf{G}\left(D_1^{1/2}\mathbf{x}^l\right) - \mathbf{G}\left(D_2^{1/2}\mathbf{x}^l\right) \right\|_{Fr}^2 \text{ by (5)} \\
&= \left(\frac{L}{m}\right)^2 \sum_{l=1}^{n} \sum_{ij=1}^{m} \left(\left\langle x_i^l, D_1 x_j^l \right\rangle - \left\langle x_i^l, D_2 x_j^l \right\rangle \right)^2 \\
&= \mathbb{E}(\Xi_{D_1} - \Xi_{D_2})^2,
\end{aligned}$$

where the last equality follows from orthonormality of the γ_{ij}^l. It then follows from Slepians lemma (Theorem 5) that $\mathbb{E}\sup_D \Omega_D \leq \mathbb{E}\sup_D \Xi_D$. Multiplying with $2/n$ this becomes

$$\Gamma\left(\mathcal{F}\left(\mathbf{X},\mathbf{Y}\right)\right) \leq \frac{2}{n}\mathbb{E}_\gamma \sup_{D\in\mathcal{D}} \frac{L}{m}\sum_{l=1}^{n}\sum_{i,j=1}^{m}\gamma_{ij}^l \left\langle x_i^l, Dx_j^l\right\rangle. \tag{6}$$

Define an operator J_{ij}^l on H by $J_{ij}^l z = \left\langle z, x_i^l\right\rangle x_j^l$. By Lemma 1 (iii) and (iv) we have $\left\|J_{ij}^l\right\|_2 = \left\|x_i^l\right\|\left\|x_j^l\right\|$ and $\left\langle x_i^l, Dx_j^l\right\rangle = \left\langle J_{ij}^l, D\right\rangle_2$. Then with Schwarz' inequality applied to the Hilbert-Schmidt inner product we obtain

$$\sum_{l=1}^{n}\sum_{i,j=1}^{m}\gamma_{ij}^l \left\langle x_i^l, Dx_j^l\right\rangle = \left\langle \sum_{l=1}^{n}\sum_{i,j=1}^{m}\gamma_{ij}^l J_{ij}^l, D\right\rangle_2 \leq \left\|\sum_{l=1}^{n}\sum_{i,j=1}^{m}\gamma_{ij}^l J_{ij}^l\right\|_2 \|D\|_2,$$

so that we get from (6), Jensen's inequality and independence of the γ_{ij}^l

$$\begin{aligned}
\Gamma\left(\mathcal{F}\left(\mathbf{X},\mathbf{Y}\right)\right) &\leq \frac{2L\left\|\mathcal{D}\right\|_2}{nm}\mathbb{E}_\gamma\left\|\sum_{l=1}^{n}\sum_{i,j=1}^{m}\gamma_{ij}^l J_{ij}^l\right\|_2 \\
&\leq \frac{2L\left\|\mathcal{D}\right\|_2}{nm}\left(\mathbb{E}_\gamma\left\|\sum_{l=1}^{n}\sum_{i,j=1}^{m}\gamma_{ij}^l J_{ij}^l\right\|_2^2\right)^{1/2} \\
&= \frac{2L\left\|\mathcal{D}\right\|_2}{nm}\left(\sum_{l=1}^{n}\sum_{i,j=1}^{m}\left\|x_i^l\right\|^2\left\|x_j^l\right\|^2\right)^{1/2} \\
&\leq \frac{2L\left\|\mathcal{D}\right\|_2}{\sqrt{n}},
\end{aligned}$$

almost surely w.r.t. $\hat{\rho}^n$, since $\|x_i\| \leq 1$ almost surely w.r.t. $\hat{\rho}^n$. The conclusion follows.

Using Theorem 4 (ii) and (iii) we immediately obtain

Theorem 8. *Let $\delta > 0$ and let f be as in the previous theorem. Then with probability greater that $1-\delta$ in the draw of the sample $\left(\left(\mathbf{x}^1,\mathbf{y}^1\right),...,\left(\mathbf{x}^n,\mathbf{y}^n\right)\right) \sim \hat{\rho}$ we have for every $D \in \mathcal{D}$ that*

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\left[f\left(D^{1/2}\mathbf{x},\mathbf{y}\right)\right] \leq \frac{1}{n}\sum_{l=1}^{n} f\left(D^{1/2}\mathbf{x}^l,\mathbf{y}^l\right) + \frac{\sqrt{2\pi}L\left\|\mathcal{D}\right\|_2}{\sqrt{n}} + \sqrt{\frac{\ln\left(1/\delta\right)}{2n}}$$

If w is 1-bounded and has kernel stability L then $\hat{\ell}_w$ can be substituted for f. Together with Theorem 6 this implies that with probability greater $1-\delta$ in

