Journal of Machine Learning Research 17 (2016) 1-32 Submitted 5/15; Revised 10/15; Published 4/16

The Benefit of Multitask Representation Learning

Andreas Maurer AM@ANDREAS-MAURER.EU
Adalbertstrasse 55, D-80799 München, Germany
Massimiliano Pontil MASSIMILIANO.PONTIL@IIT.IT
Istituto Italiano di Tecnologia, 16163, Genoa, Italy
Department of Computer Science, University College London, WC1E 6BT, UK
Bernardino Romera-Paredes BERNARD@ROBOTS.OX.AC.UK
Department of Engineering Science, University of Oxford, OX1 3PJ, UK

Editor: Urun Dogan, Marius Kloft, Francesco Orabona, and Tatiana Tommasi

Abstract

We discuss a general method to learn data representations from multiple tasks. We provide a justification for this method in both settings of multitask learning and learning-to-learn. The method is illustrated in detail in the special case of linear feature learning. Conditions on the theoretical advantage offered by multitask representation learning over independent task learning are established. In particular, focusing on the important example of half-space learning, we derive the regime in which multitask representation learning is beneficial over independent task learning, as a function of the sample size, the number of tasks and the intrinsic data dimensionality. Other potential applications of our results include multitask feature learning in reproducing kernel Hilbert spaces and multilayer, deep networks.

Keywords: learning-to-learn, multitask learning, representation learning, statistical learning theory, transfer learning

1. Introduction

Multitask learning (MTL) can be characterized as the problem of learning multiple tasks *jointly*, as opposed to learning each task in isolation. This problem is becoming increasingly important due to its relevance in many applications, ranging from modelling users' preferences for products, to multiple object classification in computer vision, to patient healthcare data analysis in health informatics, to mention but a few. Multitask learning algorithms which exploit structure and similarities across different learning problems have been studied by the machine learning community since the mid 90's, initially in connection to neural network models (see Baxter, 2000; Caruana, 1998; Thrun and Pratt, 1998, and reference therein). More recent approaches have been based on kernel methods (Evgeniou et al., 2005), structured sparsity and convex optimization (Argyriou et al., 2008), among others.

Closely related to multitask learning but more challenging is the problem of *learning-to-learn* (LTL), namely learning to perform a new task by exploiting knowledge acquired when solving previous tasks. Arguably, a solution to this problem would have major impact in

Artificial Intelligence as we could build machines which learn from experience to perform new tasks, similar to what we observe in human behavior.

An influential line of research on multitask and transfer learning is based on the idea that the tasks are related by means of a common low dimensional representation, which is learned jointly with the tasks' parameters. This approach was first advocated in (Baxter, 2000; Caruana, 1998; Thrun and Pratt, 1998) and more recently reconsidered in (Argyriou et al., 2008) from the perspective of convex optimization and sparsity regularization. Representation learning is also a key problem in AI, and in the past years there has been much renewed interest in learning nonlinear hierarchical representations from multiple tasks using multilayer, deep networks. Researchers have shown improved results in a number of empirical domains; the case of computer vision is perhaps most remarkable, (see e.g. Girshick et al., 2014, and references therein). This success has increased interest in multitask representation learning (MTRL) as it is a core component of deep networks. Still, the understanding of why this methodology works remains largely unexplored.

In this paper we analyze a general method for MTRL and discuss its potential advantage in both the MTL setting, where the learned representation is applied to the same tasks used during training, and in the domain of LTL, where the representation is applied to new tasks. We derive upper bounds on the error of these methods and quantify their advantage over independent task learning. When the original data representation is high dimensional and the number of examples provided to solve a regression or classification problem is limited, any learning algorithm which does not use any sort of prior knowledge will perform poorly because there is not enough data to reliably estimate the model parameters. We make this statement precise by considering the example of half space learning.

1.1 Previous Work

Many papers have proposed multitask learning methods and studied their applications to specific problems (see Ando and Zhang, 2005; Argyriou et al., 2008; Baxter, 2000; Ben-David and Schuller, 2003; Caruana, 1998; Cavallanti et al., 2010; Kuzborskij and Orabona, 2013; Maurer et al., 2013; Pentina and Lampert, 2014; Widmer et al., 2013, and references therein). There is a vast literature on these subjects and the list of papers provided here is necessarily incomplete.

Despite the considerable success of multitask learning and in particular multitask representation learning there are only few theoretical investigations (Ando and Zhang, 2005; Baxter, 2000; Ben-David and Schuller, 2003). Other statistical learning bounds are restricted to linear multitask learning such as (Cavallanti et al., 2010; Lounici et al., 2011; Maurer, 2006a,a).

Learning-to-learn (also called inductive bias learning or transfer learning) has been proposed by Thrun and Pratt (1998) and theoretically studied by Baxter (2000) where an error analysis is provided, showing that a common representation which performs well on the training tasks will also generalize to new tasks obtained from the same "environment". More recent papers which present dimension independent bounds appear in Maurer (2006a,b); Maurer and Pontil (2013); Maurer et al. (2013); Pentina and Lampert (2014).

1.2 Our Contributions

There are two main contributions of this work. First we present bounds to both the MTL and LTL settings, which apply to a very general MTRL method. Our analysis goes well beyond linear representation learning considered in most previous works. It improves over the analysis by Baxter (2000) based on covering numbers. We use more recent techniques of empirical process theory to achieve bounds which are independent of the input dimension (hence also valid in reproducing kernel Hilbert spaces) and to avoid logarithmic factors. Furthermore our analysis can be made fully data dependent. When specialized to subspace learning (i.e. linear feature learning) we get best bounds valid for infinite dimensional input spaces.

As the second main contribution of this paper, we explain the advantage of MTRL in terms of specificity of feature maps and expose conditions when MTRL is beneficial or when it is not worth the effort. We further specialize our upper bounds to half-space learning (noiseless binary classification) and compare them to a general lower bound for learning isolated tasks. We observe that if the number of tasks grows then the performance of the method (both in the MTL and LTL setting) matches the performance of square norm regularization with best a priori known representation. This analysis highlights the advantage of multitask learning over learning the tasks independently. We also present numerical experiments for half-space learning, which indicate the good agreement between theory and experiments.

1.3 Organization

The paper is organized as follows. In Section 2, we introduce the problem and present our main results. In Section 3, we specialize these results to subspace learning and illustrate the role played by the data covariance matrices in our bounds. In Section 3.1 we further illustrate our results in the case of half-space learning, rigorously comparing our upper bounds to a general lower bound for orthogonal equivariant algorithms. In Section 4, we present the proof of our main results, developing in particular uniform bounds on the estimation error. Finally, in Section 5 we summarize our findings and suggest directions for future research.

2. Multitask Representation Learning

The set of possible observations is denoted by $\mathcal{Z} = (\mathcal{X}, \mathbb{R})$, where the members of $\mathcal{X}$ are interpreted as inputs and the members of $\mathbb{R}$ are interpreted as outputs, or labels. A learning task is modelled by a probability measure μ on $\mathcal{Z}$ where $\mu(x, y)$ is the probability to encounter the input-output pair $(x, y) \in \mathcal{Z}$ in the context of task μ. We want to learn how to predict outputs. If we predict y while the true output is y', we suffer a loss $\ell(y, y')$, where the loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is assumed to be 1-Lipschitz in the first argument for every value of the second argument. Different Lipschitz constants can be absorbed in the scaling of the predictors and different ranges than $[0, 1]$ can be handled by a simple scaling of our results.

If g is a real function defined on $\mathcal{X}$, then the values $g(x)$ can be interpreted as predictors and the expectation $\mathbb{E}_{(X,Y)\sim\mu}[\ell(g(X),Y)]$ is the risk associated with hypothesis g on the task μ.

Multitask learning simultaneously considers many tasks $\mu_1,\ldots,\mu_T$ and hopes to exploit some suspected common property of these tasks. For the purpose of this paper this property is the existence of a representation or common feature-map, which simultaneously simplifies the learning problem for most, or all of the tasks at hand. We consider predictors g which factorize

$$g = f \circ h,$$

where "$\circ$" stands for functional composition, that is, $(f\circ h)(x) = f(h(x))$, for every $x\in\mathcal{X}$. The function $h:\mathcal{X}\to\mathbb{R}^K$ is called the representation, or feature-map, and it is used across different tasks, while f is a function defined on $\mathbb{R}^K$, a predictor specialized to the task at hand. In the sequel K will always be the dimension of the representation space.

As usual in learning theory the functions $h:\mathcal{X}\to\mathbb{R}^K$ and $f:\mathbb{R}^K\to\mathbb{R}$ are chosen from respective hypothesis classes $\mathcal{H}$ and $\mathcal{F}$, which we refer to as the class of representations and the class of specialized predictors, respectively. These classes can be quite general, but we require that the functions in $\mathcal{F}$ have Lipschitz constant at most L, for some positive real number L.

The choice of representation and specialized predictors is based on the data observed for all the tasks. This data takes the form of a multi-sample $\bar{\mathbf{Z}} = (\mathbf{Z}_1,\ldots,\mathbf{Z}_T)$, with $\mathbf{Z}_t = (Z_{t1},\ldots,Z_{tn}) \sim \mu_t^n$. Here and in the sequel an exponent on a measure indicates a product measure, so that μ_t^n is a measure on $\mathcal{Z}^n$ and $\mathbf{Z}_t$ is an iid sample of n random variables distributed as μ_t. We also write $Z_{ti} = (X_{ti}, Y_{ti})$, $\mathbf{Z}_t = (\mathbf{X}_t, \mathbf{Y}_t)$ and $\bar{\mathbf{Z}} = \left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right)$.

Multitask representation learning (MTRL) solves the optimization problem

$$\min\left\{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell\left(f_t\left(h\left(X_{ti}\right)\right),Y_{ti}\right) : h\in\mathcal{H},\ (f_1,\ldots,f_T)\in\mathcal{F}^T\right\}. \tag{1}$$

In this paper, we are not concerned with the algorithmics of this problem, but rather with the statistical properties of its solutions $\hat{h}$ and $\hat{f}_1,\ldots,\hat{f}_T$. Note that these are functional random variables in their dependence on $\bar{\mathbf{Z}}$.

We consider two possible applications of these solutions. One application, which we will refer to as multitask learning (MTL), retains both the representation $\hat{h}$ and the specializations $\hat{f}_1,\ldots,\hat{f}_T$ to be applied to the tasks at hand. The other, perhaps more important, application assumes that the tasks μ_t are related by a probabilistic law, called an *environment*, and keeps only the representation $\hat{h}$ to be used when specializing to new tasks obeying the same law. In this way the parametrization of a learning algorithm is learned, hence the name "learning-to-learn" (LTL).

We will give general statistical guarantees in both cases. Our bounds consist of three terms. The first term can be interpreted as the cost of estimating the representation h and decreases with the number T of tasks available for training. The second term corresponds to the cost of estimating task-specific predictors and decreases with the number n of training examples available for each task. The last term contains the confidence parameter and typically makes only a very small contribution.

It is not surprising that the complexity of the representation class $\mathcal{H}$ (first term in the bounds) plays a central role. We measure this complexity on the observed input data $\bar{\mathbf{X}} \in \mathcal{X}^{Tn}$. Define a random set $\mathcal{H}\left(\bar{\mathbf{X}}\right) \subseteq \mathbb{R}^{KTn}$ by

$$\mathcal{H}\left(\bar{\mathbf{X}}\right) = \left\{\left(h_k\left(X_{ti}\right)\right) : h \in \mathcal{H}\right\}.$$

The complexity measure relevant to estimation of the representation is the Gaussian average

$$G\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right) = \mathbb{E}\left[\sup_{h\in\mathcal{H}} \sum_{kti} \gamma_{kti} h_k\left(X_{ti}\right) | X_{ti}\right], \tag{2}$$

where the γ_{kti} are independent standard normal variables. The Gaussian average is of order $\sqrt{nT}$ in T and n for many classes of interest. These include kernel machines with Lipschitz kernels (e.g. Gaussian RBF) and arbitrarily deep compositions thereof, see Maurer (2014) for a discussion. As we shall see, this increase of $O(\sqrt{nT})$ is compensated in our bounds and the cost of learning the representation vanishes in the multi-task limit $T \to \infty$.

