Learning Fair and Transferable Representations

Luca Oneto
Universitá di Pisa
Department of Informatics, 56127 Pisa, Italy
luca.oneto@gmail.com

Michele Donini
Amazon Web Services, Seattle, WA 98109, USA
donini@amazon.com

Andreas Maurer
Adalbertstr 55, D-80799 Munich, Germany
am@andreas-maurer.eu

Massimiliano Pontil
Istituto Italiano di Tecnologia, 16163 Genoa, Italy
massimiliano.pontil@iit.it
and
University College London, Department of Computer Science
London WC1E 6BT, UK

December 30, 2019

Abstract

Developing learning methods which do not discriminate subgroups in the population is a central goal of algorithmic fairness. One way to reach this goal is by modifying the data representation in order to meet certain fairness constraints. In this work we measure fairness according to demographic parity. This requires the probability of the possible model decisions to be independent of the sensitive information. We argue that the goal of imposing demographic parity can be substantially facilitated within a multitask learning setting. We leverage task similarities by encouraging a shared fair representation across the tasks via low rank matrix factorization. We derive learning bounds establishing that the learned representation transfers well to novel tasks both in terms of prediction performance and fairness metrics. We present experiments on three real world datasets, showing that the proposed method outperforms state-of-the-art approaches by a significant margin.

1 Introduction

During the last decade, the widespread distribution of automatic systems for decision making is raising concerns about their potential for unfair behaviour [3, 30, 6]. As a consequence machine learning models are often required to meet fairness requirements, ensuring the correction and limitation of – for example – racist or sexist decisions.

In literature, it is possible to find a plethora of different methods to generate fair models with respect to one or more sensitive attributes (e.g. gender, ethnic group, age). These methods can be mainly divided in three families: (i) methods in the first family change a pre-trained model in order to make it more fair (while trying to maintain the classification performance) [13, 16, 27]; (ii) in the second family, we can find methods that enforce fairness directly during the training phase, e.g. [36, 11, 35, 1]; (iii) the third family of methods implements fairness by modifying the data representation, and then employs standard machine learning methods [37, 8].

All methods in the previous families have in common the goal of creating a fair model from scratch on the specific task at hand. This solution may work well in specific cases, but in a large number of real world applications, using the same model (or at least part of it) over different tasks is helpful if not mandatory. For example, it is common to perform a fine tuning over pre-trained models [10], keeping fixed the internal representation. Indeed, most modern machine learning frameworks (especially the deep learning ones) offer a set of pre-trained models that are distributed in so-called model zoos[1]. Unfortunately, fine tuning pre-trained models on novel previously unseen tasks could lead to an unexpected unfairness behaviour, even starting from an apparently fair model for previous tasks (e.g. discriminatory transfer [20] or negative legacy [19]), due to missing generalization guarantees concerning the fairness property of the model.

In order to overcome the above problem, in this paper we embrace the framework of multitask learning. We aim to leverage task similarity in order to learn a fair representation that provably generalizes well to unseen tasks. By this we mean that when the representation is used to learn novel tasks, it is guaranteed to learn a model that has both a small error and meets the fairness requirement. We measure fairness according to demographic parity [7] (for an extended analysis of the different fairness definitions see [33, 36]). It requires the probability of possible model decisions to be independent of the sensitive information. We argue that multitask methods based on low rank matrix factorization are well suited to learn a shared fair representation according to demographic parity. We show theoretically that the learned representation transfers to novel tasks both in terms of prediction performance and fairness metrics. Other papers in literature already pursued a similar goal [5, 12, 21, 22, 24, 25, 34]. They mainly rely on generating a model acting randomly when the internal representation is exploited to predict the sensitive variable. No actual constraint is imposed directly on the internal representation, but only over the output of the model.

The main contribution of this paper is to augment multitask learning methods based on low rank matrix factorization by imposing a fairness constraint directly on the representation factor matrix. We show empirically and theoretically, via learning bounds, that by imposing the fairness constraint within the multitask learning method, the learned representation can be used to train new models over different (new and possibly unseen) tasks, maintaining the desiderata of an accurate and fair model. Our learning bound improves over previous bounds for learning-to-learn and by being fully

[1]Computer Science Department, University College London, WC1E 6BT London, United Kingdom
[2]Computational Statistics and Machine Learning - Istituto Italiano di Tecnologia, 16100 Genova, Italy
[3]Electrial and Electronics Engineering Department, Imperial College London, SW7 2BT, United Kingdom.
[1]See for example the Caffe Model Zoo: `github.com/BVLC/caffe/wiki/Model-Zoo`

data dependent can be used to evaluate the transfer capability of the learned representation.

The paper is organized in the following manner. In Sec. 2, we discuss previous related work aimed at learning fair representations. In Sec. 3, we introduce the proposed method. In Sec. 4, we study the generalization properties of the method, embracing the framework of learning-to-learn. In Sec. 5, we experimentally compare the proposed method against different baselines and state-of-the-art approaches on three real world datasets. Finally, in Sec. 6 we discuss directions of future research.

2 Related work

Let us consider a composition of models $f(g(x))$ where $x \in \mathbb{R}^d$ is a vector of raw features (an element of the input space), $g : \mathbb{R}^d \to \mathbb{R}^r$ is a function mapping the input space into a new one, that we refer to as the representation. In other words, the function g synthesizes the information needed to solve a particular task (or a set of tasks) by learning a function f, chosen from a set of possible functions.

In this work – and more generally in the current literature [5, 12, 21, 22, 24, 25, 34, 18, 37] – with fair representation we refer to the concept of learning a representation function g, which does not discriminate subgroups in the data. Namely, g is conditionally independent of subgroup membership. This approach is different from most commonly used approaches [11, 16, 35], in which the focus is to solve a task (or a set of tasks) without discriminating subgroups in the data, regardless of the fairness of the representation itself. That is, in the previously mentioned work a fair model $f : \mathbb{R}^r \to \mathbb{R}$ is learned directly from the raw data, without performing any explicit representation extraction.

In particular, in [5, 12, 21, 22, 24, 25, 34], the authors propose different neural networks architectures together with modified learning strategies able to learn a representation that obscures or removes the sensitive variable. In the general case, all these methods have an input, a target variable (i.e. the task at hand) and a binary sensitive variable. The objective is to learn a representation that: (i) preserves information about the input space; (ii) is useful for predicting the target; (iii) is approximately independent of the sensitive variable. In practice, these methods pursue the goal of making the generated model act randomly when the internal representation is exploited to predict the sensitive variable. In this sense, no actual constraint is directly imposed on the internal representation, but only on the output of the model.