$\left(\left(\mathbf{x}^1,\mathbf{y}^1\right),...,\left(\mathbf{x}^n,\mathbf{y}^n\right)\right)$ we have for all $D\in\mathcal{D}$ that

$$\begin{aligned}
&R\left(D\right)-\frac{1}{n}\sum_{l=1}^{n}\hat{\ell}_{w_D}\left(\mathbf{x}^l,\mathbf{y}^l\right)\\
&=\left(R\left(D\right)-\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)\right)+\left(\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\hat{\ell}_{w_D}\left(\mathbf{x},\mathbf{y}\right)-\frac{1}{n}\sum_{l=1}^{n}\hat{\ell}_{w_D}\left(\mathbf{x}^l,\mathbf{y}^l\right)\right)\\
&\leq\sqrt{\frac{2\pi}{m}}M\left(\left\|D\right\|_{\infty}^{1/2}\right)\left\|D\right\|_q^{1/2}\left\|C\right\|_p^{1/2}+\sqrt{\frac{2\pi}{n}}L\left\|\mathcal{D}\right\|_2+\sqrt{\frac{\ln\left(1/\delta\right)}{2n}},
\end{aligned}$$

which gives Theorem 2.

5 Regularized least squares regression

For any sample $(\mathbf{x},\mathbf{y})\in\left(H\times[0,1]\right)^m$ the algorithm ω and its empirical loss $\hat{\ell}_\omega$ and minimal objective value ξ are defined by

$$\begin{aligned}
\omega\left(\mathbf{x},\mathbf{y}\right)&=\arg\min_{w\in H}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\left\langle w,x_i\right\rangle-y_i\right)^2+\left\|w\right\|^2\right),\\
\hat{\ell}_\omega\left(\mathbf{x},\mathbf{y}\right)&=\frac{1}{m}\sum_{i=1}^{m}\left(\left\langle\omega\left(\mathbf{x},\mathbf{y}\right),x_i\right\rangle-y_i\right)^2,\\
\xi\left(\mathbf{x},\mathbf{y}\right)&=\min_{w\in H}\left(\frac{1}{m}\sum_{i=1}^{m}\left(\left\langle w,x_i\right\rangle-y_i\right)^2+\left\|w\right\|^2\right).
\end{aligned}$$

The first two assertions in the following proposition state that, relative to square loss, ω satisfies the hypotheses of Theorem 2 with $L=2$. The third assertion implies that the minimal value ξ of the regularized objective can also be estimated uniformly over the set of pre-processing operators.

Proposition 1. *$\forall\mathbf{x},\mathbf{x}'\in H^m,\mathbf{y}\in[0,1]^m$ we have*
(i) $\left\|\omega\left(\mathbf{x},\mathbf{y}\right)\right\|\leq 1$ and $\hat{\ell}_\omega\left(\mathbf{x},\mathbf{y}\right)\leq 1$ and $\xi\left(\mathbf{x},\mathbf{y}\right)\leq 1$.
(ii) $\hat{\ell}_\omega\left(\mathbf{x},\mathbf{y}\right)-\hat{\ell}_\omega\left(\mathbf{x}',\mathbf{y}\right)\leq\left(2/m\right)\left\|\mathbf{G}\left(\mathbf{x}\right)-\mathbf{G}\left(\mathbf{x}'\right)\right\|_{Fr}$
(iii) $\xi\left(\mathbf{x},\mathbf{y}\right)-\xi\left(\mathbf{x}',\mathbf{y}\right)\leq\left(1/m\right)\left\|\mathbf{G}\left(\mathbf{x}\right)-\mathbf{G}\left(\mathbf{x}'\right)\right\|_{Fr}$.

For the proof we use the following lemma, the proof of which is given in [15], Lemma 11.

Lemma 3. *Let G_1 and G_2 be positive semidefinite operators on any Hilbert space and $\lambda>0$. Then*
1. $G_i+\lambda I$ is invertible,
2. $\left\|\left(G_i+\lambda I\right)^{-1}\right\|_\infty\leq 1/\lambda$ and
3. we have

$$\left\|\left(G_1+\lambda I\right)^{-1}-\left(G_2+\lambda I\right)^{-1}\right\|_\infty\leq\frac{1}{\lambda^2}\left\|G_1-G_2\right\|_\infty.$$

4. Let x_1 and x_2 satisfy $(G_i + \lambda I) x_i = y$. Then

$$\left| \|x_1\|^2 - \|x_2\|^2 \right| \leq 2\lambda^{-3} \|G_1 - G_2\|_\infty \|y\|^2 .$$

Proof (Proof of Proposition 1). We have

$$\begin{aligned} \hat{\ell}_\omega (\mathbf{x}, \mathbf{y}) + \|\omega (\mathbf{x}, \mathbf{y})\|^2 &= \min_{w \in H} \hat{\ell}_{\text{sqr}} (\langle w, \mathbf{x} \rangle, \mathbf{y}) + \|w\|^2 \\ &\leq \hat{\ell}_{\text{sqr}} (\langle 0, \mathbf{x} \rangle, \mathbf{y}) + \|0\|^2 \leq 1, \end{aligned}$$

which proves (i), since both terms on the left are nonnegative and their sum is $\xi (\mathbf{x}, \mathbf{y})$.

