Learning Fair and Transferable Representations with Theoretical Guarantees

Luca Oneto, *Member, IEEE*, Michele Donini, Massimiliano Pontil, and Andreas Maurer

***Abstract*— Developing learning methods which do not discriminate subgroups in the population is the central goal of algorithmic fairness. One way to reach this goal is by modifying the data representation in order to satisfy prescribed fairness constraints. This allows to reuse the same representation in other context (tasks) without discriminate subgroups. In this work we measure fairness according to demographic parity, requiring 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 multi-task 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.**

***Index Terms*— Algorithmic Fairness, Learning Fair Representation, Demographic Parity, Multi-Task Learning, Transfer Representation**

I. Introduction

During the last decade, the widespread distribution of automatic systems for decision making is raising concerns about their potential for unfair behaviour [1]–[4]. 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) [5]–[8]; (ii) in the second family, we can find methods that enforce fairness directly during the training phase, e.g. [9]–[12]; (iii) the third family of methods implements fairness by modifying the data representation, and then employs standard machine learning methods [13], [14].

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 [15], 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, a phenomenon which is referred to as discriminatory transfer in [16] or negative legacy in [17], 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 multi-task learning. We aim to leverage task similarities in order to learn a fair representation that 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 [18] (for an extended analysis of the different fairness definitions see [9], [19]) that requires the probability of possible model decisions to be independent of the sensitive information. We argue that multi-task 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 well to novel tasks both in terms of prediction performance and fairness metrics. Other papers in literature already pursued a similar goal [20]–[26]. 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 multi-task 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 multi-task 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 [27] and by being fully data dependent, it can be used to evaluate the transfer capability of the learned representation.

Luca Oneto - University of Genoa, Italy (email: luca.oneto@unige.it). Michele Donini - Amazon Web Services, US (email: donini@amazon.com). Massimiliano Pontil, Istituto Italiano di Teconologia & University College London, Italy (email: massimiliano.pontil@iit.it). Andreas Maurer - Self Employed, Germany & Italy (email: am@andreas-maurer.eu).

[1]See for example the Caffe Model Zoo: `github.com/BVLC/caffe/wiki/Model-Zoo`

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

II. 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 [13], [20]–[26], [28] – with fair representation we refer to the concept of learning a function g, which does not discriminate subgroups in the data. This approach is different from most commonly used approaches [7], [10], [11], 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 papers 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 [20]–[26], the authors propose different neural network 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 [28], 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 [13], 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.

Contrary to the described methods from literature, our proposal enforces the fairness constraint directly on the representation layer, i.e. not using indirect approaches such as different objective functions for the entire functional or adversarial methods. Moreover we support the proposed method with generalization guarantees in terms of both accuracy and fairness measure.

III. Method

In this section, we present our method to learn a shared fair representation from multiple tasks. We consider T supervised learning tasks (i.e. binary classification or regression problems). 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 $\mathcal{X} = \mathbb{R}^d$, the functions f_t are linear, and to simplify the presentation we do not include s in the functional form of the model, 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 multi-task learning formulation (MTL) is based on minimizing the multi-task 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,

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

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. [30] 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 [31], see e.g. [32] and references therein[3]. We follow the formulation

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

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

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 [19], [33] on each task. This means that, for every measurable subset $C \subset \mathbb{R}^r$, and for every $t \in \{1, \ldots, T\}$, 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 $t \in \{1, \ldots, T\}$, that both distributions have the same mean. Furthermore, we compute the means from empirical data. For each training sequence $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^m$ 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 distributions 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 [34] in the case of fair regression, and reported to be empirically effective. Furthermore, in Appendix VIII-A, we discuss the quality of this approximation referring also the previous results in the literature [10], [34].

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

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

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

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

which can be easily solved. In particular note that the problem decouples across the tasks, and each task specific problem amounts to run 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

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

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 whole 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, i.e. 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.

IV. LEARNING BOUND

The intuition behind MTL and in particular the method presented above is that MTL works because of the effective increase in sample size and correspondingly improved generalization through simultaneous consideration of the samples of many related tasks. In this section, we rigorously study the learning ability of the proposed method. We consider the setting of learning-to-learn [35] (a.k.a. meta-learning), 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 [27] 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. We extend this

analysis to the setting of algorithmic fairness, in which the performance of the algorithm is evaluated both relative to the risk and the fairness constraint. We show that both quantities can be bounded by their empirical counterparts evaluated on the training samples.