In [18], instead, the authors show how to formulate the problem of counterfactual inference as a domain adaptation problem, and more specifically a covariate shift problem [29]. The authors derive two new families of representation algorithms for counterfactual inference. The first one is based on linear models and variable selection, and the other one on deep learning. The authors show that learning representations that encourage similarity (i.e. balance) between the treatment and control populations leads to better counterfactual inference; this is in contrast to many methods which attempt to create balance by re-weighting samples.

Finally, in [37], the authors learn a representation of the data that is a probability distribution over clusters where learning the cluster of a datapoint contains no-information about the sensitive variable, namely fair clustering. In this sense, the clustering is learned to be fair and also discriminative for the prediction task at hand.

3 Method

In this section, we present our method to learn a shared fair representation from multiple tasks. We consider T supervised learning tasks (each could be a binary classification or regression problem). Each task $t \in \{1, \dots, T\}$ is identified by a probability distribution μ_t on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the set of non-sensitive input variables, $\mathcal{S} = \{1, 2\}$ is the set of values of a binary sensitive variable[2] and $\mathcal{Y}$ is the output space which is either $\{-1, 1\}$ for binary classification or $\mathcal{Y} \subset \mathbb{R}$ for regression. We let $\mathbf{z}_t = (x_{t,i}, s_{t,i}, y_{t,i})_{i=1}^m \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^m$ be the training sequence for task t, which is sampled independently from μ_t. The goal is to learn a predictive model $f_t : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ for each task $t \in \{1, \dots, T\}$.

Depending on the application at hand, the model may include (i.e. $f : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ or not (i.e. $f : \mathcal{X} \to \mathcal{Y}$) the sensitive feature in its functional form. In the following we consider the case that the functions f_t are linear, and to simplify the presentation we consider the case that s is not included in the functional form, that is, $f_t(x) = \langle w_t, x \rangle$, where $w_t \in \mathbb{R}^d$ is a vector of parameters. The case in which both x and s are used as predictors is obtained by adding two more components to x, representing the one-hot encoding of s, and letting $w_t \in \mathbb{R}^{d+2}$.

A general multitask learning formulation (MTL) is based on minimizing the multitask empirical error plus a regularization term which leverages similarities between the tasks. A natural choice for the regularizer which is considered in this paper is given by the trace norm, namely the sum of the singular values of the matrix $W = [w_1 \cdots w_T] \in \mathbb{R}^{d \times T}$. It is well know, that this problem is equivalent to the matrix factorization problem,

$$\min_{A,B} \quad \frac{1}{Tm} \sum_{t=1}^{T} \sum_{i=1}^{m} \left(y_{t,i} - \langle b_t, A^\top x_{t,i} \rangle \right)^2 + \frac{\lambda}{2} (\|A\|_F^2 + \|B\|_F^2) \tag{1}$$

where $A = [a_1 \dots a_r] \in \mathbb{R}^{d \times r}$ and $B = [b_1 \dots b_T] \in \mathbb{R}^{r \times T}$ and $\| \cdot \|_F$ is the Frobenius norm, see e.g. [31] and references therein. Here $r \in \mathbb{N}$ is the number of factors, that is the upper bound on the rank of $W = AB$. If $r \geq \min(d, T)$ then Problem (1) is equivalent to trace norm regularization [2], see e.g. [9] and references therein[3]. We follow the formulation of Eq. (1) since it can easily be solved by gradient descent or alternate minimization as we discuss next. Once the problem is solved the estimated parameters of the function w_t for the tasks' linear models are simply computed as $w_t = Ab_t$. We also note that for simplicity the problem is stated with the square loss function, but our observations extended to the general case of proper convex loss functions.

Note that the method can be interpreted as a 2-layer network with linear activation functions. Indeed, the matrix $A^\top$ applied to an input vector $x \in \mathbb{R}^d$ induces the linear representation $A^\top x = (a_1^\top x, \cdots, a_r^\top x)^\top$. We would like this representation to be fair w.r.t. the sensitive feature. Specifically, we require that each component of the representation vector satisfies the demographic parity constraint [14, 33] on each task. This means that, for every measurable subset $C \subset \mathbb{R}^r$, and for every $t \in \{1, \dots, T\}$, we require that

$$\mathbb{P}(A^\top x_t \in C \mid s = 1) = \mathbb{P}(A^\top x_t \in C \mid s = 2) \tag{2}$$

that is the two conditional distributions are the same. We relax this constraint by requiring, for every $t \in \{1, \dots, T\}$, that both distributions have the same mean. Furthermore, we compute the

[2] Our method naturally extends to multiple sensitive variables but for ease of presentation we consider only the binary case.

[3] If $r < \min(d, T)$ then Problem (1) is equivalent to trace norm regularization plus a rank constraint.

means from empirical data. For each training sequence $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^T$ and $s \in \mathcal{S}$, we use the notation $I_s(\mathbf{z}) = \{(x_i, y_i) : s_i = s\}$, define the empirical conditional means

$$c(\mathbf{z}) = \frac{1}{|I_1(\mathbf{z})|} \sum_{i \in I_1(\mathbf{z})} x_i - \frac{1}{|I_2(\mathbf{z})|} \sum_{i \in I_2(\mathbf{z})} x_i \tag{3}$$

and then relax the constraint of Eq. (2) to

$$A^\top c(\mathbf{z}_t) = 0. \tag{4}$$

This is a crude approximation since it corresponds to requiring the first order moment of the two distribution to be the same. However, as we shall see, it works well in practice and has the major advantage of turning a non-convex constraint in a convex one. We note that a similar approximation has been considered in [26] in the case of fair regression, and reported to be empirically effective.

Based on the above reasoning, we propose to learn a fair linear representation as a solution to the optimization problem

$$\begin{aligned} \min_{A,B} \quad & \frac{1}{Tm} \sum_{t=1}^{T} \sum_{i=1}^{m} \left(y_{t,i} - \langle b_t, A^\top x_{t,i} \rangle\right)^2 + \frac{\lambda}{2}(\|A\|_F^2 + \|B\|_F^2) \\ & A^\top c(\mathbf{z}_t) = 0, \quad t \in \{1, \ldots, T\}. \end{aligned} \tag{5}$$

where we used the shorthand notation $c_t = c(\mathbf{z}_t)$. There are many methods to tackle Problem (5). A natural approach is based on alternate minimization. We discuss the main steps below. Let $y_t = [y_{t,1}, \ldots, y_{t,m}]^\top$, the vector formed by the outputs of task t, and let $X_t = [x_{t,1}^\top, \ldots, x_{t,m}^\top]^\top$, the data matrix for task t.