A standard argument shows that $\exists \boldsymbol{\alpha} = \boldsymbol{\alpha} (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m$ such that $\omega (\mathbf{x}, \mathbf{y}) = \sum_{j=1}^m \alpha_j x_j$ with $\boldsymbol{\alpha} (\mathbf{x}, \mathbf{y}) = (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y}$ where I is the identity matrix. Also $\langle \omega (\mathbf{x}, \mathbf{y}), x_i \rangle = \sum_{j=1}^m \alpha_j \langle x_j, x_i \rangle = (\mathbf{G} (\mathbf{x}) \boldsymbol{\alpha})_i$, so

$$\begin{aligned} \hat{\ell}_\omega (\mathbf{x}, \mathbf{y}) &= \frac{1}{m} \sum_{i=1}^m ((\mathbf{G} (\mathbf{x}) \boldsymbol{\alpha})_i - y_i)^2 \\ &= \frac{1}{m} \sum_{i=1}^m (((\mathbf{G} (\mathbf{x}) + mI) \boldsymbol{\alpha})_i - y_i - m\alpha_i)^2 \\ &= m \sum \alpha_i^2 = m \left\| (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y} \right\|_{\mathbb{R}^m}^2 . \end{aligned} \tag{7}$$

We therefore also have

$$\begin{aligned} & \xi (\mathbf{x}, \mathbf{y}) \\ &= \|\omega (\mathbf{x}, \mathbf{y})\|^2 + \hat{\ell}_\omega (\mathbf{x}, \mathbf{y}) \\ &= \langle \boldsymbol{\alpha}, \mathbf{G} (\mathbf{x}) \boldsymbol{\alpha} \rangle_{\mathbb{R}^m} + m \left\| (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y} \right\|_{\mathbb{R}^m}^2 \\ &= \left\langle (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y}, \mathbf{G} (\mathbf{x}) (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y} + m (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y} \right\rangle_{\mathbb{R}^m} \\ &= \left\langle (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y}, \mathbf{y} \right\rangle_{\mathbb{R}^m} . \end{aligned} \tag{8}$$

Using identity (7) and Lemma 3 applied to $\mathbb{R}^m$ we get

$$\left| \hat{\ell}_\omega (\mathbf{x}, \mathbf{y}) - \hat{\ell}_\omega (\mathbf{x}', \mathbf{y}) \right| = m \left(\left| \left\| (\mathbf{G} (\mathbf{x}) + mI)^{-1} \mathbf{y} \right\|_{\mathbb{R}^m}^2 - \left\| (\mathbf{G} (\mathbf{x}') + mI)^{-1} \mathbf{y} \right\|_{\mathbb{R}^m}^2 \right| \right)$$
$$\leq 2m^{-2} \|\mathbf{G} (\mathbf{x}) - \mathbf{G} (\mathbf{x}')\|_{\mathbb{R}^m, \infty} \|\mathbf{y}\|_{\mathbb{R}^m}^2 .$$

Now $\|\mathbf{y}\|_{\mathbb{R}^m}^2 = \sum y_i^2 \leq m$. Also $\|\mathbf{G} (\mathbf{x}) - \mathbf{G} (\mathbf{x}')\|_{\mathbb{R}^m, \infty} \leq \|\mathbf{G} (\mathbf{x}) - \mathbf{G} (\mathbf{x}')\|_{Fr}$ so (ii) follows. Proceeding similarly we get

$$\begin{aligned} & |\xi (\mathbf{x}, \mathbf{y}) - \xi (\mathbf{x}', \mathbf{y})| = m \left| \left\langle \left((\mathbf{G} (\mathbf{x}) + mI)^{-1} - (\mathbf{G} (\mathbf{x}') + mI)^{-1} \right) \mathbf{y}, \mathbf{y} \right\rangle_{\mathbb{R}^m} \right| \\ & \leq m^{-2} \|\mathbf{G} (\mathbf{x}) - \mathbf{G} (\mathbf{x}')\|_{\mathbb{R}^m, \infty} \|\mathbf{y}\|_{\mathbb{R}^m}^2 \\ & \leq m^{-1} \|\mathbf{G} (\mathbf{x}) - \mathbf{G} (\mathbf{x}')\|_{Fr} . \end{aligned}$$

QED

We can now apply Theorem 2 to regularized least squares regression. On the interval $[-K, K]$ the square loss $\ell_{\mathrm{sqr}}(., y)$ has Lipschitz constant $2(K+1)$, so we can set $M_K = 2(K+1)$. By the above ω satisfies the hypotheses of Theorem 2 with $L = 2$.

To get Theorem 1 in the introduction we take $\mathcal{D} = \left\{\lambda^{-1} P : P \in \mathcal{P}_d\right\}$, where $\lambda > 0$ is fixed and $\mathcal{P}_d$ is the set of orthogonal projections in H with d-dimensional range. For fixed $P \in \mathcal{P}_d$ the algorithm $\omega_{\lambda^{-1}P}$ is just ordinary RLSR with regularizing parameter λ and a d-dimensional subspace constraint expressed by the projection P.