The second term in the bounds is governed by the quantity

$$\sup_{h\in\mathcal{H}} \frac{1}{n\sqrt{T}} \left\|h\left(\bar{\mathbf{X}}\right)\right\| = \frac{1}{\sqrt{n}} \sup_{h\in\mathcal{H}} \sqrt{\frac{1}{nT} \sum_{kti} h_k\left(X_{ti}\right)^2} \tag{3}$$

or an equivalent distribution-dependent expression. If the feature-maps in $\mathcal{H}$ are very specific, in the sense that their components are appreciably different from zero only for very special data, the quantity in (3) can become much smaller than $1/\sqrt{n}$, a phenomenon which can give a considerable competitive edge to MTRL, in particular if the per-task sample size n is small. We will demonstrate this in Section 3, where we apply Theorems 1 and 2 to subspace-learning and show that the above quantity is related to the operator norm of the data covariance.

2.1 Bounding the Excess Task-averaged Risk (MTL)

If we make no further assumptions on the generation of the task-measures $\mu_1, \ldots, \mu_T$, a conceptually simple performance measure for a representation h and specialized predictors $f_1, \ldots, f_T$ is the task-averaged risk

$$\mathcal{E}_{\text{avg}}\left(h, f_1, \ldots, f_T\right) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(X,Y)\sim\mu_t} \ell\left(f_t\left(h\left(X\right)\right), Y\right).$$

We want to compare this to the very best we can do using the classes $\mathcal{H}$ and $\mathcal{F}$, given complete knowledge of the distributions $\mu_1, \ldots, \mu_T$. The minimal risk is clearly

$$\mathcal{E}^*_{\text{avg}} = \min_{h\in\mathcal{H}, (f_1,\ldots,f_T)\in\mathcal{F}^T} \mathcal{E}_{\text{avg}}\left(h, f_1, \ldots, f_T\right).$$

It is a fundamental hope underlying our approach that the classes $\mathcal{H}$ and $\mathcal{F}$ are large enough for this quantity to be sufficiently small for practical purposes. We use the words "hope" and "belief" because an "assumption" would imply a statement to be used in analytical

reasoning. Instead our approach is agnostic, and our results are valid independent of the size of the minimal risk above.

Our first result bounds the excess average risk, which measures the difference between the task-averaged true risk of the solutions to (1) and the theoretical optimum above.

Theorem 1 *Let $\mu_1, \ldots, \mu_T$, $\mathcal{H}$ and $\mathcal{F}$ be as above, and assume $0 \in \mathcal{H}$ and $f(0) = 0$ for all $f \in \mathcal{F}$. Then for $\delta > 0$ with probability at least $1 - \delta$ in the draw of $\bar{\mathbf{Z}} \sim \prod_{t=1}^{T} \mu_t^n$ we have that*

$$\mathcal{E}_{\mathrm{avg}}(\hat{h}, \hat{f}_1, \ldots, \hat{f}_T) - \mathcal{E}^*_{\mathrm{avg}} \leq \frac{c_1 L \; G\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right)}{nT} + \frac{c_2 Q \sup_{h \in \mathcal{H}} \left\|h\left(\bar{\mathbf{X}}\right)\right\|}{n\sqrt{T}} + \sqrt{\frac{8 \ln\left(4/\delta\right)}{nT}},$$

where c_1 and c_2 are universal constants, $G(\mathcal{H}(\bar{\mathbf{X}}))$ is the Gaussian average in Equation (2), and Q is the quantity

$$Q \equiv Q(\mathcal{F}) \sup_{y \neq y' \in \mathbb{R}^{Kn}} \frac{1}{\|y - y'\|} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \gamma_i \left(f\left(y_i\right) - f\left(y_i'\right)\right). \tag{4}$$

Remarks:

1. The assumptions $0 \in \mathcal{H}$ and $f(0) = 0$ for all $f \in \mathcal{F}$ are made to give the result a simpler appearance. They are not essential, as the reader can verify from the proof.
2. If $G\left(\mathcal{H}\left(\bar{\mathbf{x}}\right)\right)$ is of order $\sqrt{nT}$ then the first term on the right hand side above is of order $1/\sqrt{Tn}$ and vanishes in the *multi-task limit* $T \to \infty$ even for small values of n.
3. For reasonable classes $\mathcal{F}$ one can find a bound on Q, which is independent of n, because the $\|y - y'\|$ in the denominator balances the Gaussian average depending on the class $\mathcal{F}$.
4. The quantity $\sup_h \left\|h\left(\bar{\mathbf{X}}\right)\right\|$ is of order $\sqrt{nT}$ whenever $\mathcal{H}$ is uniformly bounded, a crude bound being $\sqrt{nT} \sup_{h \in \mathcal{H}} \max_{ti} \|h\left(x_{ti}\right)\|$. The second term is thus typically of order $1/\sqrt{n}$. As explained in the discussion of Equation (3) above it can be very small if the representation components in $\mathcal{H}$ are very data-specific.

2.2 Bounding the Excess Risk for Learning-to-learn (LTL)

Now we consider the case where we only retain the representation $\hat{h}$ obtained from (1) and specialize it to future, hitherto unknown tasks. This is of course only possible, if there is some common law underlying the generation of tasks. Following Baxter (2000) we suppose that the tasks originate in a common environment η, which is by definition a probability measure on the set of probability measures on $\mathcal{Z}$. The draw of $\mu \sim \eta$ models the encounter of a learning task μ in the environment η.

The environment η induces a measure μ_η on $\mathcal{Z}$ by

$$\mu_\eta\left(A\right) = \mathbb{E}_{\mu \sim \eta}\left[\mu\left(A\right)\right] \text{ for } A \subseteq \mathcal{Z}.$$

This simple mixture plays an important role in the interpretation of our results.

The measure η also induces a measure ρ_η on $\mathcal{Z}^n$ which corresponds to the draw of an n-sample from a random task in the environment. To draw a sample $\mathbf{Z} \in \mathcal{Z}^n$ from ρ_η we first draw a task μ from η and then generate the sample $\mathbf{Z} = (Z_1, \ldots, Z_T)$ from n independent draws from μ. Formally

$$\rho_\eta(A) = \mathbb{E}_{\mu\sim\eta}\left[\mu^n(A)\right] \text{ for } A \subseteq \mathcal{Z}^n.$$

We assume that the tasks $\mu_1, \ldots, \mu_T$ are drawn independently from η and, consequently, that the multisample $\bar{\mathbf{Z}} = (\mathbf{Z}_1, \ldots, \mathbf{T}_T)$ is obtained in T independent draws from ρ_η, that is, $\bar{\mathbf{Z}} \sim \rho_\eta^T$.

The way we plan to use a representation $h \in \mathcal{H}$ on a new task $\mu \sim \eta$ is as follows: we draw a training sample $\mathbf{Z} = (Z_1, \ldots, Z_n)$ from μ^n and solve the optimization problem

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(f\left(h\left(X_i\right)\right), Y_i\right).$$

Let $\hat{f}_{h,\mathbf{Z}}$ denote the minimizer and $m_{h,\mathbf{Z}}$ the corresponding minimum. We will then use the hypothesis $a(h)_\mathbf{Z} = \hat{f}_{h,\mathbf{Z}} \circ h = \hat{f}_{h,\mathbf{Z}}(h(\cdot))$ for the new task. In this way any representation $h \in \mathcal{H}$ parametrizes a learning algorithm, which is a function $a(h) : \mathcal{Z}^n \to \mathcal{F} \circ h$, defined, for every $\mathbf{Z} \in \mathcal{Z}^n$, as

$$a(h)_\mathbf{Z} = \hat{f}_{h,\mathbf{Z}} \circ h.$$

In this sense the problem of optimizing such a representation can properly be called "learning-to-learn". It can also be interpreted as "learning a hypothesis space" as in (Baxter, 2000), namely selecting a hypothesis space $\mathcal{F} \circ h$ from the collection of hypothesis spaces $\{\mathcal{F} \circ h : h \in \mathcal{H}\}$.

We can test the algorithm $a(h)$ on the environment η in the following way:

- we draw a task $\mu \sim \eta$,
- we draw a sample $\mathbf{Z} \in \mathcal{Z}^n$ from μ^n,
- we run the algorithm to obtain $a(h)_\mathbf{Z} = \hat{f}_{h,\mathbf{Z}} \circ h$,
- finally, we measure the loss of $a(h)_\mathbf{Z}$ on a random data-point $Z = (X, Y) \sim \mu$.

To define the risk $\mathcal{E}_\eta(h)$ associated with the algorithm $a(h)$ parametrized by h we just replace all random draws with corresponding expectations, so

$$\mathcal{E}_\eta(h) = \mathbb{E}_{\mu\sim\eta}\mathbb{E}_{\mathbf{Z}\sim\mu^n}\mathbb{E}_{(X,Y)\sim\mu}\left[\ell\left(a(h)_\mathbf{Z}(X), Y\right)\right].$$

The best value for any representation h in $a(h)$, given complete knowledge of the environment, is then

$$\min_{h\in\mathcal{H}} \mathcal{E}_\eta(h).$$

But, given complete knowledge of the environment, this is still not the best we can do using the classes $\mathcal{F}$ and $\mathcal{H}$, because for given μ and h we still use the expected performance

$\mathbb{E}_{\mathbf{Z}\sim\mu^n}\mathbb{E}_{Z\sim\mu}\,\ell\left(a\left(h\right)_{\mathbf{Z}}\left(X\right),Y\right)$ of the empirical risk minimization algorithm $a\left(h\right)$, instead of using knowledge of μ to replace it by $\min_{f\in\mathcal{F}}\mathbb{E}_{Z\sim\mu}\ell\left(f\left(h\left(X\right)\right),Y\right)$. The very best we can do is thus

$$\mathcal{E}_\eta^* = \min_{h\in\mathcal{H}}\mathbb{E}_{\mu\sim\eta}\left[\min_{f\in\mathcal{F}}\mathbb{E}_{Z\sim\mu}\ell\left(f\left(h\left(X\right)\right),Y\right)\right].$$

The excess risk associated with any representation h is thus

$$\mathcal{E}_\eta\left(h\right) - \mathcal{E}_\eta^*.$$

We give the following bound for the excess risk associated with the representation $\hat{h}$ found as solution to the optimization problem (1).

Theorem 2 *Let η be an environment on $\mathcal{Z}$ and $\mathcal{H}$ and $\mathcal{F}$ as above. Then: (i) with probability at least $1-\delta$ in the draw of $\bar{\mathbf{Z}}\sim\rho_\eta^T$*

$$\mathcal{E}_\eta(\hat{h}) - \mathcal{E}_\eta^* \leq \frac{\sqrt{2\pi}L\;G\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right)}{T\sqrt{n}} + \sqrt{2\pi}Q'\sup_{h\in\mathcal{H}}\sqrt{\frac{\mathbb{E}_{(X,Y)\sim\mu_\eta}\left[\left\|h\left(X\right)\right\|^2\right]}{n}} + \sqrt{\frac{8\ln\left(4/\delta\right)}{T}},$$

and (ii) with the same probability

$$\mathcal{E}_\eta(\hat{h}) - \mathcal{E}_\eta^* \leq \frac{\sqrt{2\pi}L\;G\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right)}{T\sqrt{n}} + \frac{\sqrt{2\pi}Q'\left(1/T\right)\sum_t\sup_{h\in\mathcal{H}}\left\|h\left(\mathbf{X}_t\right)\right\|}{n} + 5\sqrt{\frac{\ln\left(8/\delta\right)}{T}},$$

where $\hat{h}$ is solution to the problem (1), $G\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right)$ is the Gaussian average introduced in (2), and Q' is the quantity

$$Q' \equiv Q'(\mathcal{F}) = \sup_{y\in\mathbb{R}^{Kn}\setminus\{0\}}\frac{1}{\|y\|}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n\gamma_i f\left(y_i\right). \tag{5}$$

We make some remarks and comparison to the previous result.

1. The constants are now explicit and small. For Theorem 1, uniform estimation had to be controlled simultaneously in $\mathcal{H}$ and $\mathcal{F}$, while for LTL the problem can be more easily decoupled.

2. The first term is equivalent to the first term in Theorem 1 except for $\sqrt{n}$ replacing n in the denominator. It is therefore typically of order $1/\sqrt{T}$ instead of $1/\sqrt{nT}$. The different order is due to the estimation of a hitherto unknown task, for which the sample sizes are irrelevant. To understand this point assume the η has the property that every $\mu\sim\eta$ is deterministic, that is supported on a single point $z_\mu\in\mathcal{Z}$. Then clearly the sample size n is irrelevant, and the problem becomes equivalent to learning a single task with a sample of size T.

3. The quantity Q' is very much like the quantity Q in Equation (4), and it is uniformly bounded in n for the classes we consider. For linear classes $Q = Q'$.