When we regard A as fixed and solve w.r.t. B, then Problem (5) can be reformulated as

$$\min_{B} \quad \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} - \begin{bmatrix} X_1 A & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & X_T A \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_T \end{bmatrix} \right\|^2 + \lambda \left\| \begin{bmatrix} b_1 \\ \vdots \\ b_T \end{bmatrix} \right\|^2 \tag{6}$$

which can be easily solved. In particular note that the problem decouples across the tasks, and each task specific problem amounts running ridge regression on the data transformed by the representation matrix $A^\top$. When instead B is fixed and we solve w.r.t. A, Problem (5) can be reformulated as

$$\min_{A} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} - \begin{bmatrix} b_{1,1} X_1 & \cdots & b_{1,r} X_1 \\ & \vdots & \\ b_{t,1} X_T & \cdots & b_{t,r} X_T \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_r \end{bmatrix} \right\|^2 + \lambda \left\| \begin{bmatrix} a_1 \\ \vdots \\ a_r \end{bmatrix} \right\|^2 \quad \text{s.t.} \quad \begin{bmatrix} a_1^T \\ \vdots \\ a_r^T \end{bmatrix} \circ \begin{bmatrix} c_1, \ldots, c_T \end{bmatrix} = 0$$

where $\circ$ is the Kronecker product for partitioned tensors (or Tracy-Singh product). Consequently by alternating minimization we can solve the original problem. Note also that we may relax the equality constraint as $\frac{1}{T} \sum_{t=1}^{T} \|A^\top c(\mathbf{z}_t)\|^2 \leq \epsilon$, where ϵ is some tolerance parameter. In fact, this may be required when the vectors $c(\mathbf{z}_t)$ span the all input space. In this case we may also add a soft constraint in the regularizer.

We conclude this section by noting that if demographic parity is satisfied at the representation level, that is, Eq. (2) holds true, then every model built from such representation will satisfy

demographic parity as well. Likewise if the representation satisfies the convex relaxation of Eq. (4), then it will also hold that $\langle w_t, c(\mathbf{z}_t)\rangle = \langle b_t, A^\top c(\mathbf{z}_t)\rangle = 0$, that is the task weight vectors will satisfy the first order moment approximation of demographic parity. More importantly, as we will show in the next section, if the tasks are randomly observed, then demographic parity will also be satisfied on future tasks with high probability. In this sense our method can be interpreted as learning a fair transferable representation.

4 Learning bound

In this section, we study the learning ability of the proposed method. We consider the setting of learning-to-learn [4], in which the training tasks (and their corresponding datasets) used to find a fair data representation are regarded as random variables from a meta-distribution. The learned representation matrix A is then transferred to a novel task, by applying ridge regression on the task dataset, in which the input x is transformed as $A^\top x$. In [23] a learning bound is presented, linking the average risk of the method over tasks from the meta-distribution (the so-called transfer risk) to the multi-task empirical error on the training tasks. This result quantifies the good performance of the representation learning method when the number of tasks grow and the data distribution on the raw input data is intrinsically high dimensional (hence learning is difficult without representation learning). We extend this analysis to the setting of algorithmic fairness, in which the performance of the algorithm is evaluated both relative to risk and the fairness constraint. We show that both quantities can be bounded by their empirical counterparts evaluated on the training tasks.

To present our result we introduce some more notation. We let $\mathcal{E}_\mu(w)$ and $\mathcal{E}_\mathbf{z}(w)$ be the expected and empirical errors of a weight vector w, that is

$$\mathcal{E}_\mu(w) = \mathbb{E}_{(x,y)\sim\mu}[(y - \langle w, x\rangle)^2], \qquad \mathcal{E}_\mathbf{z}(w) = \frac{1}{m}\sum_{i=1}^m (y_i - \langle w, x_i\rangle)^2.$$

Furthermore, for every matrix $A \in \mathbb{R}^{d\times r}$ and for every data sample $\mathbf{z} = (x_i, y)_{i=1}^m$, we define $b_A(\mathbf{z}) = \operatorname{argmin}_{b\in\mathbb{R}^r} \frac{1}{m}\sum_{i=1}^m (y_i - \langle b, A^\top x_i\rangle)^2 + \lambda\|b\|^2$ be the minimizer of ridge regression with modified data representation, that is where "+" is the pseudo-inverse operation.

Theorem 1. *Let A be the representation learned by solving Problem (1) and renormalized so that $\|A\|_F = 1$. Let tasks $\mu_1, \ldots, \mu_T$ be independently sampled from a meta-distribution ρ, and let $\mathbf{z}_t$ be sampled from μ_t^m for $t \in \{1, \ldots, T\}$. Assume that the input marginal distribution of random tasks from ρ is supported on the unit sphere and that the outputs are in the interval $[-1, 1]$, almost surely. Let $r = \min(d, T)$. Then, for any $\delta \in (0, 1]$ it holds with probability at least $1 - \delta$ in the drawing of the datasets $\mathbf{z}_1, \ldots, \mathbf{z}_T$, that*

$$\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{\mathbf{z}\sim\mu^m}\ \mathcal{R}_\mu(w_A(\mathbf{z})) - \frac{1}{T}\sum_{t=1}^T \mathcal{R}_{\mathbf{z}_t}(w_A(\mathbf{z}_t)) \le \frac{4}{\lambda}\sqrt{\frac{\|\hat{C}\|_\infty}{m}} + \frac{24}{\lambda m}\sqrt{\frac{\ln\frac{8mT}{\delta}}{T}} + \frac{14}{\lambda}\sqrt{\frac{\ln(mT)\|\hat{C}\|_\infty}{T}} + \sqrt{\frac{2\ln\frac{4}{\delta}}{T}}$$

and

$$\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{\mathbf{z}\sim\mu^m}\|Ac(\mathbf{z})\|^2 - \frac{1}{T}\sum_{t=1}^T \|Ac(\mathbf{z}_t)\|^2 \le 96\frac{\ln\frac{8r^2}{\delta}}{T} + 6\sqrt{\frac{\|\hat{\Sigma}\|_\infty \ln\frac{8r^2}{\delta}}{T}}.$$

Proof. Let $D = \frac{1}{\lambda} A^\top A$. Note that algorithm $\mathbf{z} \mapsto w_D(\mathbf{z}) = A b_A(\mathbf{z})$ is equivalent to run regularized least squares on the original dataset, constraining the paramter vector w to be in the range of D and using the regularizer $w^\top D^+ w$, where "$^+$" denote the pseudo-inverse. The first claim follows from Theorem 6 stated in the appendix, with $\mathcal{D} = \{D \succeq 0, \mathrm{tr} D \leq 1/\lambda\}$, noting that the algorithm has kernel stability 2, the function $M(K) = 2K + 1$, $\|D\|_\infty \leq \|D\|_1 = 1/\lambda$. We then use the first inequality in Corollary 3 in the appendix to upper bound $\sqrt{\|C\|}$ by $\sqrt{\|\hat{C}\| + 6\sqrt{(\ln(4mT)/\delta)/(mT)}}$ and a union bound.