We have $\|\mathcal{D}\|_2 = \lambda^{-1}\sqrt{d}$ and $\|D\|_\infty = \lambda^{-1}$ and $\|D\|_1 = \lambda^{-1} d$ for all $D \in \mathcal{D}$. Let us also set $q = 1$ and $p = \infty$ and assume $\lambda \leq 1$. The bound (slightly loosened for the sake of simplicity) then says that with probability greater $1 - \delta$ we have for all $P \in \mathcal{P}_d$ that

$$R(\omega_{\lambda^{-1}P}) \leq \frac{1}{n}\sum_{l=1}^{n} \hat{\ell}_{\omega_{\lambda^{-1}P}}\left(\mathbf{x}^l, \mathbf{y}^l\right) + \frac{\sqrt{8\pi d}}{\lambda}\left(2\sqrt{\frac{\|C\|_\infty}{m}} + \sqrt{\frac{1}{n}}\right) + \sqrt{\frac{\ln(1/\delta)}{2n}},$$

which is the bound of Theorem 1, obtained as a corollary to Theorem 2.

Another interesting consequence of Proposition 1 regards the minimal objective function ξ. From Theorem 8 we obtain

Corollary 2. *For $\delta > 0$, with probability greater that $1 - \delta$ in the draw of the sample $\left(\left(\mathbf{x}^1, \mathbf{y}^1\right), ..., \left(\mathbf{x}^n, \mathbf{y}^n\right)\right) \sim \hat{\rho}$ we have for every $D \in \mathcal{D}$ that*

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\rho}}\left[\xi\left(D^{1/2}\mathbf{x}, \mathbf{y}\right)\right] \leq \frac{1}{n}\sum_{l=1}^{n} \xi\left(D^{1/2}\mathbf{x}^l, \mathbf{y}^l\right) + \frac{\sqrt{\pi} L \|\mathcal{D}\|_2}{\sqrt{2n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

It follows that the expected values of $\left\|\omega\left(D^{1/2}\mathbf{x}, \mathbf{y}\right)\right\|$, $\hat{\ell}_\omega\left(D^{1/2}\mathbf{x}, \mathbf{y}\right)$ and $\xi\left(D^{1/2}\mathbf{x}, \mathbf{y}\right)$ for future samples $(\mathbf{x}, \mathbf{y})$ can all be estimated uniformly for $D \in \mathcal{D}$.

6 Related work

There has been considerable interest in transfer learning, learning-to-learn and multi-task learning, and many encouraging experimental results have been obtained ([9], see also [22]). The study of learning beyond the confinements of single tasks is motivated by the study of biological learning systems, where it is observed that, in addition to the rewards obtained for learning the task at hand, learners can benefit by improving their learning ability for future, yet unknown tasks.

6.1 Multi-task vs transfer-learning

The theoretical properties of corresponding machine learning systems were studied in depth by J. Baxter [6], who introduced the notion of an environment (essential for the statement of our results) and realized that there are two distinct aspects to his findings

- "Learning multiple related tasks reduces the sampling burden required for good generalization, at least on a number-of-examples-required-per-task basis.
- Bias that is learnt on sufficiently many training tasks is likely to be good for learning novel tasks drawn from the same environment."[6]

The first observation regards the subject of *multi-task learning* (MTL), where the task-distributions $\mu_1, ..., \mu_n$ are fixed in advance and potential relations between the tasks are exploited to minimize the *expected loss for the same set of tasks*. From the available data the MTL algorithm produces a *multi-hypothesis*, that is a vector of hypotheses, one for each task at hand. Theoretical studies of MTL typically bound the task-average expected loss of such a multi-hypothesis (see [6],[1],[16]), but under certain specific assumptions on the relationships between the various tasks better results are obtainable, which uniformly bound the expected loss for each task in the set (as in Ben-David [7]).

In either case the bounds can be expressed in terms of an empirical average of the loss plus a bound $B(n,m)$ on the estimation error, where n is the number of tasks and m the number of examples per task. If such bounds are any good, we better have $B(n,m) \to 0$ as $m \to \infty$ for all n, because this is what we already get for single task learning. The specific benefit of multi-task learning is exposed in the limit $n \to \infty$, for fixed and possibly rather small m. If we have $\limsup_{n\to\infty} B(n,m) \leq B_0(m)$, where B_0 would be a bound for some competing single-task algorithm, then multi-task learning is preferable to ordinary learning, for a large number of tasks and the given number of examples per task.

The second point raised by Baxter addresses learning-to-learn (LTL) and transfer, and - if the term *bias* is specialized to *linear feature map* - the subject of this paper. Here the notion of an environment, which is irrelevant to the discussion of multi-task learning, becomes crucial. The quantity to be bounded is the transfer-risk introduced above, or an intermediate construction as in [6], and again the bound is formulated as the sum of an empirical average and a bound $B(n,m)$ on the estimation error. But, owing to the fact that we now bound the expected loss for a new task and not for an ensemble of known tasks, there are essential differences. We will not have $B(n,m) \to 0$ as $m \to \infty$ for all n, because there is an extra estimation difference between the expected loss on a new task and the average expected loss on the sampled tasks. For the bounds to be tight we should have $B(n,m) \to 0$ as $m \to \infty$ *and* $n \to \infty$. Theorem 2 satisfies this requirement.