4. The bound in part (i) is not fully data-dependent, but more convenient for our applications below. The quantity

$$\sup_{h\in\mathcal{H}} \sqrt{\mathbb{E}_{(X,Y)\sim\mu_\eta} \|h(X)\|^2} = \sup_{h\in\mathcal{H}} \sqrt{\sum_k \mathbb{E}_{(X,Y)\sim\mu_\eta} \left[h_k(X)^2\right]}$$

plays a similar role to (3), which is its empirical counterpart. Again, if the features h_k are very specific, as the dictionary atoms of the next section or the atoms in a radial basis function network, then the above quantity can become very small.

2.3 Comparison to Previous Bounds

The first and most important theoretical study of MTL and LTL was carried out by Baxter (2000), where sample complexity bounds are given for both settings. Instead of a feature map a hypothesis space is selected from a class of hypothesis spaces. Clearly every feature map with values in $\mathbb{R}^K$ defines a hypothesis space while the reverse is not true in general, so Baxter's setting is certainly more general than ours. On the other hand the practical applications discussed in (Baxter, 2000) can be cast in the language of feature learning.

To prove his sample complexity bounds Baxter uses covering numbers. This classical method requires to cover a (meta-)hypothesis space (or its evaluation on a sample) with a set of balls in an appropriately chosen metric. The uniform bound is then obtained as a union bound over the cover and bounds valid on the individual balls. The latter bounds follow from Lipschitz properties L of the loss function relative to the chosen metric. For a bound of order ϵ the radius of the balls has to be of order ϵ/L. This leads to covering numbers of order ϵ^{-d}, where d is some exponent (see the last inequalities in the proof of in (Baxter, 2000), and has the consequence that the dominant term in the bound has an additional factor of $\ln(1/\epsilon)$. This is manifest in Theorem 8, Theorem 12 and Corollary 13 in (Baxter, 2000) and constitutes an essential weakness of the method of covering numbers. For bounds on the excess risk it implies that the orders of $\sqrt{1/T}$ and $\sqrt{1/n}$ obtained from Rademacher or Gaussian complexities have to be replaced by $\sqrt{\ln(T)/T}$ and $\sqrt{\ln(n)/n}$.

Rademacher and Gaussian complexities make it easy to handle infinite dimensional input spaces (see our Theorems 4 and 5 below). They also lead to data dependent bounds, which allows us to explain the benefits of multi-task learning in terms of the spectrum of the data covariance operator and the effective input dimension. Bounding Gaussian complexities for linear classes is comparatively simple, see the proof of our Lemma 3. There is a wealth of recent literature on the Rademacher complexity of matrices with spectral regularizers (see e.g. Kakade et al., 2012; Maurer and Pontil, 2013, and references therein), while it is unclear to us how Baxter's method could be applied if the feature map is constrained by a bound on, say, the trace norm of the associated matrix. In the case of LTL, our approach also leads to explicit and small constant factors.

On the other hand it must be admitted, that it is relatively easy to obtain bounds (also provided by Baxter) of order $\ln(n)/n$ or $\ln(T)/T$ with covering numbers in the realizable case. Such bounds would be more difficult to obtain with our techniques.

The work of Ando and Zhang (2005) proposes the use of MTL as a method of semi-supervised learning through the creation of artificial tasks from unlabelled data, for example predicting concealed components of vectors. They analyze a specific algorithm where the class of feature maps can be seen as a linear mixture of a fixed feature map with subspace projections as discussed in our paper. The bounds given apply to the task-averaged risk and not to LTL. The analysis is based on Rademacher averages and is independent of the input dimension. The bound itself is expressed as an entropy integral as given by Dudley (see e.g. Van Der Vaart and Wellner, 1996) but it is not very explicit. In particular the role of the spectrum of the data covariance is not apparent.

3. Multi-task Subspace Learning

We illustrate the general results of the previous section with an important special case. We assume that the input space $\mathcal{X}$ is a bounded subset of a Hilbert space H, which could for example be a reproducing kernel Hilbert space. We denote by $\langle \cdot, \cdot \rangle$ the inner product in H and by $\| \cdot \|$ the induced norm. We hope that sufficiently good results can be obtained by predictors of the form g, where $g : H \to \mathbb{R}$ is linear with bounded norm. We also suspect that only few linear features in H suffice for most tasks, so that the vectors defining the hypotheses g can all be chosen from one and the same, albeit unknown, K-dimensional subspace M of H.

Consequently we will factorize predictors as $f \circ h$, where h is a partial isometry $h : H \to \mathbb{R}^K$ and f is a linear functional on $\mathbb{R}^K$ chosen from some ball of bounded radius. Specifically, we introduce the classes

$$\begin{aligned} \mathcal{H} &= \left\{ H \ni x \mapsto \left(\langle d_1, x \rangle, \ldots, \langle d_K, x \rangle \right) \in \mathbb{R}^K \ : \ D = (d_1, \ldots, d_K) \in H^K \text{ orthonormal} \right\} \\ \mathcal{F} &= \left\{ \mathbb{R}^K \ni y \mapsto \sum_k w_k y_k \in \mathbb{R} \ : \ \sum_k w_k^2 \leq B^2 \right\}. \end{aligned}$$

The D's appearing in the definition of $\mathcal{H}$ are also called dictionaries and the individual d_k are called atoms (see Maurer et al., 2013).

It does no harm to our analysis if we immediately generalize the class $\mathcal{H}$ so as to include certain two-layer neural networks by allowing a nonlinear activation function ϕ with Lipschitz constant L_ϕ and satisfying $\phi(0) = 0$, to be applied with each atom. We can also drop the condition of orthonormality and allow the atoms to trade some of their norms when needed. The enlarged class of representations is

$$\mathcal{H} = \left\{ x \in H \mapsto \left(\phi\left(\langle d_1, x \rangle \right), \ldots, \phi\left(\langle d_K, x \rangle \right) \right) \in \mathbb{R}^K : d_1, \ldots, d_K \in H, \ \sum_k \|d_k\|^2 \leq K \right\}.$$

The results can then be re-specialized to subspace learning by setting ϕ to the identity and L_ϕ to one.

When applied to subspace learning, our bounds are expressed in terms of covariances. If ν is a probability measure on H the corresponding covariance operator C_ν is defined by

$$\langle C_\nu v, w \rangle = \mathbb{E}_{X \sim \nu} \langle v, X \rangle \langle X, w \rangle \ \text{ for } v, w \in H.$$

For an environment η we denote the covariance operator corresponding to the data-marginal of the mixture measure μ_η simply by C.

If $\mathbf{x} = (x_1, \ldots, x_m) \in H^m$ we define the empirical covariance operator $\hat{C}(\mathbf{x})$ by

$$\left\langle \hat{C}(\mathbf{x})\, v, w \right\rangle = \frac{1}{m} \sum_i \langle v, x_i \rangle \langle x_i, w \rangle \text{ for } v, w \in H,$$

in particular

$$\left\langle \hat{C}\left(\bar{\mathbf{X}}\right) v, w \right\rangle = \frac{1}{nT} \sum_{ti} \langle v, X_{ti} \rangle \langle X_{ti}, w \rangle .$$

The following lemma establishes the necessary ingredients for the application of Theorems 1 and 2 to the case of subspace learning. Recall that if A is a selfadjoint positive linear operator on H, we denote by $\|A\|_\infty$ and $\|A\|_1$ its spectral and trace norms, respectively. They are defined as $\|A\|_\infty = \sup_{\|z\|\leq 1} \|Az\|$ and $\|A\|_1 = \sum_{i\in\mathbb{N}} \langle e_i, Ae_i \rangle$, where $\{e_i\}_{i\in\mathbb{N}}$ is an orthonormal basis in H. Recall also the definition of $Q(\mathcal{F})$ and $Q'(\mathcal{F})$ given in Equations (4) and (5), respectively.

Lemma 3 *Let $\bar{\mathbf{x}} = (x_{ti})$ be a $T \times n$ matrix with values in a Hilbert space and let ϕ, $\mathcal{H}$ and $\mathcal{F}$ be defined as above. Then*

(i) $G(\mathcal{H}(\bar{\mathbf{x}})) \leq L_\phi K \sqrt{nT \left\| \hat{C}(\bar{\mathbf{x}}) \right\|_1}$.

(ii) For every $h \in \mathcal{H}$, $\|h(\bar{\mathbf{x}})\| \leq L_\phi \sqrt{KnT \left\| \hat{C}(\bar{\mathbf{x}}) \right\|_\infty}$.

(iii) For an environment η *and every* $h \in \mathcal{H}$

$$\mathbb{E}_{(X,Y)\sim\mu_\eta} \|h(X)\|^2 \leq L_\phi^2 K \|C\|_\infty .$$

(iv) $L(\mathcal{F}) \leq B$.

(v) $Q(\mathcal{F}) \leq B$ *and* $Q'(\mathcal{F}) \leq B$.

Proof *(i)* Using the contraction lemma, Corollary 11, in the first inequality and Cauchy-Schwarz and Jensen's inequality in the second we get

$$\begin{aligned}
G(\mathcal{H}(\bar{\mathbf{x}})) &\leq L_\phi \mathbb{E} \sup_{d\in\mathcal{H}} \sum_{kti} \gamma_{kti} \langle d_k, x_{ti} \rangle \\
&= L_\phi \mathbb{E} \sup_{d\in\mathcal{H}} \sum_k \left\langle d_k, \sum_{ti} \gamma_{kti} x_{ti} \right\rangle \\
&\leq L_\phi \sqrt{K} \left(\sum_k \mathbb{E} \left\| \sum_{ti} \gamma_{ti} x_{ti} \right\|^2 \right)^{1/2} \\
&\leq L_\phi K \left(\sum_{ti} \|x_{ti}\|^2 \right)^{1/2} = L_\phi K \sqrt{nT \left\| \hat{C}(\mathbf{x}) \right\|_1}.
\end{aligned}$$

(ii) For any $D \in \mathcal{H}$

$$\begin{aligned}\sum_{kti} \phi\left(\langle d_k, x_{ti}\rangle\right)^2 &\leq L_\phi^2 \sum_{kti} \langle d_k, x_{ti}\rangle^2 \\ &= L_\phi^2 \sum_k \|d_k\|^2 \sum_{ti} \left\langle \frac{d_k}{\|d_k\|}, x_{ti} \right\rangle^2 \\ &\leq L_\phi^2 K \sup_{v:\|v\|\leq 1} \sum_{ti} \langle v, x_{ti}\rangle^2 \\ &= L_\phi^2 K n T \left\| \hat{C}\left(\bar{\mathbf{x}}\right)\right\|_\infty,\end{aligned}$$

where we used $\phi(0) = 0$ in the first step.
(iii) Similarly, we have that

$$\begin{aligned}\mathbb{E}_{(X,Y)\sim\mu_\eta} \sum_k \phi\left(\langle d_k, X\rangle\right)^2 &\leq L_\phi^2 \sum_k \|d_k\|^2 \mathbb{E}_{(X,Y)\sim\mu_\eta} \left\langle \frac{d_k}{\|d_k\|}, X \right\rangle^2 \\ &\leq L_\phi^2 K \sup_{\|v\|\leq 1} \mathbb{E}_{(X,Y)\sim\mu_\eta} \langle v, X\rangle^2 \\ &= L_\phi^2 K \|C\|_\infty.\end{aligned}$$

(iv) Let $y, y' \in \mathbb{R}^K$. Then

$$\sup_{w\in\mathcal{F}} \left\{ \sum_k w_k y_k - \sum_k w_k y'_k \right\} \leq \left(\sum_k w_k^2 \right)^{1/2} \left\| y - y' \right\| \leq B \left\| y - y' \right\|,$$

so $L \leq B$.
(v) Similarly, we have that

$$\begin{aligned}\mathbb{E} \sup_{w\in\mathcal{F}} \sum_i \gamma_i \left(\sum_k w_k y_{ki} - \sum_k w_k y_{ki} \right) &= \mathbb{E} \sup_{w\in\mathcal{F}} \sum_k w_k \sum_i \gamma_i \left(y_{ki} - y_{ki}\right) \\ &\leq \sup_{w\in\mathcal{F}} \sqrt{\sum_k w_k^2 \sum_k \mathbb{E} \left(\sum_i \gamma_i \left(y_{ki} - y_{ki}\right) \right)^2} \\ &\leq B \sqrt{\sum_{ki} \left(y_{ki} - y_{ki}\right)^2} = B \left\| y - y' \right\|,\end{aligned}$$

so $Q \leq B$. The same proof works for Q'. ■

Substitution in Theorem 1 immediately gives

Theorem 4 (subspace MTL) *With probability at least $1 - \delta$ in $\bar{\mathbf{X}}$ the excess risk is bounded by*

$$\mathcal{E}_{\text{avg}}(\hat{h}, \hat{f}_1, \ldots, \hat{f}_T) - \mathcal{E}^*_{\text{avg}} \leq c_1 L_\phi B K \sqrt{\frac{\left\| \hat{C}\left(\bar{\mathbf{X}}\right)\right\|_1}{nT}} + c_2 L_\phi B \sqrt{\frac{K \left\| \hat{C}\left(\bar{\mathbf{X}}\right)\right\|_\infty}{n}} + \sqrt{\frac{8 \ln\left(2/\delta\right)}{nT}}. \tag{6}$$

We remark that in the linear case the best competing bound for MTL, obtained by Maurer and Pontil (2013) from noncommutative Bernstein inequalities, is

$$2B\sqrt{\frac{K\left\|\hat{C}\left(\bar{\mathbf{X}}\right)\right\|_1 \ln\left(Tn\right)}{nT}} + B\sqrt{\frac{8K\left\|\hat{C}\left(\bar{\mathbf{X}}\right)\right\|_\infty}{n}} + \sqrt{\frac{8\ln\left(2/\delta\right)}{nT}}. \tag{7}$$

If we disregard the constants this is worse than the bound (6) whenever $K < \ln(Tn)$. Its approach to the multitask limit is slower ($\sqrt{\ln(T)/T}$ as opposed to $\sqrt{1/T}$), but of course it has the advantage of smaller constants. The methods used to obtain (7), however, break down for nonlinear dictionaries.