To prove the second claim we note that

$$\frac{1}{T} \sum_{t=1}^{T} \|Ac(\mathbf{z}_t)\|^2 = \mathrm{tr} D \hat{\Sigma}, \quad \hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} c(\mathbf{z}_t) \otimes c(\mathbf{z}_t) \tag{7}$$

and similarly

$$\mathbb{E}_{\mu \sim \rho} \mathbb{E}_{\mathbf{z} \sim \mu^m} \|Ac(\mathbf{z})\|^2 = \mathrm{tr} D \Sigma, \quad \Sigma = \mathbb{E}_{\mu \sim \rho} \mathbb{E}_{\mathbf{z} \sim \mu^m} c(\mathbf{z}) \otimes c(\mathbf{z}). \tag{8}$$

Then

$$\mathbb{E}_{\mu \sim \rho} \mathbb{E}_{\mathbf{z} \sim \mu^m} \|Ac(\mathbf{z})\|^2 - \frac{1}{T} \sum_{t=1}^{T} \|Ac(\mathbf{z}_t)\|^2 = \mathrm{tr} D (\Sigma - \hat{\Sigma}) \leq \|D\|_1 \|\Sigma - \hat{\Sigma}\|_\infty = \|\Sigma - \hat{\Sigma}\|_\infty.$$

The second inequality then follows immediately from inequality (12) in Cor. 3, with $N = T$ and $A_t = (1/4) c(\mathbf{z_t}) \otimes \mathbf{c}(\mathbf{z_t})$. ∎

We make some remarks on the above result:

1. The first bound in Theorem 1 improves Theorem 2 in [23]. The improvement is due to the introduction of the empirical total covariance in the second term in the RHS of the inequality. The result in [23] instead contains the term $\sqrt{1/T}$, which can be considerably larger when the raw input is distributed on a high dimensional manifold.

2. The bounds in Theorem 1 can be extended to hold with variable sample size per task. In order to simplify the presentation, we assume that all datasets are composed of the same number of points m. The general setting can be addressed by letting the sample size be a random variable and introducing the slightly different definition of the transfer risk in which we also take the expectation w.r.t. the sample size.

3. The hyperparameter λ is regarded as fixed in the analysis. In practice it will be chosen by cross-validation as in our experiments below.

4. The bound on fairness measure contains two terms in the right hand side, in the spirit of Bernstein's inequality. The slow term $O(1/\sqrt{T})$ contains the spectral norm of the covariance of difference of means across the sensitive groups. Notice that $\|\Sigma\|_\infty \leq 1$ but it can be much smaller when the means are close to each other, that is, when the original representation is already approximately fair.

5 Experiments

In this section, we compare the proposed method against different baselines and state-of-the-art-methods.

Settings. In order to better understand the performance of the proposed method we performed two sets of experiments.

In the first set (Table 1) we compare the following methods: (a) Unconstrained single task learning (STL), (b) Fair constrained STL, (c) Unconstrained MTL, (d) Fair constrained MTL, that is the proposed method. We test each method either on the same tasks exploited during the training phase, or on novel tasks. Furthermore, we consider both the case where the sensitive feature is present, and not in the functional form of the model (i.e. the sensitive feature is known or not in the testing phase).

In the second set of experiments (Table 2) we compare, in the same setting that we just described, (a) Standard MTL with the fairness constraints on the outputs (M1), (b) feed-forward neural network (FFNN) with linear activation and the fair shared representation method presented in [22] (M2), (c) FFNN with linear activation by exploiting a fair shared representation as presented in [12] (M3), (d) Fair constrained MTL (Our Method). We used linear activation functions in FFNN for fair comparison, since the proposed method learns linear models.

Concerning the experiments on the same task setting, we train the model with all the tasks and then we measure results on an independent test set of the same tasks. In the case of novel task experiments, we train the model with all the tasks minus one (randomly selected). Then, we fix the representation found by our method and we use a subset of the data (70%) for the excluded task to train the last layer, maintaining fixed the representation layer. Finally, we used the remaining data (30%) of the novel task as test set, measuring both error and fairness measure.

We repeated all the experiments both with and without the sensitive feature in the functional form of the model. We validated the hyperparameters using a grid search with $\lambda \in \{10^{-6.0}, 10^{-5.8}, \cdots, 10^{+4.0}\}$ and $r \in \{2^j d \mid j = -4, -3, \ldots, 10\}$, following the validation procedure in [11]. Specifically, in the first step, the classical 10-fold CV error for each of the combination of the hyperparameters is computed. In the second step, we shortlist all the hyperparameters' combinations with error close to the best one (in our case, above 90% of the smallest error). Finally, from this list, we select the hyperparameters with the smallest fairness risk. Concerning the error (ERR) we used mean average precision error as the performance index, and concerning the fairness of our model (FAIR), we compute the the difference of demographic parity as $\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} |P(f(x) = y|s = 1) - P(f(x) = y|s = 2)|$, since in our datasets the output space is finite. For all the experiments, we report performance over 30 repetitions with the corresponding standard deviation.

Datasets. In our comparisons we used three datasets. The first one is the School data set [15] – made available by the Inner London Education Authority (ILEA) – formed by examination records from 139 secondary schools in years 1985, 1986 and 1987. It is a random 50% sample with 15362 students. Each task in this setting is to predict exam scores for students in one school, based on eight inputs. The first four inputs (year of the exam, gender, VR band and ethnic group) are student-dependent, the next four (percentage of students eligible for free school meals, percentage of students in VR band one, school gender (mixed or single-gender and school denomination) are school-dependent. The categorical variables (year, ethnic group and school denomination) were split up in one-hot variables, one for each category, making a new total of 16 student-dependent inputs, and six school-dependent inputs. We scaled each covariate and output to have zero mean and unit variance. The sensitive attribute is the gender of the student. The second dataset we propose has

Table 1: Comparison between the following method: (a) Unconstrained single task learning (STL), (b) Fair constrained STL, (c) Unconstrained MTL, (d) Fair constrained MTL, that is the proposed method.