To highlight the differences of bounds for LTL and bounds for MTL we give an example of the latter, taken from [16] and modified for easy comparison to Theorem 2 (for the case $p = q = 2$):

Theorem 9. *Suppose $M(.)$ and ℓ are as in Theorem 2.*

Then for every set of tasks $(\mu_1, ..., \mu_n) \in M_1(B_1(H) \times [0,1])^n$ and every $\forall \delta > 0$, with probability greater than $1-\delta$ in the draw $\left(\left(\mathbf{x}^1, \mathbf{y}^1\right), ..., \left(\mathbf{x}^n, \mathbf{y}^n\right)\right) \sim$

$\prod_{l=1}^{n}(\mu_l)^m$ *it holds for all* $D \in \mathcal{D}$ *and all* $\left(v^1, ..., v^n\right) \in B_1(H)^n$ *that*

$$\frac{1}{n}\sum_{l=1}^{n}\mathbb{E}_{(x,y)\sim\mu^l}\left[\ell\left(\left\langle v^l, D^{1/2}x\right\rangle, y\right)\right] \leq \frac{1}{nm}\sum_{l=1}^{n}\sum_{i=1}^{m}\ell\left(\left\langle v^l, D^{1/2}x_i^l\right\rangle, y_i^l\right) + \frac{2}{\sqrt{m}}M\left(\|\mathcal{D}\|_\infty^{1/2}\right)\|\mathcal{D}\|_2^{1/2}\left(\|C\|_2^2 + \frac{3}{n}\right)^{1/4} + \sqrt{\frac{\ln(1/\delta)}{2nm}}.$$

Here the environment ρ is replaced by the fixed task distributions $\mu_1, ..., \mu_n$ and the right hand side of the above inequality can only be determined from realizations of *all* of these distributions. The MTL Theorem 9 does therefore not imply generalization beyond this fixed set of tasks and is weaker than the LTL Theorem 2 in this respect. But the MTL bound works for any algorithm selecting its weights from the unit ball, while our LTL bound requires the weight vectors v^l to be obtained by an algorithm with specific Lipschitz properties. Theorem 9 can therefore not be derived by applying Theorem 2 to the environment $\rho = (1/n)\sum_l \delta_{\mu_l}$. The two results are inequivalent.

For any number of tasks n the bound on the estimation error in the MTL Theorem 9 goes to zero as $m \to \infty$, as we required for multitask learning. But for fixed sample size m and $n \to \infty$ the limit is

$$\frac{2}{\sqrt{m}}M\left(\|\mathcal{D}\|_\infty^{1/2}\right)\|\mathcal{D}\|_2^{1/2}\|C\|_2^{1/2}$$

which is, apart from the factor $\sqrt{\pi/2} \approx 1.25$, the same as in the LTL Theorem 2, which shows an important similarity of the two results: In the presence of algorithms with appropriate Lipschitz properties and for a large number of tasks, any argument in favour of MTL over single task learning should also give an argument in favour of LTL over single task learning.

Another similarity of MTL and LTL is the following: In practice multi-task learning and learning to learn both operate on samples drawn from multiple tasks, and for clever MTL algorithms learning to learn can be a by-product (as in [1] and [3], see also section 6.3 below). These similarities are the reason why the two subjects are easily confused.

6.2 Bounds for learning to learn

The comparison of our result to other bounds is made easy by the fact that, despite the fascinating quality of the subject, to the author's knowledge there are only the pioneering work by Baxter [6] and a precursor of the present work [15] where bounds for learning to learn are proposed.

Baxter's paper follows the paradigm of empirical risk minimization, where the choice of a learning algorithm is equivalent to the choice of the hypothesis space $\mathcal{H}$ from which the algorithm chooses its hypothesis, so as to minimize empirical

risk. This is similar to the learning of a d-dimensional subspace constraint as put forward in the introduction. In the general case studied here, where we select from general classes $\mathcal{D}$ of Hilbert-Schmidt operators, this analogy no longer holds, because the ranges of all $D \in \mathcal{D}$ may be the same and equal to H.

Let us write $w_{\mathcal{H}}$ for the ERM algorithm (assuming all hypothesis spaces to be finite for simplicity)

$$w_{\mathcal{H}}(\mathbf{x}, \mathbf{y}) = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h(x), y).$$

If $\mathbb{H}$ is a collection of hypothesis spaces, then $\{w_{\mathcal{H}} : \mathcal{H} \in \mathbb{H}\}$ is a collection of algorithms and within the framework of the present paper we would look for a high probability bound on the transfer risk $R(w_{\mathcal{H}})$ uniformly over all $\mathcal{H} \in \mathbb{H}$. Baxter does not give such a bound directly, but gives a high probability bound on

$$\mathrm{er}_{\rho}(\mathcal{H}) = \mathbb{E}_{\mu \sim \rho} \min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mu} \ell(h(x), y),$$

uniformly over all $\mathcal{H} \in \mathbb{H}$, and we can write

$$\begin{aligned} & R(w_{\mathcal{H}}) - \frac{1}{n} \sum_{l=1}^{n} \hat{\ell}_{w_{\mathcal{H}}}\left(\mathbf{x}^l, \mathbf{y}^l\right) \\ &= \left(R(w_{\mathcal{H}}) - \mathrm{er}_{\rho}(\mathcal{H})\right) + \left(\mathrm{er}_{\rho}(\mathcal{H}) - \frac{1}{n} \sum_{l=1}^{n} \hat{\ell}_{w_{\mathcal{H}}}\left(\mathbf{x}^l, \mathbf{y}^l\right)\right). \end{aligned}$$

If every hypotheses space $\mathcal{H} \in \mathbb{H}$ has bounded capacity, then we can find a bound of $O\left(m^{-1/2}\right)$ on the first term, and a bound on the transfer risk results, making the results comparable.