For the LTL setting, we use the distribution dependent bound, Theorem 2 (i), and obtain

Theorem 5 (subspace LTL) *With probability at least $1-\delta$ in $\bar{\mathbf{X}}$, the excess risk is bounded by*

$$\mathcal{E}_\eta(\hat{h}) - \mathcal{E}_\eta^* \leq \sqrt{2\pi} L_\phi B \left(\frac{K\sqrt{\left\|\hat{C}\left(\bar{\mathbf{X}}\right)\right\|_1}}{\sqrt{T}} + \sqrt{\frac{K\left\|C\right\|_\infty}{n}} \right) + \sqrt{\frac{8\ln\left(4/\delta\right)}{T}}.$$

The two most important common features of Theorems 4 and 5 are the decay to zero of the first term, as $T \to \infty$, and the occurrence of the operator norm of the empirical or true covariances in the second term. The first implies that for very large numbers of tasks the bounds are dominated by the second term.

To understand the second term we must first realize that the ratio of trace and operator norms of the true covariances can be interpreted as an effective dimension of the distribution. This is easily seen if the mixture of task-marginals is concentrated and uniform on a d-dimensional unit-sphere. In this case $\|C\|_1 = 1$ and by isotropy all eigenvalues are equal, so $\|C\|_\infty = 1/d$, whence $\|C\|_1 / \|C\|_\infty = d$. In such a case the second term in Theorem 5 above becomes

$$B\sqrt{\frac{K}{dn}}. \tag{8}$$

The appropriate standard bound for learning the tasks independently would be $B\sqrt{1/n}$ (see Bartlett and Mendelson, 2002). The ratio $\sqrt{K/d}$ of the two bounds in the multitask limit is the quotient of utilized information (the dimension of the representation space) to available information (the dimension of the data). This highlights the potential advantages of MTRL: if the data is already low-dimensional in the order of K then multi-task learning isn't worth the extra computational labour. If the data is high dimensional however, then multi-task learning may be superior.

The expression (8) above might suggest that there really is a benefit of high dimensions for learning-to-learn. This is of course not the case, because the regularizer B has to be chosen large, in fact proportional to $\sqrt{d}$ to allow a small empirical error. The correct interpretation of (8) is that the burden of high dimensions vanishes in the limit $T \to \infty$. In the next section we will explain this point in more detail.

3.1 Learning to Learn Half-spaces

In this section, we illustrate the benefit of MTRL over independent task learning (ITL) in the case of noiseless linear binary classification (or half-space learning). We compare our upper bounds for LTL to a general lower bound on the performance of ITL algorithms and quantify the parameter regimes where LTL is superior to ITL.

We assume that all the input marginals are given by the uniform distribution σ on the unit sphere $\mathcal{S}_d$ in $\mathbb{R}^d$, and the objective is for each task μ to classify membership in the half-space $\{x : \langle x, u_\mu \rangle > 0\}$ defined by a task-specific (unknown) unit vector u_μ. In the given environment all the vectors u_μ are assumed to lie in some (unknown) K-dimensional subspace M of $\mathbb{R}^d$. We are interested in the regime that

$$K \ll n \ll d$$

and T grows. This is the safe regime in which our upper bounds for MTL or LTL (cf. Theorems 4 and 5) are smaller than a uniform lower bound for independent task learning, which we discuss below. We need $n \ll d$ for the lower bound to be large and $K \ll n$ for the middle term in our upper bounds to be small. If T is large enough, the second term in our upper bounds dominates the first (task dependent) term. A safe choice is $T \gg K^2 d$, see Equation (9) below.

The 0-1-loss is unsuited for our bounds because it is not Lipschitz. Instead we will use the truncated hinge loss with unit margin given by $\ell(y', y) = \xi(y'y)$, where ξ is the real function

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

This loss is an upper bound of the 0-1-loss, so upper bounds for this loss function are also upper bounds for the classification error.

Let $\mathcal{H}$ and $\mathcal{F}$ be as given at the beginning of Section 3 in its linear variant, where $\mathcal{H}$ is defined by orthonormal dictionaries without activation functions. Thus, $\mathcal{H}$ can be viewed as the set of partial isometries $D : H \to \mathbb{R}^K$.

Recall the definition of the minimal risk for LTL

$$\begin{aligned} \mathcal{E}_\eta^* &= \min_{h \in \mathcal{H}} \mathbb{E}_{\mu \sim \eta} \left[\min_{f \in \mathcal{F}} \mathbb{E}_{Z \sim \mu} \ell\Big(f\big(h(X)\big), Y \Big) \right] \\ &= \min_{D \in \mathcal{H}} \mathbb{E}_{\mu \sim \eta} \left[\min_{\|w\| \leq B} \mathbb{E}_{Z \sim \mu} \xi\big(\langle w, DX \rangle \operatorname{sgn}(\langle u_\mu, X \rangle) \big) \right]. \end{aligned}$$

Let D_M be the partial isometry mapping M onto $\mathbb{R}^K$. Then $D_M \in \mathcal{H}$ and for every unit vector $u \in H$ we have $D_M(Bu) \in \mathcal{F}$. Thus

$$\begin{aligned} \mathcal{E}_\eta^* &\leq \mathbb{E}_{\mu \sim \eta} \left[\mathbb{E}_{Z \sim \mu} \xi\big(\langle D_M(Bu_\mu), D_M X \rangle \operatorname{sgn}(\langle u_\mu, X \rangle) \big) \right] \\ &= \mathbb{E}_{\mu \sim \eta} \left[\mathbb{E}_{X \sim \sigma} \xi\left(B |\langle u_\mu, X \rangle| \right) \right] \\ &\leq \sup_{\|u\| \leq 1} \mathbb{E}_{X \sim \sigma} \xi\left(B |\langle u, X \rangle| \right). \end{aligned}$$

For any unit vector $u \in H$ the density of the distribution of $|\langle u, X \rangle|$ under σ has maximum A_{d-1}/A_d, where A_d is the volume of $\mathcal{S}_d$ in the metric inherited from $\mathbb{R}^d$. This density can

therefore be bounded by $\sqrt{d}/2$. Thus

$$\mathcal{E}_\eta^* \leq \sqrt{d} \int_{-\infty}^{\infty} \xi\left(B\left|s\right|\right) ds = \frac{\sqrt{d}}{2B} = \epsilon,$$

if we set $B = \sqrt{d}/\left(2\epsilon\right)$. This choice is made to ensure that the Lipschitz loss upper bounds the 0-1-loss.

Now let $\bar{\mathbf{Z}}$ be a multi-sample generated from the environment η and assume that we have solved the optimization problem (1) to obtain the representation (or feature-map) $\hat{D} \in \mathcal{H}$. Using the excess risk bound, Theorem 5, and the fact that $\|C\|_\infty = 1/d$ and $\|C\|_1 = 1$, we get with probability at least $1-\delta$ in the draw of $\bar{\mathbf{Z}}$, that

$$\begin{aligned} \mathcal{E}_\eta(\hat{D}) &\leq \epsilon + \frac{\sqrt{2\pi}}{2\epsilon}\left(K\sqrt{\frac{d}{T}} + \sqrt{\frac{K}{n}}\right) + \sqrt{\frac{8\ln\left(4/\delta\right)}{T}} \\ &\leq \sqrt{\sqrt{2\pi}\left(K\sqrt{\frac{d}{T}} + \sqrt{\frac{K}{n}}\right)} + \sqrt{\frac{8\ln\left(4/\delta\right)}{T}}, \end{aligned} \tag{9}$$

if we optimize ϵ. This guarantees the expected performance of future uses of the representation $\hat{D}$. The high dimension still is a hindrance to the estimation of the representation, but, as announced, its effect vanishes in the limit $T \to \infty$. The individual samples must only well outnumber the dimension K, roughly the number of shared features.

We compare this upper bound to a lower bound for a large class of algorithms which learn the tasks independently.

Definition 6 *An algorithm $f : \mathcal{S}_d \times \{-1,1\}^n \to \mathcal{S}_d$ is called* orthogonally equivariant *if*

$$f\left(V\mathbf{x}, \mathbf{y}\right) = Vf\left(\mathbf{x}, \mathbf{y}\right), \text{ for every orthogonal matrix } V \in \mathbb{R}^{d\times d}. \tag{10}$$

For data transformed by an orthogonal transformation an orthogonally equivariant algorithm produces a correspondingly transformed hypothesis. Any algorithm which does not depend on a specific coordinate system is orthogonally equivariant. This class of algorithms includes all kernel methods, but it excludes the Lasso (L1-norm regularization). If the known properties of the problem posses a rotation symmetry only equivariant algorithms make sense.

Below we denote by $\mathrm{err}(u,v)$ the misclassification error between the half-spaces associated with unit vectors u and v, that is $\mathrm{err}(u,v) = \Pr_{x\sim\sigma}\{\langle u,x\rangle\langle v,x\rangle < 0\}$. The following lower error bound is given in (Maurer and Pontil, 2008).

Theorem 7 *Let $n < d$ and suppose that $f : \mathcal{S}_d^n \times \{-1,1\}^n \to \mathcal{S}_d$ is an orthogonally equivariant algorithm. Then for $\delta > 0$ with probability at least $1-\delta$ in the draw of $\mathbf{X} \sim \boldsymbol{\sigma}^n$ we have for every $u \in \mathcal{S}_d$ that*

$$\mathrm{err}\big(u, f(\mathbf{X}, u(\mathbf{X}))\big) \geq \frac{1}{\pi}\left(\sqrt{\frac{d-n}{d}} - \sqrt{\frac{\ln\left(1/\delta\right)}{d}}\right),$$

where $u\left(\mathbf{X}\right) = \left(sgn\left\langle u, X_1\right\rangle, \ldots, sgn\left\langle u, X_n\right\rangle\right)$.

If we use a union bound to subtract the upper bound (9) from this lower bound we obtain high probability guarantees for the advantage of representation learning over other algorithms.

In the following section we plot the phase diagram derived here, namely the difference between the uniform lower bound and our upper bound, and compare it with empirical results (see Figure 4).

3.2 Numerical Experiments

The purpose of the experiments is to compare MTL and LTL to independent task learning (ITL) in the simple setting of linear feature learning (or subspace learning)[1]. We wish to study the regime in which MTL/LTL learning is beneficial over ITL as a function of the number of tasks T and the sample size per task n.

We consider noiseless linear binary classification tasks, namely halfspace learning. We generated the data in the following way. The ground truth weight vectors $u_1, \ldots, u_T$ are obtained by the equation $u_t = Dc_t$, where $c_t \in \mathbb{R}^K$ is sampled from the uniform distribution on the unit sphere in $\mathbb{R}^K$, and the dictionary $D \in \mathbb{R}^{d \times K}$ is created by first sampling a d-dimension orthonormal matrix from the Haar measure, and then selecting the first K columns (atoms). We create all input marginals by sampling from the uniform distribution on the $\sqrt{d}$ radius sphere in $\mathbb{R}^d$. For each task we sample n instances to build the training set, and 1000 instances for the test set.