		STL - UnCons		STL - Cons		MTL - UnCons		MTL - Cons	
	Dataset	ERR	FAIR	ERR	FAIR	ERR	FAIR	ERR	FAIR
Same Tasks		Sensitive feature not in the functional form of the model							
	School	15.30±0.60	0.110±0.005	16.37±0.34	0.044±0.003	10.71±0.57	0.077±0.003	11.78±0.75	0.011±0.000
	UNIGE	19.50±0.94	0.100±0.006	20.87±1.16	0.040±0.002	13.65±0.47	0.070±0.003	15.02±0.54	0.010±0.001
	Movielens	30.30±1.98	0.160±0.008	32.42±1.14	0.048±0.002	15.15±0.60	0.112±0.008	17.27±0.76	0.000±0.000
		Sensitive feature in the functional form of the model							
	School	14.23±0.70	0.118±0.006	15.30±0.81	0.052±0.003	9.64±0.40	0.085±0.004	10.71±0.52	0.019±0.001
	UNIGE	18.13±0.83	0.107±0.005	19.50±0.71	0.047±0.003	12.29±0.67	0.077±0.004	13.65±0.82	0.017±0.001
	Movielens	28.18±1.35	0.171±0.010	30.30±1.28	0.059±0.002	13.03±0.47	0.123±0.007	15.15±0.73	0.011±0.001
New Tasks		Sensitive feature not in the functional form of the model							
	School	18.36±1.12	0.121±0.007	19.43±0.80	0.055±0.003	13.77±0.52	0.088±0.003	14.84±0.74	0.022±0.001
	UNIGE	21.45±1.16	0.105±0.006	22.82±1.22	0.045±0.002	15.60±0.83	0.075±0.003	16.97±0.70	0.015±0.001
	Movielens	33.33±2.14	0.176±0.009	35.45±1.84	0.064±0.004	18.18±0.76	0.128±0.007	20.30±1.18	0.016±0.001
		Sensitive feature in the functional form of the model							
	School	17.29±0.73	0.129±0.007	18.36±0.88	0.063±0.004	12.70±0.50	0.096±0.005	13.77±0.76	0.030±0.002
	UNIGE	20.08±1.21	0.112±0.005	21.45±1.04	0.052±0.002	14.23±0.67	0.082±0.001	15.60±0.61	0.022±0.001
	Movielens	31.21±1.63	0.187±0.007	33.33±1.28	0.075±0.004	16.06±0.92	0.139±0.011	18.18±0.79	0.027±0.001

been collected at the University of Genoa[4] (UNIGE) and is also exploited in [26]. This dataset is a proprietary and highly sensitive dataset containing all the data about the past and present students enrolled at the UNIGE. In this study we take into consideration students who enrolled, in the academic year (a.y.) 2017-2018. The dataset contains 5000 instances, each one described by 35 attributes (both numeric and categorical) about ethnicity, gender, financial status, and previous school experience. The scope is to predict the grades at the end of the first semester being fair with respect to the gender of the student. The sensitive attribute is the gender of the student. Finally, the third dataset is Movielens [17]. Specifically, we considered Movielens 100k (ml100k), which consists of ratings (1 to 5) provided by 943 users for a set of 1682 movies, with a total of 100,000 ratings available. Additional features for each movie, such as the year of release or its genre, are provided. The sensitive attribute is the gender of the user.

Discussion. From our experimental results, different interesting aspects and comparisons can be extracted. Firstly, the results in Table 1 confirm the benefit of using a MTL approach in comparison to STL, in that accuracy has a significant improvement, both on same and novel tasks, thanks to the shared representation. Achieving less error has the positive side effect of producing a more fair model, even in the unconstrained case (i.e. fair unaware).

In the case of constrained methods, learning a fair shared representation slightly increases the final error but brings a large decrease of the fairness measure. From Table 1, we observe that this benefit is maintained also by tackling new and unseen (during the training of the shared representation) tasks. In this sense, our method (constrained MTL) obtains the best performance among all the others.

In general, the same analysis of the results applies to both having and not having the sensitive

[4]The data and the research are related to the project DROP@UNIGE of the University of Genoa.

Table 2: Comparison of the following methods: (M1) Standard MTL with the fairness constraints on the outputs, (M2) FFNN with linear activation and the fair shared representation method presented in [22], (M3) FFNN with fair shared representation [12], (M4) Fair constrained MTL (Our Method).

		M1		M2		M3		M4 (OURS)	
	Dataset	ERR	FAIR	ERR	FAIR	ERR	FAIR	ERR	FAIR
Same Tasks	Sensitive feature not in the functional form of the model								
	School	12.34±0.75	0.013±0.001	13.44±1.04	0.017±0.002	12.93±0.79	0.018±0.002	11.78±0.75	0.011±0.000
	UNIGE	18.12±0.98	0.012±0.001	21.23±1.34	0.021±0.004	26.19±1.76	0.027±0.004	15.02±0.54	0.010±0.001
	Movielens	17.12±0.65	0.009±0.000	19.21±0.87	0.014±0.002	18.01±0.76	0.012±0.002	17.27±0.76	0.007±0.000
	Sensitive feature in the functional form of the model								
	School	11.01±0.91	0.020±0.001	12.01±1.01	0.022±0.002	13.31±1.23	0.025±0.002	10.71±0.52	0.019±0.001
	UNIGE	13.75±0.82	0.017±0.001	20.13±1.24	0.029±0.005	25.92±1.76	0.032±0.006	13.65±0.82	0.017±0.001
	Movielens	15.65±0.73	0.010±0.001	18.97±0.67	0.017±0.004	17.11±0.78	0.015±0.003	15.15±0.73	0.011±0.001
New Tasks	Sensitive feature not in the functional form of the model								
	School	15.64±0.79	0.032±0.002	16.43±1.11	0.044±0.004	17.21±1.32	0.041±0.004	14.84±0.74	0.022±0.001
	UNIGE	16.21±0.97	0.021±0.002	21.98±1.47	0.029±0.004	27.31±1.23	0.033±0.005	16.97±0.70	0.015±0.001
	Movielens	19.20±1.35	0.025±0.002	21.21±1.35	0.031±0.004	20.12±1.43	0.030±0.003	20.30±1.18	0.016±0.001
	Sensitive feature in the functional form of the model								
	School	14.72±0.87	0.038±0.002	18.02±1.07	0.042±0.003	17.92±0.87	0.056±0.003	13.77±0.76	0.030±0.002
	UNIGE	15.89±0.68	0.029±0.002	19.21±1.04	0.035±0.005	25.87±1.23	0.038±0.006	15.60±0.61	0.022±0.001
	Movielens	19.98±0.74	0.038±0.002	20.12±1.12	0.037±0.003	19.93±1.53	0.038±0.004	18.18±0.79	0.027±0.001

feature in the functional form of the model. In order to better interpret our results, and due to our higher interest in the case of a fair constrained model without the sensitive feature in the functional form of the model, we compared in Figure 1 the constrained STL approach to the constrained MTL approach (our method) both on the same and the novel tasks. In this figure it is easier to note the benefits of our algorithm in decreasing both the error and the fairness measure.