The bound on $\mathrm{er}_{\rho}(\mathcal{H})$ is formulated in terms of covering numbers of $\mathbb{H}$, which depend on the unknown distribution ρ. When converted to a concrete bound (e.g. in Theorem 8 in [6]) it becomes dimension dependent and becomes trivial for many kernelized forms of transfer learning (for example with Gaussian RBF-kernels), in contrast to the dimension free results in this paper.

In [15] the alternative decomposition

$$\begin{aligned} & R(w_{\mathcal{H}}) - \frac{1}{n} \sum_{l=1}^{n} \hat{\ell}_{w_{\mathcal{H}}}\left(\mathbf{x}^l, \mathbf{y}^l\right) \\ &= \left(R(w_{\mathcal{H}}) - \mathbb{E}_{(\mathbf{x},\mathbf{y})} \hat{\ell}_{w_{\mathcal{H}}}(\mathbf{x}, \mathbf{y})\right) + \left(\mathbb{E}_{(\mathbf{x},\mathbf{y})} \hat{\ell}_{w_{\mathcal{H}}}(\mathbf{x}, \mathbf{y}) - \frac{1}{n} \sum_{l=1}^{n} \hat{\ell}_{w_D}\left(\mathbf{x}^l, \mathbf{y}^l\right)\right) \end{aligned}$$

is proposed, as it is used in this paper. Again the first term is bounded in $O\left(m^{-1/2}\right)$ if all the $\mathcal{H} \in \mathbb{H}$ have bounded capacity. To bound the second term one seeks to find a bound on the capacity of the function class

$$\mathcal{F}_{\mathbb{H}} = \left\{(\mathbf{x}, \mathbf{y}) \mapsto \hat{\ell}_{w_{\mathcal{H}}}(\mathbf{x}, \mathbf{y}) : \mathcal{H} \in \mathbb{H}\right\}.$$

In [15] (Proposition 9) it is shown, that the $\mathcal{N}_1$-covering numbers of this function class are always bounded by the capacities used in [6], and may be finite when the latter become infinite, indicating that the proposed decomposition is preferable. Instead of using covering numbers [15] proposes stability arguments to bound the second term in the decomposition. The stability requirement however imposes severe restrictions on the class of LTL algorithms to which the results apply. In [15] a heuristically motivated meta-algorithm (the "chorus of prototypes" taken from [10]) is proposed, which satisfies these requirements.

The restriction to linear feature maps in combination with the use of empirical process theory to bound the second term in the decomposition above, as in the present paper, leads to generalization bounds for LTL extensions of popular algorithms, such as ridge regression (see the next section), still applicable in a kernelized, infinite-dimensional setting.

6.3 Convex multi-task learning

While the problems of finding bounds for transfer and multi-task learning are somewhat distinct, algorithms for multi-task learning can be used for transfer, if, along with the requested multi-hypothesis, they also output some common structural parameter, which can be used in future learning.

An algorithm with particularly attractive properties is given in [3]. The square-loss version of it minimizes the functional

$$F(W,D) = \frac{1}{nm}\sum_{l=1}^{n}\sum_{i=1}^{m}\ell_{\mathrm{sqr}}\left(\left\langle w^l, x_i^l\right\rangle, y_i^l\right) + \lambda\frac{1}{n}\sum_{l=1}^{n}\left\langle D^+ w^l, w^l\right\rangle,$$

where $\left(w^l\right)_{l=1}^n$ is the multi-hypothesis returned by the algorithm. $\lambda > 0$ is a regularization constant and $D \in \mathcal{L}^+(H)$ is positive and satisfies $\|D\|_1 \leq 1$ and D^+ is the pseudoinverse of D. We can absorb λ in the constraint on D by demanding $D \in \mathcal{D} = \left\{D \in \mathcal{L}^+(H) : \|D\|_1 \leq \lambda^{-1}\right\}$. With a change of variables $v^l = D^{1/2}w^l$ we find that the algorithm selects the pre-processor $D(\mathbf{X},\mathbf{Y})^{1/2}$, where $D(\mathbf{X},\mathbf{Y})^{1/2}$ is given by

$$\begin{aligned} D(\mathbf{X},\mathbf{Y}) &= \arg\min_{D\in\mathcal{D}}\min_{v^1,\ldots,v^n\in H}\frac{1}{n}\sum_{l=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}\ell_{\mathrm{sqr}}\left(\left\langle v^l, D^{1/2}x_i^l\right\rangle, y_i^l\right) + \left\langle v^l, v^l\right\rangle\right) \\ &= \arg\min_{D\in\mathcal{D}}\frac{1}{n}\sum_{l=1}^{n}\left(\hat{\ell}_{\omega_D}\left(\mathbf{x}^l,\mathbf{y}^l\right) + \left\|\omega_D\left(\mathbf{x}^l,\mathbf{y}^l\right)\right\|^2\right), \end{aligned}$$

so it minimizes an upper bound to $(1/n)\sum_{l=1}^n \hat{\ell}_{\omega_D}\left(\mathbf{x}^l,\mathbf{y}^l\right)$. Since $\|\mathcal{D}\|_\infty \leq \|\mathcal{D}\|_2 \leq \|\mathcal{D}\|_1 = \lambda^{-1}$ we obtain for this algorithm from Theorem 2 the generalization guarantee that w.h.p.