We train the methods with the hinge loss function $h(z) := \max\{0, 1 - z/c\}$, where c is the margin. We choose $c = 2/\epsilon$, so that the true error relative to the best hypothesis is of order ϵ. We fixed the value of ϵ to be $(K/n)^{1/2}$. For ITL we optimize that loss function constraining the ℓ_2-norm of the weights, for MTL and LTL we constrain D to have a Frobenius norm less or equal than 1, and each c_t is constrained to have an ℓ_2 norm less or equal than 1. During testing we use the 0-1 loss. For example the task-average error is evaluated as

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{1000}\sum_{i=1}^{1000} 1\{\text{sign}(\langle u_t, x_i\rangle) \neq \text{sign}(\langle \hat{u}_t, x_i\rangle)\} \tag{11}$$

where $\hat{u}_t$ are the weight vectors learned by the assessed method.

3.3 MTL Experiment

We first discuss the MTL experiment. We let $d = 50$, and vary $T \in \{5, 10, \ldots, 150\}$, $n \in \{5, 10, \ldots, 150\}$ considering the cases $K = 2$ and $K = 5$. In Figure 1 we report the difference between the classification error of the two methods. These results are obtained by repeating the experiment 10 times, reporting the average difference. In each trial a different set of input points and underlying weight vectors are generated for each task. In the MTL case the training error was always below 0.1 and on average it was smaller than 0.04. This suggests that despite the problem being non-convex, the gradient optimization algorithm finds a good suboptimal solution.

1. The code used for the experiments presented in this section is available at `http://romera-paredes.com/multitask-representation`.

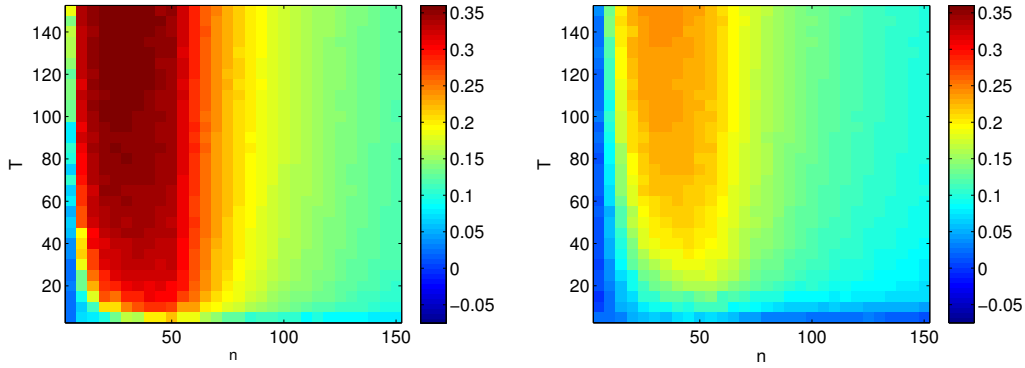

Figure 1: Difference of test classification error, computed according to eq. (11), between ITL and MTL. The vertical axis represents the number of training tasks, and the horizontal axis the number of training instances per task. In the left column $K = 2$, and in the right column $K = 5$.

We have made further experiments to assess the influence of other data settings on the difference between ITL and MTL. In the first of those experiments we have explored the cases in which the dictionary size is overestimated and underestimated. The results are shown in Figure 2. In the left plot the dictionary size is overestimated, in particular the ground truth number of atoms is 2, and the number of atoms used in the MTL method is 5. We can appreciate a similar pattern as the one we saw in Figure 1, although differences between ITL and MTL are not as high. The performance is slightly hampered, as expected due to an overestimation of the number of atoms. On the other hand in Figure 2 (right) we show the results when the number of atoms in the ground truth dictionary is 5, whereas the number of atoms used in the MTL approach is 2. In this case we see that the performance is severely affected by the underestimation of the size of the dictionary, yet we observe that MTL performs better than ITL in the same regime as in the previous experiments.

In the second of these experiments we study how the results are affected when the data are noisy. To do so, we have generated the data so that the ground truth label for instance x_i for task t is given by $\text{sign}(\langle u_t, x_i \rangle + \varepsilon_{ti})$, where $\varepsilon_{ti} \sim \mathcal{N}(0, 1)$. The dictionary size, for both the ground truth and the MTL approach, is $K = 2$. The results are shown in Figure 3, and we can see a similar behaviour as the one in Figure 1, with somewhat smaller differences between ITL and MTL.

3.4 LTL Experiment

In this experiment we test how the dictionary learned at the training stage helps learning new tasks, and we assess how similar the resultant figure is in comparison to the phase diagram derived in the previous section.

The data is generated according to the settings given in the MTL experiment. Furthermore, 50 new tasks are sampled following the same scheme previously described for the purpose of computing the LTL test error. We present the results in Figure 4 (Top). Similar

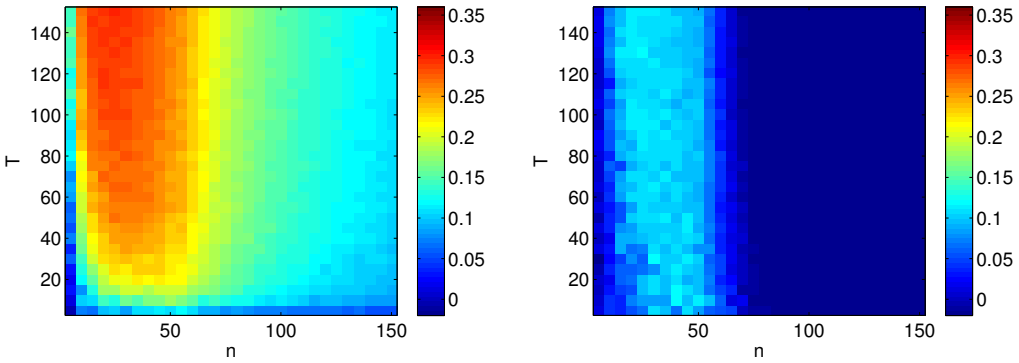

Figure 2: Difference of test classification error, computed according to eq. (11), between ITL and MTL, when the number of atoms of the ground truth dictionary does not match the number of atoms of the MTL model. The plot in the left shows the experiment in which the ground truth number of atoms is 2, whereas the number of atoms used in the MTL approach is 5. The plot in the right shows the opposite scenario: 5 atoms as ground truth, and 2 atoms in the MTL model. The vertical axis represents the number of training tasks, and the horizontal axis the number of training instances per task.

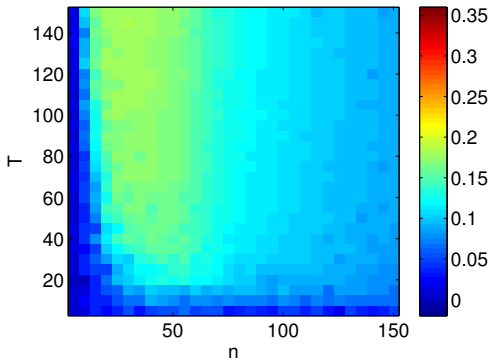

Figure 3: Difference of test classification error, computed according to eq. (11), between ITL and MTL, when adding Gaussian noise to the ground truth labels. The vertical axis represents the number of training tasks, and the horizontal axis the number of training instances per task.

to the previous experiment, we report the average difference between the test error of ITL and LTL after 10 trials.

In Figure 4 (Bottom) we present the theoretical phase diagram, which was generated using $1 \leq T \leq 10^{11}$, $1 \leq n \leq 10^5$, $d = 10^5$, $\delta = 0.0001$. We also plot as a dark line the points in which there is no difference in the performances between ITL and LTL.

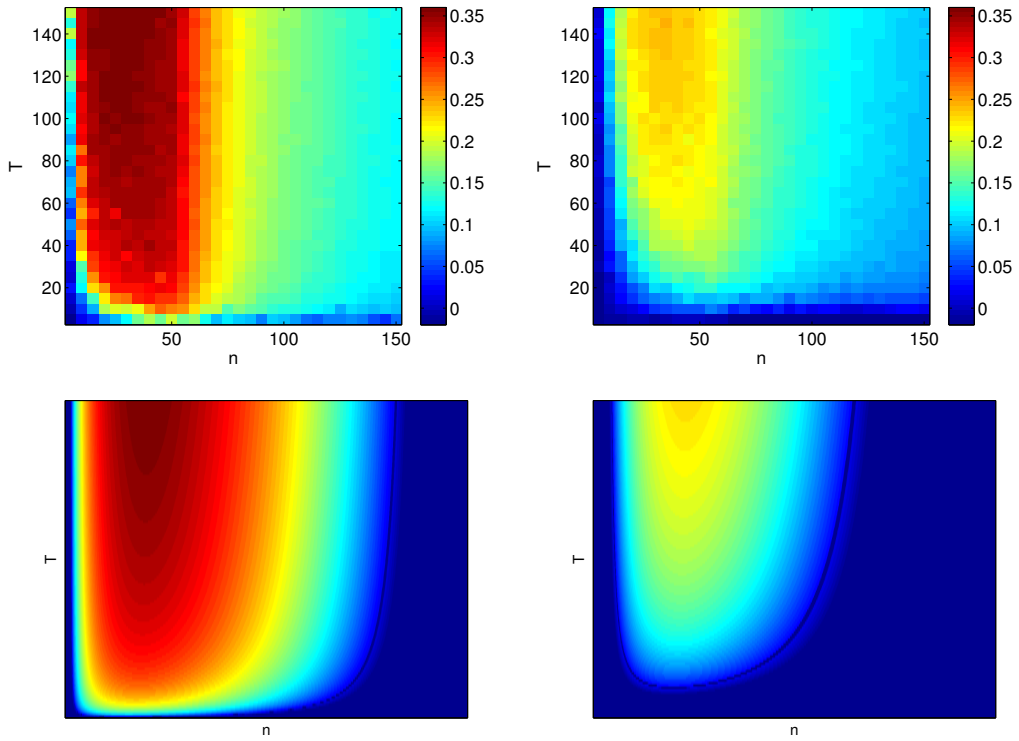

Figure 4: The vertical axis represents the number of training tasks, and the horizontal axis the number of training instances per task. Plots in the top row show the difference of test classification error, computed on 50 new tasks, between ITL and LTL. Plots in the bottom row show the region where the upper bound for LTL is smaller than the lower bound for any equivariant algorithm for ITL (see the discussion in Section 3.1, in particular Equation 9) using $1 \leq T \leq 10^{11}$, $1 \leq n \leq 10^5$, $d = 10^5$, and $\delta = 0.0001$. In the left column $K = 2$, and in the right column $K = 5$.

The reader may object about the much larger parameter values used to generate the plots of theoretical differences, in comparison to the experimental settings. These large parameters are partly a consequence of an accumulation of somewhat loose estimates in the derivation of both the upper and lower bounds. Another reason is that in applying it to a noiseless, finite-dimensional problem (for clarity) we have sacrificed two strong points of our results: independence of input dimension and its agnostic nature. Apart from the large parameter values the theoretical prediction shown in Figure 4 (Bottom) is in very good agreement with the experimental results in Figure 4 (Top).

We have also performed experiments in order to evaluate the influence of noise and under/overestimation of the dictionary size on the difference between ITL and LTL. We obtained similar results as the ones reported for MTL in Figures 2 and 3.

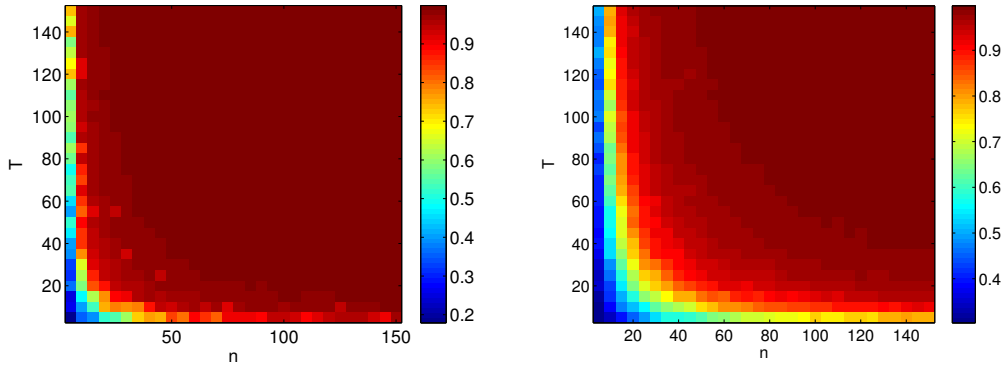

Figure 5: Similarity between the learned dictionary $\hat{D}$ and the ground truth dictionary D, according the similarity measure $s(\hat{D}, D)$ in Equation (12). The vertical axis represents the number of training tasks, and the horizontal axis the number of training instances per task. Left plot: $K = 2$. Right plot: $K = 5$.