To present our result we introduce some more notation[4]. We let $\mathcal{R}_\mu(w)$ and $\mathcal{R}_{\mathbf{z}}(w)$ be the expected and empirical errors of a weight vector w, that is

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

Throughout the paper sometimes we also refer to the expected error as the risk. Furthermore, for every matrix $A \in \mathbb{R}^{d\times r}$ and for every data sample $\mathbf{z} = (x_i, y_i)_{i=1}^m$, we define $b_A(\mathbf{z}) = \operatorname{argmin}_{b\in\mathbb{R}^r} \frac{1}{m}\sum_{i=1}^m (y_i - \langle b, A^\top x_i\rangle)^2 + \lambda\|b\|^2$ as the minimizer of ridge regression with modified data representation.

In the statistical learning-to-learn setting the tasks $\mu_1, \ldots, \mu_T$ are independently sampled from a meta-distribution ρ on the set of probability distributions on $\mathcal{X}\times\mathcal{Y}$. In turn, for every task $t \in \{1, \ldots, T\}$, we are provided with a training dataset $\mathbf{z}_t$ of m points sampled independently from μ_t. We also require the following assumption, which is standard in the learning-to-learn literature.

Assumption 1: We 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 $\hat{C}$ be the total (uncentered) empirical covariance of the input points of all tasks,

$$\hat{C} = \frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m x_{t,i}\otimes x_{t,i}.$$

Notice that when the raw input is intrinsically high dimensional (hence learning is difficult without representation learning), the spectrum of $\hat{C}$ tends to be flat. In particular if the data are uniformly distributed on the unit d-dimensional sphere then with high probability $\|\hat{C}\|_\infty \le 1/d + O(1/Tm)$, where $\|\cdot\|_\infty$ is the spectral norm (largest singular value) of a matrix. We also define the empirical covariance of the difference of empirical means of the sensitive groups,

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^T c(\mathbf{z}_t)\otimes c(\mathbf{z}_t).$$

Our main result, presented in the theorem below, bounds with high probability the expected risk (transfer risk) and the expected violation of the fairness constraint for a new task drawn from the random environment, and gives a theoretical justification of the method used in the paper.

Theorem 1: Let A be the representation learned by solving Problem (1) and renormalized so that $\|A\|_F = 1$. 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

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

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

To prove the second claim we note that

$$\frac{1}{T}\sum_{t=1}^T \|Ac(\mathbf{z}_t)\|^2 = \operatorname{tr} D\hat{\Sigma}$$

and similarly

$$\mathbb{E}_{\mu\sim\rho}\mathbb{E}_{\mathbf{z}\sim\mu^m}\|Ac(\mathbf{z})\|^2 = \operatorname{tr} D\Sigma$$

where Σ is the true covariance of the difference of the sensitive groups means,

$$\Sigma = \mathbb{E}_{\mu\sim\rho}\mathbb{E}_{\mathbf{z}\sim\mu^m} c(\mathbf{z})\otimes c(\mathbf{z}).$$

Then

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

The second inequality then follows immediately from inequality (10) in Corollary 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 [27], due to the introduction of the empirical total covariance in the third term in the RHS of the inequality. The result in [27] instead contains the term $\sqrt{1/T}$, which can be considerably larger when the raw input is distributed on a high dimensional manifold.

[4]See also suplementary material here `https://www.dropbox.com/s/rt5q1buluv4l7wo/Sup.pdf`

2) The bounds in Theorem 1 can be extended to hold with variable sample size per task. However, 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 the 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) Finally, we note that although the theorem provides a bound for the linear approximation of demographic parity, we may pass to a bound on demographic parity following the reasoning in [10]. This is discussed in detail in Appendix VIII-A, where we also address the quality of the above approximation.

V. Experiments

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

A. Settings

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

In the first set (Table I) we compare the following methods: (a) Unconstrained single task learning[5] (STL), (b) Fair constrained STL (i.e. STL with additional demographic parity constraint), (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 II) 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 [23] (M2), (c) FFNN with linear activation by exploiting a fair shared representation as presented in [21] (M3), (d) Fair constrained MTL (Our Method). We used linear activation functions in FFNN for fair comparison, since the proposed method learns linear models. It is important to note that, for the sake of completeness, we performed all same experiments also in the non-linear case, namely the case when non-linear (sigmoid) activation function is exploited. The complete set of results for the non-linear case can be found in the supplementary material, Appendix VIII-B.

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 [10]. 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 measure. Concerning the error (ERR) we used mean average precision error as the performance index, and concerning the unfairness of our model (UNFAIR), 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.

B. Datasets

In our comparisons we used three datasets. The first one is the School data set [36] – 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 been collected at the University of Genoa[6] (UNIV) and is also exploited in [34]. This dataset is a proprietary and highly sensitive

[5]STL is equivalent to learning the tasks independently with no representation constraint.