$$R(D) \leq (1/n)\sum_{l=1}^{n}\hat{\ell}_{\omega_D}\left(\mathbf{x}^l,\mathbf{y}^l\right) + \lambda^{-1}\left(\frac{8\|C\|_\infty^{1/2}}{\sqrt{m}} + \frac{\sqrt{8\pi}}{\sqrt{n}}\right) + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

It also follows from Corollary 2 that the expected value of the minimal objective for least squares regression on the future task can be well estimated from the quantity which is being minimized for D. The algorithm has an efficient implementation and is shown to converge in [3].

Note that this algorithm's primary purpose is to find a good multi-hypothesis for the given tasks, so as to satisfy a set of current customers of the machine learning agency. The utility of the feature map $D^{1/2}$ (or the regularizer D^{+}) for future learning is a pleasant by-product of this effort.

References

1. R. K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, *Journal of Machine Learning Research*, 6: 1817-1853, 2005.
2. M. Anthony, P. Bartlett, *Learning in Neural Networks: Theoretical Foundations*, Cambridge University Press 1999
3. A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, to appear.
4. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3: 463-482, 2002.
5. P.Bartlett, O.Bousquet and S.Mendelson. Local Rademacher complexities. Available online: http://www.stat.berkeley.edu/~bartlett/papers/bbm-lrc-02b.pdf.
6. J.Baxter, A Model of Inductive Bias Learning, *Journal of Artificial Intelligence Research* 12: 149-198, 2000
7. S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *COLT 03*, 2003.
8. O. Bousquet, A. Elisseeff, "Stability and Generalisation", *Journal of Machine Learning Research*, 2: 499-526, 2002.
9. R.Caruana, Multitask Learning, *Machine Learning*, 28: 41-75, 1997.
10. S.Edelman, Representation, similarity and the chorus of prototypes. *Minds and Machines*, 45-68, 1995
11. T. Evgeniou and M. Pontil, Regularized multi-task learning. *Proc. Conference on Knowledge Discovery and Data Mining*, 2004.
12. T. Evgeniou, C. Micchelli and M. Pontil, Learning multiple tasks with kernel methods. *JMLR*, 6: 615-637, 2005.
13. 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.
14. M. Ledoux, M. Talagrand, *Probability in Banach Spaces*, Springer 1991.
15. A. Maurer, Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6:967–994, 2005.
16. A. Maurer, Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.
17. C.McDiarmid, *Concentration*, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, (1998) 195-248. Springer, Berlin
18. Michael Reed and Barry Simon. *Fourier Analysis, Self-Adjointness*, part II of *Methods of Mathematical Physics, Academic Press*, 1980.
19. Michael Reed and Barry Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.

20. J. Shawe-Taylor, N. Cristianini, Estimating the moments of a random vector, *Proceedings of GRETSI 2003 Conference*, I: 47–52, 2003.
21. J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, J.S. Kandola: On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory* 51(7): 2510-2522, 2005.
22. S.Thrun, Lifelong Learning Algorithms, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998
23. W. van der Vaart, J. A. Wellner, *Weak convergence and empirical processes*, Springer 1996
24. L. Zwald, O. Bousquet and G. Blanchard. Statistical properties of kernel principal component analysis. *Machine Learning* 66: 259–294, 2007.