Finally, we compared our method with three different state-of-the-art methods. In Table 2 and Figure 2, we show these results. We note how our method, in all the possible settings, obtains better or comparable performance. In fact, it is able to maintain a larger accuracy (comparable to the other methods) and simultaneously a smaller fairness risk.

6 Conclusion

We have presented a method to learn a fair shared representation among different tasks in a MTL setting. Our method is able to provide good generalization performance both in accuracy and fairness over novel and unseen tasks. We studied the learning ability of our method in theory and we analyzed the performance over several experimental scenarios in practice. The obtained results corroborate our theoretical findings and proved that our approach overcomes common benchmark algorithms and current state-of-the-art methods. Our next step will be to study (explicit) fair representation learning in the context of shallow and deep neural networks, basically a generalization to the non-linear case of the proposed approach, with particular attention to the interpretability of the learned representation, in the context of transparency and trust of the machine learning model.

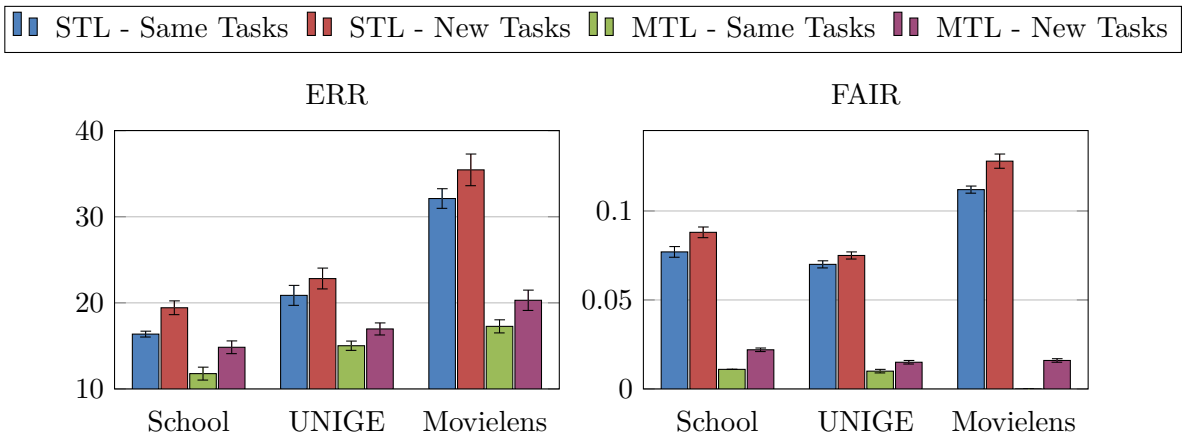

Figure 1: Graphical representation of the results in Table 1, when the sensitive feature is not included in the functional form of the model and the fairness constraint is active.

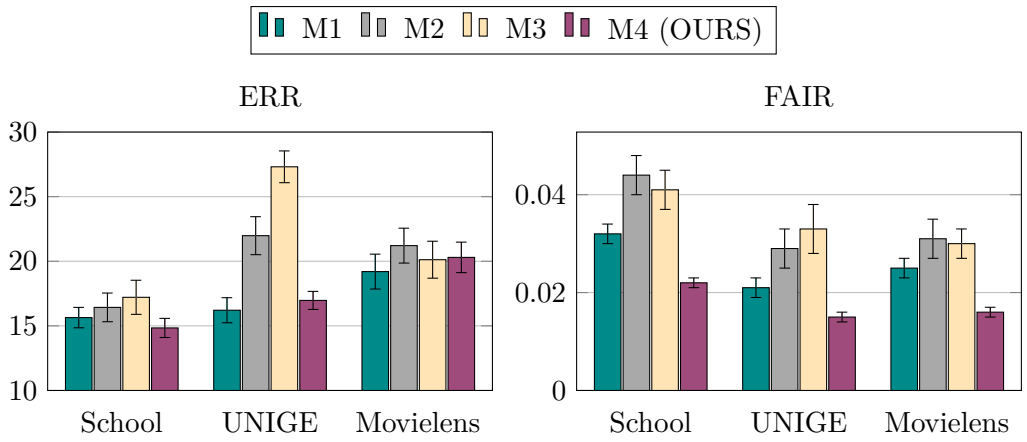

Figure 2: Graphical representation of the results in Table 2 for new tasks when the sensitive feature is not included in the functional form of the model.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, 2018.

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

[3] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

[4] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(149-198):3, 2000.

[5] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[6] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 2018.

[7] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE international conference on Data mining*, 2009.

[8] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 2017.

[9] C. Ciliberto, D. Stamos, and M. Pontil. Reexamining low rank matrix factorization for trace norm regularization. *arXiv preprint arXiv:1706.08934*, 2017.

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, 2014.

[11] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018.

[12] H. Edwards and A. Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[14] P. Gajane and M. Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

[15] H. Goldstein. Multilevel modelling of survey data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40(2):235–244, 1991.

[16] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 2016.

[17] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19, 2016.

[18] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, 2016.

[19] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012.

[20] C. Lan and J. Huan. Discriminatory transfer. *arXiv preprint arXiv:1707.00780*, 2017.

[21] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[22] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[23] A. Maurer. Transfer bounds for linear feature learning. *Machine learning*, 75(3):327–350, 2009.

[24] D. McNamara, C. S. Ong, and R. C. Williamson. Provably fair representations. *arXiv preprint arXiv:1710.04394*, 2017.

[25] D. McNamara, C. Soon Ong, and B. Williamson. Costs and benefits of fair representation learning. In *AAAI Conference on Artificial Intelligence, Ethics and Society*, 2019.

[26] L. Oneto, M. Donini, and M. Pontil. General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080*, 2019.

[27] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

[28] M. Pontil and A. Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, 2013.

[29] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[30] I. Deborah Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conference on AI Ethics and Society*, 2019.