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

TABLE I: Feed Forward Single Layered Neural Network with linear activation functions. (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	UNFAIR	ERR	UNFAIR	ERR	UNFAIR	ERR	UNFAIR
		Sensitive feature not in the functional form of the model							
Same Tasks	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±0.001
	UNIV	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	0.001±0.001
		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
	UNIV	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
		Sensitive feature not in the functional form of the model							
New Tasks	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
	UNIV	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
	UNIV	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

TABLE II: Feed Forward Single Layered Neural Network with linear activation functions. Comparison of the following methods: (M1) Standard MTL with the fairness constraints on the outputs, (M2) FFNN with the fair shared representation method presented in [23], (M3) FFNN with the fair shared representation method presented in [21], (M4) Fair constrained MTL (Our Method).

		M1		M2		M3		M4 (OURS)	
	Dataset	ERR	UNFAIR	ERR	UNFAIR	ERR	UNFAIR	ERR	UNFAIR
		Sensitive feature not in the functional form of the model							
Same Tasks	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±0.001
	UNIV	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±0.001	19.21±0.87	0.014±0.002	18.01±0.76	0.012±0.002	17.27±0.76	0.007±0.001
		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
	UNIV	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
		Sensitive feature not in the functional form of the model							
New Tasks	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
	UNIV	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
	UNIV	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

dataset containing all the data about the past and present students enrolled at the UNIV. 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. Finally, the third dataset is Movielens [37]. 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. In this case, the sensitive attribute is the gender of the user.

VI. Discussion

From our experimental results, different interesting aspects and comparisons can be extracted. Firstly, the results in Table I confirm the benefit of using a MTL approach in comparison

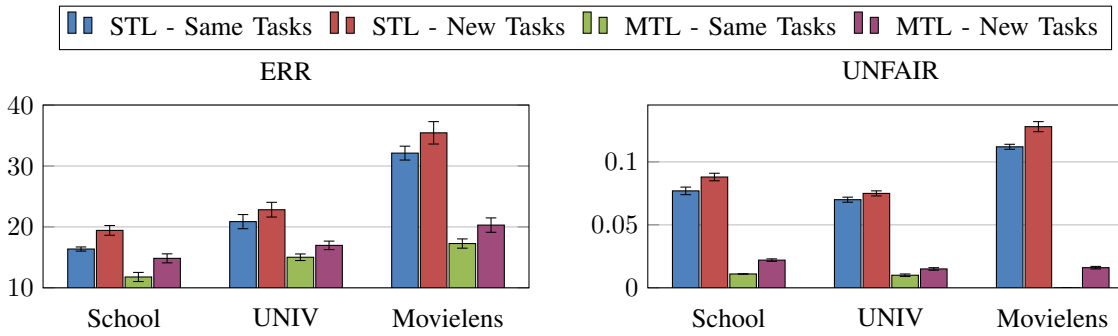

Fig. 1: Graphical representation of the results in Table I, when the sensitive feature is not included in the functional form of the model and the fairness constraint is active.

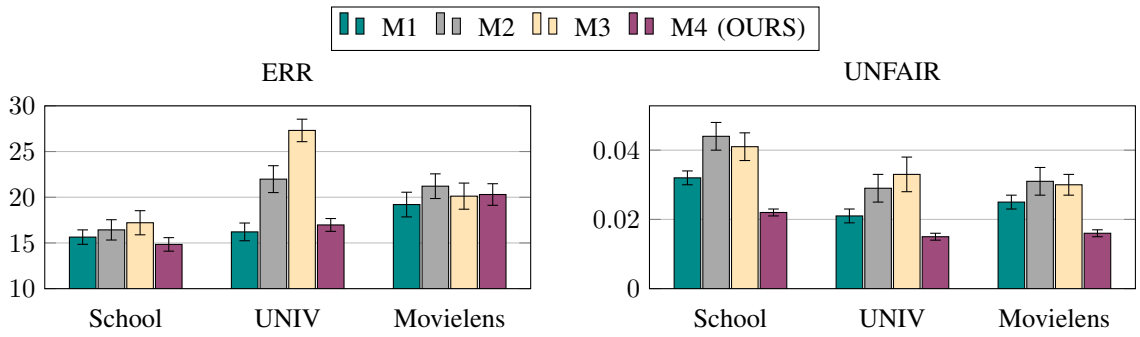

Fig. 2: Graphical representation of the results in Table II for new tasks when the sensitive feature is not included in the functional form of the model.

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 I, 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 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 II 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 measure.