Finally, we have compared the learned dictionary, $\hat{D}$, with the ground truth, D, in the same regime of parameters used for the previous experiments. Note that a dictionary could be correct up to permutations and changes of sign of its atoms. To overcome this issue we use the similarity measure

$$s(\hat{D}, D) = \frac{1}{K} \left\| D^\top \hat{D} \right\|_{\mathrm{tr}}, \tag{12}$$

where $\|\cdot\|_{\mathrm{tr}}$ is the sum of singular values of a matrix. Note that $s(\hat{D}, D) = 1$ if $\hat{D}$ and D are the same matrix up to permutation of columns and changes of sign, as requested. The results are found in Figure 5.

Figure 5 indicate that the learned dictionary is close to the true dictionary even for small sample sizes, provide T is large. This supports the results in Figure 1 and the top plots in Figure 4, where MTL or LTL are found to be superior to ITL in this regime, respectively.

4. Proofs of the Main Theorems

In this section we prove our principal results, Theorem 1 and Theorem 2. In preparation for the proofs we will first present some important auxiliary results.

4.1 Tools

We denote by γ a generic vector of independent standard normal variables, whose dimension will be clear from context. A central role in this paper is played by the Gaussian average $G(Y)$ of a set $Y \subseteq \mathbb{R}^n$, which is defined as

$$G(Y) = \mathbb{E} \sup_{y \in Y} \langle \gamma, y \rangle = \mathbb{E} \sup_{y \in Y} \sum_{i=1}^{n} \gamma_i y_i.$$

The reader who is concerned about the measurability of the random variable on the right hand side should replace Y by a countable dense subset of Y, with similar adjustments wherever the Gaussian averages occur.

Rademacher averages, where the γ_i are replaced by uniform $\{-1,1\}$-distributed variables, are somewhat more popular in the literature. We use Gaussian averages instead, because in most cases they are just as easy to bound and possess special properties (Theorem 10 and Theorem 12 below) which we need in our analysis.

The first result is a standard tool to prove uniform bounds on the estimation error in terms of Gaussian averages (Bartlett and Mendelson, 2002).

Theorem 8 *Let $\mathcal{F}$ be a real-valued function class on a space $\mathcal{X}$ and let $\mathbf{X} = (X_1, ..., X_n)$ be a vector of independent random variables and $\mathbf{X}'$ iid to $\mathbf{X}$. Then*

(i)

$$\mathbb{E}_{\mathbf{X}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{\mathbf{X}'} \left[f\left(X_i'\right) \right] - f\left(X_i\right) \right) \leq \frac{\sqrt{2\pi} \mathbb{E}_{\mathbf{X}} G\left(\mathcal{F}\left(\mathbf{X}\right)\right)}{n}$$

(ii) if the members of $\mathcal{F}$ have values in $[0,1]$ then with probability greater than $1-\delta$ in $\mathbf{X}$ for all $f \in \mathcal{F}$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{\mathbf{X}'} \left[f\left(X_i'\right) \right] - f\left(X_i\right) \right) \leq \frac{\sqrt{2\pi} G\left(\mathcal{F}\left(\mathbf{X}\right)\right)}{n} + \sqrt{\frac{9 \ln\left(2/\delta\right)}{2n}}.$$

The following theorem is a vector-valued version of the above, is useful for bounds on the task-averaged estimation error (Ando and Zhang (2005), Maurer (2006b)).

Theorem 9 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to [0,1]^T$, and let $\mu_1, ..., \mu_T$ be probability measures on $\mathcal{X}$ with $\bar{\mathbf{X}} = (\mathbf{X}_1, ..., \mathbf{X}_T) \sim \prod_{t=1}^{T} \left(\mu_t\right)^n$ where $\mathbf{X}_t = (X_{t1}, ..., X_{tn})$. Then with probability greater than $1-\delta$ in $\bar{\mathbf{X}}$ for all $f \in \mathcal{F}$*

$$\frac{1}{T} \sum_t \left(\mathbb{E}_{X \sim \mu_t} \left[f_t\left(X\right) \right] - \frac{1}{n} \sum_{ti} f_t\left(X_{ti}\right) \right) \leq \frac{\sqrt{2\pi} G\left(Y\right)}{nT} + \sqrt{\frac{9 \ln\left(2/\delta\right)}{2nT}},$$

where $Y \subset \mathbb{R}^{Tn}$ is the random set defined by $Y = \left\{ \left(f_t\left(X_{ti}\right) \right) : f \in \mathcal{F} \right\}$.

The previous two theorems replace the problem of proving uniform bounds by the problem of bounding Gaussian averages. One key result in the latter direction is known as Slepian's Lemma (Slepian (1962), Ledoux and Talagrand (1991)).

Theorem 10 *Let Ω and Ξ be mean zero, separable Gaussian processes indexed by a common set $\mathcal{S}$, such that*

$$\mathbb{E}\left(\Omega_{s_1} - \Omega_{s_2}\right)^2 \leq \mathbb{E}\left(\Xi_{s_1} - \Xi_{s_2}\right)^2 \text{ for all } s_1, s_2 \in \mathcal{S}.$$

Then

$$\mathbb{E} \sup_{s \in \mathcal{S}} \Omega_s \leq \mathbb{E} \sup_{s \in \mathcal{S}} \Xi_s.$$

The following corollary is the key to our bound for LTL.

Corollary 11 *Let $Y \subseteq \mathbb{R}^n$ and let $\phi : Y \to \mathbb{R}^m$ be (Euclidean) Lipschitz with Lipschitz constant L. Then*

$$G(\phi(Y)) \leq LG(Y).$$

Proof Define two Gaussian processes indexed by Y as

$$\Omega_y = \sum_{k=1}^{m} \gamma_k \phi(y)_k \text{ and } \Xi_y = L \sum_{i=1}^{n} \gamma_i' y_i,$$

with independent γ_k and γ_i'. Then for any $y, y' \in Y$

$$\mathbb{E}\left(\Omega_y - \Omega_{y'}\right)^2 = \left\|\phi(y) - \phi\left(y'\right)\right\|^2 \leq L^2 \left\|y - y'\right\|^2 = \mathbb{E}\left(\Xi_{s_1} - \Xi_{s_2}\right)^2,$$

so that, by Slepian's Lemma,

$$G(\phi(Y)) = \mathbb{E} \sup_{y \in Y} \Omega_y \leq \mathbb{E} \sup_{y \in Y} \Xi_y = LG(Y).$$

■

In many applications this is applied when $n = m$ and ϕ is defined by $\phi(y_1, ..., y_n) = (\phi_1(y_1), ..., \phi_n(y_n))$ where the real functions $\phi_1, ..., \phi_n$ have Lipschitz constant L.

At one point we will need a generalization of the above corollary, which allows to select ϕ from an entire class of Lipschitz functions. We will use the following result, which is taken from Maurer (2014). It will play an important role in the proof of Theorem 13 below.

Theorem 12 *Let $Y \subseteq \mathbb{R}^n$ have (Euclidean) diameter $D(Y)$ and let $\mathcal{F}$ be a class of functions $f : Y \to \mathbb{R}^m$, all of which have Lipschitz constant at most $L(\mathcal{F})$. Then for any $y_0 \in Y$*

$$G(\mathcal{F}(Y)) \leq c_1 L(\mathcal{F}) G(Y) + c_2 D(Y) Q(\mathcal{F}) + G(\mathcal{F}(y_0)),$$

where c_1 and c_2 are universal constants and

$$Q(\mathcal{F}) = \sup_{\mathbf{y}, \mathbf{y}' \in Y,\ \mathbf{y} \neq \mathbf{y}'} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{\langle \boldsymbol{\gamma}, f(\mathbf{y}) - f(\mathbf{y}') \rangle}{\|\mathbf{y} - \mathbf{y}'\|}.$$

Note that the result allows us to minimize the right hand side in y_0. Analogs of Theorem 10 and Theorem 12 are not available for Rademacher averages. This is the reason why we use the slightly more exotic Gaussian averages.

4.2 Proof of the Excess Risk Bound for the Average Risk

We first establish the following uniform bound. It is of some interest in its own right, in particular since the problem (1) is often non-convex, so that the excess risk bound may not be meaningful in practice. Recall the definition of Q given in Equation (4).

Theorem 13 *Let $\mu_1, \ldots, \mu_T$ be probability measures on $\mathcal{Z}$ and let $Z_{t1}, \ldots, Z_{tn}$ be i.i.d. from μ_t, for $t = 1, \ldots, T$. Let $\delta \in (0,1)$. With probability at least $1-\delta$ in the draw of a multisample $\bar{\mathbf{Z}}$, it holds for every $h \in \mathcal{H}$ and every $f_1, \ldots, f_T \in \mathcal{F}$ that*

$$\mathcal{E}_{\text{avg}}\left(h, f_1, ..., f_T\right) - \frac{1}{Tn}\sum_{ti} \ell\left(f_t\left(h\left(X_{ti}\right)\right), Y_{ti}\right) \leq c_1 \frac{LG(\mathcal{H}(\bar{\mathbf{X}}))}{nT} + c_2 \frac{Q \sup_{h\in\mathcal{H}} \left\|h\left(\bar{\mathbf{X}}\right)\right\|}{n\sqrt{T}} + \sqrt{\frac{9\ln\left(2/\delta\right)}{2nT}},$$

where c_1 and c_2 are universal constants.

Proof By Theorem 9, with probability at least $1-\delta$ in $\bar{\mathbf{Z}}$, for all $h \in \mathcal{H}$ and all $f_1, ..., f_T \in \mathcal{F}$, we have that

$$\mathcal{E}_{\text{avg}}\left(h, f_1, ..., f_T\right) - \frac{1}{Tn}\sum_{ti} \ell\left(f_t\left(h\left(X_{ti}\right)\right), Y_{ti}\right) \leq \frac{\sqrt{2\pi}}{nT} G\left(S\right) + \sqrt{\frac{9\ln\left(2/\delta\right)}{2nT}}, \tag{13}$$

where $S = \left\{\left(\ell\left(f_t\left(h\left(X_{ti}\right)\right), Y_{ti}\right)\right) : f \in \mathcal{F}^T \text{ and } h \in \mathcal{H}\right\} \subseteq \mathbb{R}^{Tn}$. By the Lipschitz property of the loss function ℓ and the contraction lemma Corollary 11 (recall the remark which follows its proof) we have $G\left(S\right) \leq G\left(S'\right)$, where $S' = \left\{\left(f_t\left(h\left(X_{ti}\right)\right)\right) : f \in \mathcal{F}^T \text{ and } h \in \mathcal{H}\right\} \subseteq \mathbb{R}^{Tn}$.