[31] N. Srebro. Learning with matrix factorizations. *PhD thesis, Massachusetts Institute of Technology*, 2004.

[32] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[33] S. Verma and J. Rubin. Fairness definitions explained. In *IEEE/ACM International Workshop on Software Fairness*, 2018.

[34] Y. Wang, T. Koike-Akino, and D. Erdogmus. Invariant representations from adversarially censored autoencoders. *arXiv preprint arXiv:1805.08097*, 2018.

[35] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017.

[36] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[37] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

Appendix

In this appendix, we first collect some tools used in our analysis. We then present an improvement of Theorem 2 in [23], which is instrumental in the proof of Theorem 1, and explain how to pass to a fully data dependent bound.

A Matrix concentration inequalities

Theorem 2 (Part (i) of Theorem 7 of [28])**.** *Let $A_1, \dots, A_N$ be independent random operators on a Hilbert space satisfying $0 \preceq A_k \preceq I$ and suppose that for some $d \in \mathbb{N}$*

$$\dim \operatorname{Span}\left(\operatorname{Ran}\left(A_1\right), \dots, \operatorname{Ran}\left(A_N\right)\right) \leq d \tag{9}$$

almost surely. Then

$$\Pr\left\{\left\|\sum_k \left(A_k - \mathbb{E}A_k\right)\right\|_\infty > s\right\} \leq 4d^2 \exp\left(\frac{-s^2}{9\left\|\sum_k \mathbb{E}A_k\right\|_\infty + 6s}\right). \tag{10}$$

Corollary 3. *Under the conditions of the Theorem we have (i) with probability at least $1-\delta$ that*

$$\sqrt{\left\|\mathbb{E}\sum A_k\right\|_\infty} \leq \sqrt{\left\|\sum A_k\right\|_\infty} + 6\sqrt{\ln\left(\frac{4d^2}{\delta}\right)}, \tag{11}$$

and (ii) with the same bound on the probability

$$\left\|\sum_k \left(A_k - \mathbb{E}A_k\right)\right\|_\infty \leq 3\sqrt{\left\|\sum A_k\right\|_\infty \ln\frac{8d^2}{\delta}} + 24\ln\frac{8d^2}{\delta} \tag{12}$$

Proof. Equating the RHS of (10) to δ we obtain with probability at least $1-\delta$ that

$$\left\|\sum_k \left(A_k - \mathbb{E}A_k\right)\right\|_\infty \leq \sqrt{9\left\|\sum_k \mathbb{E}A_k\right\|_\infty \ln\frac{4d^2}{\delta}} + 6\ln\frac{4d^2}{\delta} \tag{13}$$

Denoting for brevity $a := \left\|\sum_k \mathbb{E}A_k\right\|_\infty$, $b := \left\|\sum_k A_k\right\|_\infty$ and $c = 4d^2$ we have with probability at least $1-\delta$

$$a - b \leq 2\sqrt{a}\sqrt{\frac{9}{4}\ln\frac{c}{\delta}} + 6\ln\frac{c}{\delta},$$

or, subtracting $2\sqrt{a}\sqrt{\frac{9}{4}\ln\frac{c}{\delta}}$ and adding $b + \frac{9}{4}\ln\frac{c}{\delta}$,

$$\begin{aligned}\left(\sqrt{a} - \sqrt{\frac{9}{4}\ln\frac{c}{\delta}}\right)^2 &= a - 2\sqrt{a}\sqrt{\frac{9}{4}\ln\frac{c}{\delta}} + \frac{9}{4}\ln\frac{c}{\delta} \\ &\leq b + \frac{9}{4}\ln\frac{c}{\delta} + 6\ln\left(c/\delta\right).\end{aligned}$$

Taking the positive squareroot and adding $\sqrt{\frac{9}{4}\ln\frac{c}{\delta}}$ gives

$$\sqrt{a} \leq \sqrt{\frac{9}{4}\ln\frac{c}{\delta}} + \sqrt{b + \frac{9}{4}\ln\frac{c}{\delta} + 6\ln\frac{c}{\delta}}$$

$$\leq \quad \sqrt{b} + \sqrt{9 \ln \frac{c}{\delta}} + \sqrt{6 \ln \frac{c}{\delta}} \leq \sqrt{b} + 6\sqrt{\ln \frac{c}{\delta}},$$

which is (11). Part (ii) then follows from a union bound of (13) with (11). ■

The second matrix concentration inequality we need is

Theorem 4 (Theorem 1.5 in [32]). *Let $A_1, \ldots, A_N$ be fixed matrices with dimension $d_1 \times d_2$ and $\gamma_1, ..., \gamma_N$ independent standard normal variables. Let σ^2 be the variance patrameter*

$$\sigma^2 = \max \left\{ \left\| \sum_k A_k \right\|_\infty , \left\| \sum_k A_k^* \right\|_\infty \right\}.$$

Then for $s > 0$

$$\Pr \left\{ \left\| \sum_k \gamma_k A_k \right\|_\infty > s \right\} \leq (d_1 + d_2)\, e^{-s^2/(2\sigma^2)}.$$

Corollary 5. *Under above assumptions, if $d_1 + d_2 \geq 3$ then*

$$\mathbb{E} \left\| \sum_k \gamma_k A_k \right\|_\infty \leq \frac{5}{2} \sqrt{\sigma^2 \ln (d_1 + d_2)}.$$

Proof. Let $\delta = \sqrt{2\sigma^2 \ln (d_1 + d_2)}$. Integration by parts gives

$$\begin{aligned} \mathbb{E} \left\| \sum_k \gamma_k A_k \right\|_\infty &\leq \delta + (d_1 + d_2) \int_\delta^\infty e^{-t^2/2\sigma^2} dt \\ &\leq \delta + (d_1 + d_2) \frac{\sigma^2}{\delta} \exp\left(\frac{-\delta^2}{2\sigma^2} \right) \\ &\leq \frac{5}{2} \sqrt{\sigma^2 \ln (d_1 + d_2)}. \end{aligned}$$

■

B Improved bound for the transfer risk

We give an improvement of Theorem 2 in [23] for the case of trace-norm regularization.