Appendix: Glossary of terms and notation

Notation	Short Description	Section
H	real, separable Hilbert space	3.1,2
$\langle .,. \rangle$ and $\|.\|$	inner product and norm on H	3.1,2
$B_1(H)$	unit ball in $H : \{x \in H : \|x\| \leq 1\}$	3.1,2
H_2	Hilbert-Schmidt operators on H	3.1
$\mathcal{L}(H)$	bounded linear operators on H	3.1
$\|.\|_\infty$	operator norm $\|T\|_\infty = \sup_{\|x\| \leq 1} \|Tx\|$	
$\mathcal{L}^+(H)$	positive semidefinite members of $\mathcal{L}(H)$	3.1
$\mathcal{L}_2(H)$	Hilbert Schmidt operators on H	3.1
$\mathcal{L}_2^+(H)$	positive semidefinite members of $\mathcal{L}_2(H)$	3.1
$\langle .,. \rangle_2$	Frobenius inner product $\langle T, S \rangle_2 = tr(S^*T)$	3.1
$\|.\|_p$	Schatten norm $\|T\|_p = tr\left(\|T\|^p\right)^{1/p}$, $p \geq 1$	3.1
$\mathcal{P}_d$	d-dimensional orthogonal projections in H	3.1,1
Q_x, for $x \in H$	operator $Q_x z = \langle z, x \rangle x$, $\forall z \in H$	3.1
$J_{x,y}$, for $x, y \in H$	operator $J_{x,y} z = \langle z, x \rangle y$, $\forall z \in H$	3.1,4.2
$C = \mathbb{E}Q_X$	covariance operator of r.v. X in H	3.1,1
$\|.\|_{Fr}$	Frobenius norm for $m \times m$ matrices $\|A\|_{Fr} = \left(\sum_{ij} A_{ij}^2\right)^{1/2}$	3.1
$\|.\|_{\mathbb{R}^m}$	euclidean norm on $\mathbb{R}^m$	3.1
$M_1(\mathcal{X})$	probability distributions on $\mathcal{X}$	2.1
$M_1(B_1(H) \times [0,1])$	set of learning tasks	2.1
μ	$\in M_1(B_1(H) \times [0,1])$ a generic learning task	2.1
$(x, y) \sim \mu$	the draw of an input-output pair from μ	2.1
C_μ	covariance operator $C_\mu = \mathbb{E}_{\mu \sim \rho} \mathbb{E}_{(x,y) \sim \mu} Q_x$	3.1
$(\mathbf{x}, \mathbf{y})$	$\in (H \times [0,1])^m$ a *sample* of m input-output pairs $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), ..., (x_m, y_m))$	2.1
$\mathbf{G}(\mathbf{x})$	Gramian matrix $\mathbf{G}(\mathbf{x})_{ij} = \langle x_i, x_j \rangle$	2.1
$(\mathbf{x}, \mathbf{y}) \sim \mu^m$	an iid draw of a sample	2.1
ρ	a prob. distribution on $M_1(B_1(H) \times [0,1])$	2.4
$\mu \sim \rho$	the draw of a learning task from ρ	2.4
C	total covariance operator $\mathbb{E}_{\mu \sim \rho} C_\mu$	
$\hat{\rho}$	$\in M_1\left((B_1(H) \times [0,1])^m\right)$ the distribution $\hat{\rho}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mu \sim \rho} \mu^m(\mathbf{x}, \mathbf{y})$	2.4
$(\mathbf{x}, \mathbf{y}) \sim \hat{\rho}$	draw of task, followed by iid draw of sample	2.4
$(\mathbf{X}, \mathbf{Y})$	$\in (H \times [0,1])^{mn}$ $\left((\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^n, \mathbf{y}^n)\right)$ a meta-sample	2.4
$(\mathbf{X}, \mathbf{Y}) \sim \hat{\rho}^n$	iid draw of meta-sample in n trials of $\hat{\rho}$	2.4

$w(\mathbf{x},\mathbf{y})$	weight computed by algorithm w from $(\mathbf{x},\mathbf{y})$	2.2
ℓ	loss function $\ell : \mathbb{R} \times [0,1] \to \mathbb{R}^+$	2
ℓ_{sqr}	square loss $\ell_{\text{sqr}}(y', y) = (y' - y)^2$	2
$\hat{\ell}_w(\mathbf{x},\mathbf{y})$	empirical loss of algorithm w on $(\mathbf{x},\mathbf{y})$ $\hat{\ell}_w(\mathbf{x},\mathbf{y}) = (1/m)\sum_{i=1}^m \ell(\langle w(\mathbf{x},\mathbf{y}), x_i\rangle, y_i)$	2.2
$R_\rho(w)$	transfer risk of algorithm w in environment ρ $= \mathbb{E}_{\mu\sim\rho}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mu^m}\mathbb{E}_{(x,y)\sim\mu}\ell(\langle w(\mathbf{x},\mathbf{y}), x\rangle, y)$	2.4
w_D	for $D \in \mathcal{L}^+(H)$ the algorithm w using $D^{1/2}$ as a preprocessor $w_D(\mathbf{x},\mathbf{y}) = D^{1/2} w\left(D^{1/2}\mathbf{x},\mathbf{y}\right)$	2.3
$\mathcal{D}$	a generic subset $\mathcal{D} \subset \mathcal{L}^+(H)$	2.3
$\|\mathcal{D}\|_p$	$\sup\left\{\|D\|_p : D \in \mathcal{D}\right\}$	
ω	regularized least squares: $\arg\min_{w\in H}(1/m)\sum_i(\langle w, x_i\rangle - y_i)^2 + \|w\|^2$	2.2,5
ξ	minimal objective for ω	5
σ, σ_i	independent random var's, uniform on $\{-1, 1\}$	3.2
$\gamma, \gamma_i, \gamma_{ij}^l$	independent random var's, $N(0,1)$-distributed	3.2
$\mathcal{R}(A)$	Rademacher average of $A \subseteq \mathbb{R}^k$	3.2
$\Gamma(A)$	Gaussian average of $A \subseteq \mathbb{R}^k$	3.2
$\mathcal{F}$	generic function class on a space $\mathcal{X}$	
$\mathcal{F}(\mathbf{x})$	The set $\{(f(x_1), ..., f(x_k)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^k$	3.2