Recall that $\mathcal{H}\left(\bar{\mathbf{X}}\right) \subseteq \mathbb{R}^{KTn}$ is defined by

$$\mathcal{H}\left(\bar{\mathbf{X}}\right) = \left\{\left(h_k\left(X_{ti}\right)\right) : h \in \mathcal{H}\right\},$$

and define a class of functions $\mathcal{F}' : \mathbb{R}^{KTn} \to \mathbb{R}^{Tn}$ by

$$\mathcal{F}' = \left\{y \in \mathbb{R}^{KTn} \mapsto \left(f_t\left(y_{ti}\right)\right) : \left(f_1, ..., f_T\right) \in \mathcal{F}^T\right\}.$$

Then $S' = \mathcal{F}'\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right)$, and by Theorem 12 for universal constants c_1' and c_2'

$$G\left(S'\right) \leq c_1' L\left(\mathcal{F}'\right) G\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right) + c_2' D\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right) Q\left(\mathcal{F}'\right) + \min_{y\in Y} G\left(\mathcal{F}\left(y\right)\right). \tag{14}$$

We now proceed by bounding the individual terms in the right hand side above. Let $y, y' \in \mathbb{R}^{KTn}$, where $y = (y_{ti})$ with $y_{ti} \in \mathbb{R}^K$ and $y' = (y'_{ti})$ with $y'_{ti} \in \mathbb{R}^K$. Then for $f = (f_1, ..., f_T) \in \mathcal{F}^T$

$$\begin{aligned}\left\|f\left(y\right) - f\left(y'\right)\right\|^2 &= \sum_{ti}\left(f_t\left(y_{ti}\right) - f_t\left(y'_{ti}\right)\right)^2 \\ &\leq L^2 \sum_{ti}\left\|y_{ti} - y'_{ti}\right\|^2 = L^2\left\|y - y'\right\|^2,\end{aligned}$$

so that $L(\mathcal{F}') \leq L$. Also

$$\begin{aligned}
& \mathbb{E} \sup_{g \in \mathcal{F}'} \left\langle \gamma, g(y) - g\left(y'\right) \right\rangle \\
= \; & \mathbb{E} \sup_{(f_1,\ldots,f_T)\in\mathcal{F}^T} \sum_{ti} \gamma_{ti} \left(f_t(y_{ti}) - f_t\left(y'_{ti}\right) \right) \\
= \; & \sum_t \mathbb{E} \sup_{f\in\mathcal{F}} \sum_i \gamma_i \left(f(y_{ti}) - f\left(y'_{ti}\right) \right) \\
\leq \; & \sqrt{T} \left(\sum_t \left(\mathbb{E} \sup_{f\in\mathcal{F}} \sum_i \gamma_i \left(f(y_{ti}) - f\left(y'_{ti}\right) \right) \right)^2 \right)^{1/2} \\
\leq \; & \sqrt{T} \left(\sum_t Q^2 \sum_i \left\| y_{ti} - y'_{ti} \right\|^2 \right)^{1/2} \\
= \; & \sqrt{T} Q \left\| y - y' \right\|,
\end{aligned}$$

whence $Q(\mathcal{F}') = \sqrt{T}Q$. Finally we take $y_0 = 0$ and the last term in (14) vanishes since $f(0) = 0$ for all $f \in \mathcal{F}$. Substitution in (14) and using $G(S) \leq G(S')$ we arrive at

$$G(S) \leq c'_1 L G\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right) + c'_2 \sqrt{T} D\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right) Q.$$

Bounding $D\left(\mathcal{H}\left(\bar{\mathbf{X}}\right)\right) \leq 2 \sup_h \left\| h\left(\bar{\mathbf{X}}\right) \right\|$ and substitution in (13) gives the result. ■

Proof of Theorem 1 Let h^* and $f_1^*, ..., f_T^*$ be the minimizers in the definition of $\mathcal{E}^*_{\text{avg}}$. Then

$$\begin{aligned}
& \mathcal{E}_{\text{avg}}(\hat{h}, \hat{f}_1, ..., \hat{f}_T) - \mathcal{E}^*_{\text{avg}} \\
= \; & \left(\mathcal{E}_{\text{avg}}(\hat{h}, \hat{f}_1, ..., \hat{f}_T) - \frac{1}{nT} \sum_{ti} \ell(\hat{f}_t(\hat{h}(X_{ti})), Y_{ti}) \right) \\
+ \; & \left(\frac{1}{nT} \sum_{ti} \ell(\hat{f}_t(\hat{h}(X_{ti})), Y_{ti}) - \frac{1}{nT} \sum_{ti} \ell\left(f_t^*\left(h^*\left(X_{ti}\right)\right), Y_{ti}\right) \right) \\
+ \; & \left(\frac{1}{nT} \sum_{ti} \ell\left(f_t^*\left(h^*\left(X_{ti}\right)\right), Y_{ti}\right) - \frac{1}{T} \sum_t \mathbb{E}_{(X,Y)\sim\mu_t} \ell\left(f_t^*\left(h^*\left(X\right)\right), Y\right) \right).
\end{aligned}$$

The last term involves only the nT random variables $\ell\left(f_t^*\left(h^*\left(X_{ti}\right)\right), Y_{ti}\right)$ with values in $[0,1]$. It can be bounded with probability $1 - \delta/2$ by $\sqrt{\ln(2/\delta)/(2Tn)}$ using Hoeffding's inequality. The middle term is non-positive by definition of $\hat{h}, \hat{f}_1, ..., \hat{f}_T$ being the corresponding minimizers. There remains the first term which we bound by

$$\sup_{h\in\mathcal{H}, f_1,\ldots,f_T\in\mathcal{F}} \mathcal{E}_{\text{avg}}(h, f_1, ..., f_T) - \frac{1}{Tn} \sum_{ti} \ell\left(f_t\left(h\left(X_{ti}\right)\right), Y_{ti}\right).$$

and appeal to Theorem 13 to bound the supremum. A union bound then completes the proof. ■

4.3 Proof of the Excess Risk Bound for Learning-to-learn

Recall the definition of the algorithm parametrized by $h \in \mathcal{H}$

$$a(h)_{\mathbf{Z}} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_i \ell(f(h(X_i)), Y_i) \text{ for } \mathbf{Z} \in \mathcal{Z}^n$$

and the associated minimum $m(h)_{\mathbf{Z}}$. Also recall that

$$\mathcal{E}_\eta(h) = \mathbb{E}_{\mu\sim\eta}\mathbb{E}_{\mathbf{Z}\sim\mu^n}\mathbb{E}_{(X,Y)\sim\mu}\ell(a(h)_{\mathbf{Z}}(X), Y)$$

and the two measures μ_η and ρ_η induced by the environment η and defined by

$$\mu_\eta(A) = \mathbb{E}_{\mu\sim\eta}\mu(A) \text{ for } A \subseteq \mathcal{Z} \text{ and } \rho_\eta(A) = \mathbb{E}_{\mu\sim\eta}\mu^n(A) \text{ for } A \subseteq \mathcal{Z}^n.$$

Also recall the definition of Q' given in Equation (5). Again we begin with a uniform bound.

Theorem 14 *Let $\delta \in (0,1)$. (i) With probability at least $1-\delta$ in $\bar{\mathbf{Z}} \sim \rho_\eta^T$ it holds for every $h \in \mathcal{H}$ that*

$$\mathcal{E}_\eta(h) - \frac{1}{T}\sum_t m(h)_{\mathbf{Z}_t} \leq$$

$$\frac{\sqrt{2\pi}LG(\mathcal{H}(\bar{\mathbf{x}}))}{T\sqrt{n}} + \sqrt{2\pi}Q' \sup_{h\in\mathcal{H}} \sqrt{\frac{\mathbb{E}_{(X,Y)\sim\mu_\eta}\left[\|h(X)\|^2\right]}{n}} + \sqrt{\frac{9\ln(2/\delta)}{2T}}.$$

(ii) With probability at least $1-\delta$ in $\bar{\mathbf{Z}} \sim \rho_\eta^T$ it holds for every $h \in \mathcal{H}$ that

$$\mathcal{E}_\eta(h) - \frac{1}{T}\sum_t m(h)_{\mathbf{Z}_t} \leq$$

$$\frac{\sqrt{2\pi}LG(\mathcal{H}(\bar{\mathbf{x}}))}{T\sqrt{n}} + \frac{\sqrt{2\pi}Q'\sum_t \sup_{h\in\mathcal{H}}\|h(\mathbf{X}_t)\|}{nT} + \sqrt{\frac{16\ln(4/\delta)}{T}}.$$

Proof The key to the proof is the decomposition bound

$$\begin{aligned}\sup_{h\in\mathcal{H}} \mathcal{E}_\eta(h) - \frac{1}{T}\sum_t m(h)_{\mathbf{Z}_t} \quad &\leq \quad \sup_{h\in\mathcal{H}} \mathbb{E}_{\mu\sim\eta}\mathbb{E}_{\mathbf{Z}\sim\mu^n}\left[\mathbb{E}_{(X,Y)\sim\mu}\ell(a(h)_{\mathbf{Z}}(X), Y) - m(h)_{\mathbf{Z}}\right] \\ &\quad + \sup_{h\in\mathcal{H}}\left[\mathbb{E}_{\mathbf{Z}\sim\rho_\eta}[m(h)_{\mathbf{Z}}] - \frac{1}{T}\sum_{t=1}^T m(h)_{\mathbf{Z}_t}\right]. \qquad (15)\end{aligned}$$

In turn we will bound both terms on the right hand side above. A bound on the second term means that we can predict the empirical risk on the data of a future task uniformly in h. A bound on the first term means that we can predict the true risk from the empirical risk on the future task.

We first bound the second term in the right hand side of (15), and use Theorem 8-(ii) on the class of functions

$$\{\mathbf{z} \in \mathcal{Z}^n \mapsto m(h)_{\mathbf{z}} : h \in \mathcal{H}\}$$

to get with probability at least $1-\delta$ in $\bar{\mathbf{Z}} \sim \rho_\eta^T$ that

$$\sup_{h\in\mathcal{H}}\left[\mathbb{E}_{\mathbf{Z}\sim\rho_\eta}\left[m\left(h\right)_{\mathbf{Z}}\right]-\frac{1}{T}\sum_{t=1}^{T}m\left(h\right)_{\mathbf{Z}_t}\right]\leq\frac{\sqrt{2\pi}}{T}G\left(S\right)+\sqrt{\frac{9\ln\left(2/\delta\right)}{2T}},$$

where S is the subset of $\mathbb{R}^T$ defined by

$$S=\left\{\left(m\left(h\right)_{\mathbf{Z}_1},...,m\left(h\right)_{\mathbf{Z}_T}\right):h\in\mathcal{H}\right\}.$$

We will bound the Gaussian average of S using Slepian's inequality (Theorem 10). Define two Gaussian processes indexed by $\mathcal{H}$ as

$$\Omega_h=\sum_t\gamma_t m\left(h\right)_{\mathbf{z}_t}\ \text{ and }\ \Xi_h=\frac{L}{\sqrt{n}}\sum_{kti}\gamma_{kti}h_k\left(x_{ti}\right).$$

Now for any $\mathbf{z}\in\mathcal{Z}^n$ and representations $h,h'\in\mathcal{H}$

$$\begin{aligned}\left(m\left(h\right)_{\mathbf{z}}-m\left(h'\right)_{\mathbf{z}}\right)^2 &= \left(\min_{f\in\mathcal{F}}\frac{1}{n}\sum_i\ell\left(f\left(h\left(x_i\right)\right),y_i\right)-\min_{f\in\mathcal{F}}\frac{1}{n}\sum_i\ell\left(f\left(h'\left(x_i\right)\right),y_i\right)\right)^2\\ &\leq \left(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_i\ell\left(f\left(h\left(x_i\right)\right),y_i\right)-\ell\left(f\left(h'\left(x_i\right)\right),y_i\right)\right)^2\\ &\leq \frac{1}{n}\sup_{f\in\mathcal{F}}\sum_i\left(\ell\left(f\left(h\left(x_i\right)\right),y_i\right)-\ell\left(f\left(h'\left(x_i\right)\right),y_i\right)\right)^2\\ &\leq \frac{L^2}{n}\sum_{ki}\left(h_k\left(x_i\right)-h'_k\left(x_i\right)\right)^2,\end{aligned}$$

where in the last step we used the Lipschitz properties of the loss function ℓ and of the members in the class $\mathcal{F}$. It follows that

$$\begin{aligned}\mathbb{E}\left(\Omega_h-\Omega_{h'}\right)^2 &= \sum_t\left(m\left(h\right)_{\mathbf{z}_t}-m\left(h'\right)_{\mathbf{z}_t}\right)^2\\ &\leq \frac{L\left(\mathcal{F}\right)^2}{n}\sum_{kti}\left(h_k\left(x_{ti}\right)-h'_k\left(x_{ti}\right)\right)^2=\mathbb{E}\left(\Xi_h-\Xi_{h'}\right)^2,\end{aligned}$$

so by Theorem 10

$$G\left(S\right)=\mathbb{E}\sup_k\Omega_h\leq\mathbb{E}\sup_k\Xi_h=\frac{L}{\sqrt{n}}G\left(\mathcal{H}\left(\bar{\mathbf{x}}\right)\right).$$

The second term in the right hand side of (15) is thus bounded with probability $1-\delta$ by

$$\frac{\sqrt{2\pi}LG\left(\mathcal{H}\left(\bar{\mathbf{x}}\right)\right)}{T\sqrt{n}}+\sqrt{\frac{9\ln\left(2/\delta\right)}{2T}}. \tag{16}$$

We now bound the first term on the right hand side of (15) by

$$\begin{aligned}
&\sup_{h\in\mathcal{H}} \mathbb{E}_{\mu\sim\eta}\mathbb{E}_{\mathbf{Z}\sim\mu^n}\left[\mathbb{E}_{(X,Y)\sim\mu}\ell\left(a\left(h\right)_{\mathbf{X}}\left(X\right),Y\right) - m\left(h\right)_{\mathbf{Z}}\right] \\
\leq\ &\sup_{h\in\mathcal{H}} \mathbb{E}_{\mu\sim\eta}\mathbb{E}_{\mathbf{Z}\sim\mu^n}\sup_{f\in\mathcal{F}}\left[\mathbb{E}_{(X,Y)\sim\mu}\ell\left(f\left(h\left(X\right)\right),Y\right) - \frac{1}{n}\sum_i \ell\left(f\left(h\left(X_i\right)\right),Y_i\right)\right].
\end{aligned}$$