Most of the notation, definitions and assumptions are taken from [23], except that here we denote by T the number of tasks and let $t \in \{1, \ldots, T\}$ be the task index; these correspond to n and $l \in \{1, \ldots, n\}$ in [23]. We used this notation because it is common in the multitask literature. Our results are dimension free, so they hold in the general case that the input space is Hilbert space $\mathbb{H}$ and D a bounded linear operator on $\mathbb{H}$. However to simplify the presentation here we take $\mathbb{H} = \mathbb{R}^d$ and D a $d \times d$ PSD matrix. We also use $R_\mu(w)$ and $R_{\mathbf{z}}(w)$ as the expected and empirical error of a weight vector w, that is

$$\mathcal{E}_\mu(w) = \mathbb{E}_{(x,y)\sim\mu}[\ell(\langle w, x\rangle, y)], \qquad \mathcal{E}_{\mathbf{z}}(w) = \frac{1}{m} \sum_{i=1}^m \ell(\langle w, x_i\rangle, y_i)$$

For every PSD matrix D we define the following quantities (see [23, Secs. 2.2 & 2.3])

$$w(\mathbf{x},\mathbf{y}) = \operatorname*{argmin}_{w\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \ell(\langle w, x_i\rangle, y_i) + \|w\|^2, \qquad w_D(\mathbf{x},\mathbf{y}) = D^{\frac{1}{2}} w(D^{\frac{1}{2}}\mathbf{x},\mathbf{y}).$$

Note that the vector $w_D(\mathbf{x},\mathbf{y})$ corresponds to the minimizer of ridge regression with modified regularizer,

$$w_D(\mathbf{x},\mathbf{y}) = \operatorname*{argmin}_{w\in \mathrm{Ran}(D)} \frac{1}{m}\sum_{i=1}^{m} \ell(\langle w, x_i\rangle, y_i) + w^\top D^+ w$$

where D^+ is the pseudo-inverse of D.

Theorem 6. *Let $\mathcal{D}$ a subset of $d\times d$ PSD matrices. Suppose the algorithm w is 1-bounded and has kernel stability L relative to the loss function ℓ and that for every $K<\infty$ there exists $M(K)$ such that for all $y\in[0,1]$ and for all $s,t\in[-K,K]$ we have*

$$\ell(s,y) - \ell(t,y) \le M(K)\,|s-t|\,. \tag{14}$$

Then for every $\delta>0$, with probability greater $1-\delta$ in the data $(\mathbf{X},\mathbf{Y})\sim\hat{\rho}^T$ we have for all $D\in\mathcal{D}$

$$\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{\mathbf{z}\sim\mu^m}\mathcal{E}_\mu(w_D(\mathbf{z})) \le \frac{1}{T}\sum_{t=1}^{T}\mathcal{E}_{\mathbf{z}_t}(w_D(\mathbf{z}_t)) + 2M\left(\|D\|_\infty^{1/2}\right)\sqrt{\frac{\|D\|_1\,\|C\|_\infty}{m}}$$
$$+ 7L\max_{D\in\mathcal{D}}\{\mathrm{tr}D\}\sqrt{\frac{\ln(2mT)\,\|\hat{C}\|_\infty}{T}} + \sqrt{\frac{\ln(1/\delta)}{2T}},$$

where and C and $\hat{C}$ are respectively the true and empirical total covariance operator for the data.

The proof uses the following theorem to bound the Gaussian complexity in the exactly the same way as Theorem 7 is used to prove Theorem 2 in [23].

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

$$f(\mathbf{x},\mathbf{y}) - f(\mathbf{x}',\mathbf{y}) \le \frac{L}{m}\left\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}')\right\|_{Fr},$$

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. Fix a meta-sample $(\mathbf{X},\mathbf{Y}) = ((\mathbf{x}^1,\mathbf{y}^1),...,(\mathbf{x}^n,\mathbf{y}^n))$. 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

$$\Gamma(\mathcal{F},(\mathbf{X},\mathbf{Y})) = \frac{2}{T}\mathbb{E}_\gamma\sup_{f\in\mathcal{F}}\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}_t) \le 5L\,\|\mathcal{D}\|_1\sqrt{\frac{\ln(2mT)\left\|\hat{C}\right\|_\infty}{T}}.$$

Proof. The proof follows exactly the proof of Theorem 7 in [23] up to the statement of the following inequality (in [23] this is equation (6))

$$\Gamma(\mathcal{F},(\mathbf{X},\mathbf{Y})) \le \frac{2}{T}\mathbb{E}_\gamma\sup_{D\in\mathcal{D}}\frac{L}{m}\sum_{t=1}^{T}\sum_{i,j=1}^{m}\gamma_{ij}^t\left\langle x_i^t, Dx_j^t\right\rangle. \tag{15}$$

Define an operator J_{ij}^t on H by $J_{ij}^t z = \langle z, x_i^t \rangle x_j^t$. Then by Hölder's inequality

$$\sum_{t=1}^{T}\sum_{i,j=1}^{m}\gamma_{ij}^t\left\langle x_i^t, Dx_j^t\right\rangle = \left\langle \sum_{t=1}^{T}\sum_{i,j=1}^{m}\gamma_{ij}^t J_{ij}^t, D\right\rangle_2 \leq \left\|\sum_{t=1}^{T}\sum_{i,j=1}^{m}\gamma_{ij}^t J_{ij}^t\right\|_\infty \|D\|_1 .$$

Now

$$\begin{aligned}\left\|\sum_{t=1}^{T}\sum_{i,j=1}^{m}J_{ij}^{t*}J_{ij}^t\right\|_\infty &= \sup_{\|u\|\leq 1,\|v\|\leq 1}\sum_{t=1}^{T}\sum_{i,j=1}^{m}\left\langle J_{ij}^t u, J_{ij}^t v\right\rangle \\ &= \sup_{u,v}\sum_{t=1}^{T}\sum_{i,j=1}^{m}\left\langle u, x_i^t\right\rangle\left\langle v, x_i^t\right\rangle\left\|x_j^t\right\|^2 \\ &\leq m\sup_{u,v}\sum_{t=1}^{T}\sum_{i=1}^{m}\left\langle u, x_i^t\right\rangle\left\langle v, x_i^t\right\rangle = m^2T\left\|\hat{C}\right\|_\infty .\end{aligned}$$

The same bound holds for the norm of the adjoint. All the J_{ij}^t operate on the mT-dimensional subspace generated by the x_i^t, so by the corollary to the Olivera-Tropp inequality (Corollary 5) with $d_1 = d_2 = mT$ and $\sigma^2 = m^2T\left\|\hat{C}\right\|_\infty$

$$\mathbb{E}\left\|\sum_{t=1}^{T}\sum_{i,j=1}^{m}\gamma_{ij}^t J_{ij}^t\right\| \leq \frac{5}{2}\sqrt{m^2T\left\|\hat{C}\right\|_\infty \ln(2mT)}.$$

Division by mT and (15) give the conclusion. ■