For $\mathbf{Z}=(\mathbf{X},\mathbf{Y})\in\mathcal{Z}^n$ and $h\in\mathcal{H}$ denote with $\ell\left(\mathcal{F}\circ h\left(\mathbf{X}\right),\mathbf{Y}\right)$ the subset of $\mathbb{R}^n$ defined by

$$\ell\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right),\mathbf{Y}\right) = \left\{\left(\ell\left(f\left(h\left(X_i\right)\right),Y_i\right)\right) : f\in\mathcal{F}\right\}.$$

Using Theorem 8-(i) and the contraction lemma, Corollary 11, we can bound the last expression above by

$$\begin{aligned}
&\sup_{h\in\mathcal{H}} \mathbb{E}_{\mu\sim\eta}\mathbb{E}_{\mathbf{Z}\sim\mu^n}\sup_{f\in\mathcal{F}}\left[\mathbb{E}_{(X,Y)\sim\mu}\ell\left(f\left(h\left(X\right)\right),Y\right) - \frac{1}{n}\sum_i \ell\left(f\left(h\left(X_i\right)\right),Y_i\right)\right] \\
\leq\ &\sqrt{2\pi}\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{Z}\sim\boldsymbol{\rho}_\eta}\frac{G\left(\ell\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right),\mathbf{Y}\right)\right)}{n} \\
\leq\ &\sqrt{2\pi}\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{Z}\sim\boldsymbol{\rho}_\eta}\frac{G\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right)\right)}{n} \\
=\ &\frac{\sqrt{2\pi}}{n}\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{Z}\sim\boldsymbol{\rho}_\eta}\frac{G\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right)\right)}{\left\|h\left(\mathbf{X}\right)\right\|}\left\|h\left(\mathbf{X}\right)\right\| \\
\leq\ &\frac{\sqrt{2\pi}}{n}Q'\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{Z}\sim\boldsymbol{\rho}_\eta}\left\|h\left(\mathbf{X}\right)\right\|,
\end{aligned}$$

using Hoelder's inequality and the definition of Q' in the last step. But, using Jensen's inequality,

$$\mathbb{E}_{\mathbf{Z}\sim\boldsymbol{\rho}_\eta}\left\|h\left(\mathbf{X}\right)\right\| \leq \sqrt{\mathbb{E}_{\mathbf{Z}\sim\rho_\eta}\sum_i\left\|h\left(X_i\right)\right\|^2} = \sqrt{n\ \mathbb{E}_{(X,Y)\sim\mu_\eta}\left\|h\left(X\right)\right\|^2},$$

since $\mathbf{Z}\sim\rho_\eta$ is iid. Inserting this in the previous chain of inequalities and combining with (16) gives the first part of the theorem.

To obtain the data dependent bound we use the fact that, with probability at least $1-\delta/4$,

$$\begin{aligned}
\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{Z}\sim\boldsymbol{\rho}_\eta}\frac{G\left(\ell\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right),\mathbf{Y}\right)\right)}{n} &\leq \mathbb{E}_{\mathbf{X}\sim\rho_\eta}\sup_{h\in\mathcal{H}}\frac{G\left(\ell\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right),\mathbf{Y}\right)\right)}{n} &(17)\\
&\leq \frac{1}{T}\sum_t\sup_{h\in\mathcal{H}}\frac{G\left(\ell\left(\mathcal{F}\left(h\left(\mathbf{X}_t\right)\right),\mathbf{Y}_t\right)\right)}{n} + \sqrt{\frac{\ln\left(4/\delta\right)}{2T}} &(18)
\end{aligned}$$

The last inequality follows from Hoeffding's inequality since for any $h\in\mathcal{H}$ and any sample $\mathbf{Z}\in\mathcal{Z}^n$

$$\begin{aligned}
0\ &\leq\ \frac{G\left(\ell\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right),\mathbf{Y}\right)\right)}{n} = \frac{1}{n}\mathbb{E}_\gamma\sup_f\sum_i\gamma_i\ell\left(f\left(h\left(X_i\right)\right),Y_i\right) \\
&\leq\ \frac{1}{n}\mathbb{E}_\gamma\left(\sum_i\gamma_i^2\right)^{1/2}\sup_f\left(\sum_i\ell\left(f\left(h\left(X_i\right)\right),Y_i\right)^2\right)^{1/2} \leq 1,
\end{aligned}$$

where we have also used the fact that the loss function ℓ has range in $[0,1]$. Bounding

$$G\left(\ell\left(\mathcal{F}\circ h\left(\mathbf{X}_t\right),\mathbf{Y}_t\right)\right)\leq Q'\left\|h\left(\mathbf{X}_t\right)\right\|$$

as above and combining (18) and (16) in (15) with a union bound gives the second inequality of the theorem.

■

Remark: In the proof of the fully data-dependent part above the bound on

$$\sup_{h\in\mathcal{H}}\mathbb{E}_{\mathbf{Z}\sim\boldsymbol{\rho}_\eta}\frac{G\left(\ell\left(\mathcal{F}\left(h\left(\mathbf{X}\right)\right),\mathbf{Y}\right)\right)}{n}$$

is very crude. Instead we could have again invoked Theorem 8 to get a better bound with a more complicated expression involving nested Gaussian averages. We have chosen the simpler path for greater clarity.

Proof of Theorem 2 Recall that

$$\mathcal{E}_\eta^*=\min_{h\in\mathcal{H}}\mathbb{E}_{\mu\sim\eta}\left[\min_{f\in\mathcal{F}}\mathbb{E}_{(X,Y)\sim\mu}\ell\left(f\left(h\left(X\right)\right),Y\right)\right].$$

We denote with h^* the minimizer in $\mathcal{H}$ occurring in the definition of $\mathcal{E}_\eta^*$. We have the following decomposition

$$\begin{aligned}
\mathcal{E}_\eta(\hat{h})-\mathcal{E}_\eta^* &= \left(\mathcal{E}_\eta(\hat{h})-\frac{1}{T}\sum_t m(\hat{h})_{\mathbf{Z}_t}\right) && (19)\\
&+\left(\frac{1}{T}\sum_t m(\hat{h})_{\mathbf{Z}_t}-\frac{1}{T}\sum_t m(h^*)_{\mathbf{Z}_t}\right) && (20)\\
&+\left(\frac{1}{T}\sum_t m(h^*)_{\mathbf{Z}_t}-\mathbb{E}_{\mathbf{Z}\sim\rho_\eta}\left[m\left(h^*\right)_{\mathbf{Z}}\right]\right) && (21)\\
&+\mathbb{E}_{\mu\sim\eta}\left[\mathbb{E}_{\mathbf{Z}\sim\mu^n}\left[m\left(h^*\right)_{\mathbf{Z}}\right]-\min_{f\in\mathcal{F}}\mathbb{E}_{(X,Y)\sim\mu}\ell\left(f\left(h^*\left(X\right)\right),Y\right)\right]. && (22)
\end{aligned}$$

For a fixed distribution μ let f_μ^* be the minimizer in $\min_{f\in\mathcal{F}}\mathbb{E}_{(X,Y)\sim\mu}\ell\left(f\left(h^*\left(X\right)\right),Y\right)$. By definition of $m\left(h^*\right)_{\mathbf{Z}}$ we have for every $\mu\sim\eta$ that

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}\sim\mu^n}\left[m\left(h^*\right)_{\mathbf{Z}}\right] &= \mathbb{E}_{\mathbf{Z}\sim\mu^n}\min_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\ell\left(f\left(h^*\left(X_i\right)\right),Y_i\right)\\
&\leq \mathbb{E}_{\mathbf{Z}\sim\mu^n}\frac{1}{m}\sum_{i=1}^m\ell\left(f_\mu^*\left(h^*\left(X_i\right)\right),Y_i\right)\\
&= \mathbb{E}_{(X,Y)\sim\mu}\ell\left(f_\mu^*\left(h^*\left(X\right)\right),Y\right),
\end{aligned}$$

since $\mathbf{Z}$ is iid. The term in (22) is therefore non-positive.

The term in (21) involves the deviation of the empirical and true averages of the T iid $[0,1]$-valued random variables $m\left(h^*\right)_{\mathbf{Z}_t}$. With Hoeffding's inequality this can be bounded with probability at least $1-\delta/8$ by $\sqrt{\ln\left(8/\delta\right)/\left(2T\right)}$. The term (20) is non-positive by the definition of $\hat{h}$.

There remains the term (19), which we bound by Theorem 14. The result now follows by combining this bound with the bound on (21) in a union bound and some numerical simplifications. ■

5. Conclusion

Several works have advocated that sharing features among tasks as a means to learning representations which capture invariant properties to tasks can be highly beneficial. In this paper, we studied the statistical properties of a general MTRL method, presenting bounds on its learning performance in both settings of MTL and LTL. Our work provides a rigorous justification of the benefit offered by MTRL over learning the tasks independently. To give the paper a clear focus we have illustrated this advantage in the case of linear feature learning. Our results however apply to fairly general classes of representations $\mathcal{H}$ and specifications $\mathcal{F}$, and similar conclusions may be derived for other nonlinear MTRL methods. We conclude by sketching specific cases which deserve a separated study:

- *Deep networks.* As we noted our bounds directly apply to multilayer, deep architectures obtained by iteratively composing linear transformations with nonlinear activation functions, such as the rectifier linear unit or the sigmoid functions. The representations learned by such methods tend to be specific in that only a subset of components are "active" on each given input, which makes our bounds particularly attractive for further analysis.

- *Sparse coding.* Another interesting case of our framework is obtained when the specialized class $\mathcal{F}$ consists of sparse linear predictors. This case has been considered in Maurer et al. (2013); Ruvolo and Eaton (2014) when the representation class consists of linear functions. Different choices of sparse classes $\mathcal{F}$ could lead to interesting learning methods.

- *Representations in RKHS.* As we already noted the feature maps forming the class $\mathcal{H}$ could be vector-valued functions in a reproducing kernel Hilbert space. Although kernel methods are more difficult to apply to large datasets required for MTRL and need additional approximation steps, the representations learned using for example Gaussian kernels would be very specific and suitable for our bounds.

Acknowledgments

We thank the reviewers for their helpful comments.

References

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.

M. Anthony, P. Bartlett. *Learning in Neural Networks: Theoretical Foundations*. Cambridge University Press, 1999.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Machine Learning*, 73(3):243-272, 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, N. Eiron, H. U. Simon. Limitations of Learning Via Embeddings in Euclidean Half Spaces, Journal of Machine Learning Research, (3):441-461, 2002.

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.

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.

S. Dasgupta, A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22: 60-65, 2003.

A. Ehrenfeucht, D. Haussler, M. Kearns, L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3): 247-251, 1989.

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

R. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2014.

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 classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

I. Kuzborskij, and F. Orabona. Stability and Hypothesis Transfer Learning. *Proc. International Conference on Machine Learning (ICML)*, 2013.

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

P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556-1559, 1995.

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. A chain rule for the expected suprema of Gaussian processes. *Proc. 25th International Conference on Algorithmic Learning Theory*, 2014.

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.

A. Maurer and M. Pontil. Excess risk bounds for multitask learning with trace norm regularization. *Proc. 26th Annual Conference on Computational Learning Theory (COLT)*, 2013.

A. Maurer, M. Pontil, B. Romera-Paredes. Sparse coding for multitask and transfer learning. *Proc. 30th International Conference on Machine Learning*, JMLR W&CP 28 (2):343-351, 2013

A. Maurer, M. Pontil, B. Romera-Paredes. An inequality with applications to structured sparsity and multitask dictionary learning. *Proc. 27th Annual Conference on Computational Learning Theory (COLT)*, 2014.

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

A. Pentina and C.H. Lampert. A PAC-Bayesian bound for lifelong learning. *Proc. International Conference on Machine Learning (ICML)*, 2014.

P. Ruvolo and E. Eaton. Online multi-task learning via sparse dictionary optimization. In Proc. of the 28th AAAI Conference on Artificial Intelligence, 2014.

D. Slepian. The one-sided barrier problem for gaussian noise. *Bell System Tech. J.*, 41:463-501, 1962.

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

C. Widmer, M. Kloft, X. Lou X and G.Rätsch. Regularization-based multitask learning: With applications to genome biology and biomedical imaging. *German Journal on Artificial Intelligence*, 2013.

A.W. Van Der Vaart and J.A. Wellner. *Weak Convergence of Empirical Processes*. Springer, 1996