Analogous conclusions can be drawn for the non-linear scenario, where our method outperforms all the other ones in all the experiments except for M3 [23] on the UNIV dataset. Due to space constraint, the complete results are reported in Appendix VIII-B. We believe our method is able to achieve these results enforcing the fairness constraint directly on the representation layer, instead of using different objective functions for the entire functional, or with adversarial methods, and so indirectly. This is a key difference in our approach compared to the current state-of-the-art ones. Moreover, theoretical guarantees behind the generalization performance of our method, in terms of accuracy and fairness, also support our findings and promising results.

VII. Conclusions

We presented a method to learn a fair shared representation among different tasks in a MTL setting. Within a meta-learning setting, we argue theoretically that the learned representation transfers well to novel tasks, both in terms of accuracy and fairness. Up to our knowledge this is the first representation learning method that is supported by learning guarantees. We then studied the learning ability of our method in practice, in a number of experimental scenarios. The obtained results corroborate our theoretical findings and proved that our approach overcomes common benchmark algorithms and current state-of-the-art methods. A valuable direction of future research would be to extend our learning bounds to 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 final model.

VIII. Appendix

A. Approximation of Demographic Parity

The approximation of demographic parity that we employed in the paper corresponds to matching the first order moment, see Equations (2)–(4). As we have shown in Section V and Appendix VIII-B this approximation works well in practice and it allows us to outperform current state-of-the-art approaches. In the the rest of this section, we show one case in which the approximation is crude. Nevertheless, we will also show that the cases where the approximation fails can be easily detected computing an empirical quantity following the same reasoning in [10].

Let us introduce here our example. Exploiting the proposed method, every linear function based on a representation of x defined by Ax, will be orthogonal to $c(\mathbf{z})$ and the mean of the conditional distributions of $\langle b, A^\top x\rangle$ for both sensitive groups will be equal. But this does not guarantee demographic parity for binary classifiers obtained by thresholding $\langle b, A^\top x\rangle$ at 0. To see this let the two conditional distributions of $\langle b, A^\top x\rangle$ be for $t \in (0, 1/2)$

$$p_{s=1} = t\delta_{-1} + (1-t)\,\delta_{t/(1-t)}$$
$$p_{s=2} = t\delta_{1} + (1-t)\,\delta_{-t/(1-t)},$$

where δ_i is the Dirac delta centered in i. Then both means are zero, but

$$\mathbb{P}\left\{\langle b, A^\top x\rangle > 0 | s = 1\right\} = 1 - t$$
$$\neq t = \mathbb{P}\left\{\langle b, A^\top x\rangle > 0 | s = 2\right\}.$$

The closer t is to zero, the more pronounced is the violation of demographic parity. On the other hand suppose that Eq. (2) holds only for all half-spaces C instead of for all measurable sets. This is then equivalent to demographic parity for all classifiers obtained by thresholding linear functions $\langle b, A^\top x\rangle$ at 0. The relaxation of Eq. (4) does not ensure this, as shown by the above example, not even for the linear functions $b_A(\mathbf{z}_t)$ found for the training data.

The case when the approximation is crude can be detected computing an empirical quantity. In fact if

$$\frac{1}{2}\sum_{g\in\{1,2\}} \left|\mathbb{E}\left[\operatorname{sign}\left(\langle b, A^\top x\rangle\right) - \langle b, A^\top x\rangle \;\middle|\; s = g\right]\right| \leq \Delta,$$

then it also holds that

$$\left|\mathbb{P}\left\{\langle b, A^\top x\rangle > 0 | s = 1\right\} - \mathbb{P}\left\{\langle b, A^\top x\rangle > 0 | s = 2\right\}\right|$$
$$\leq \left|\mathbb{E}\left\{\langle b, A^\top x\rangle \,|s = 1\right\} - \mathbb{E}\left\{\langle b, A^\top x\rangle \,|s = 2\right\}\right| + \Delta. \quad (6)$$

In the above mentioned case $\Delta = 1$ which points out problems with the approximation. Note that the statement of Eq. (6) also hold in its empirical counterpart using $\hat{\Delta}$. In many real scenarios, see [10], [34], this $\hat{\Delta}$ has shown to be very small (smaller than 0.05), suggesting that cases like the artificial one depicted above are uncommon.

B. Results in the non-linear case

Tables III and IV report the equivalent results of Tables I and II when, in the representation layer, a sigmoidal non-linear activation functions is added. These new tables represent the non-linear counterpart of the results presented in the main text. As it is possible to see from the results, the same considerations derived from the linear case can be derived for the non-linear one, supporting the proposed method. Notice that the proposed method outperforms all state-of-the-art methods in all experiments except for M3 [23] on the UNIV dataset. The results are also visualized in Figures 3 and 4, which are the counterpart of Figures 1 and 2 for the non-linear case.

Acknowledgements

This work was supported in part by both SAP SE and Amazon Web Services.

References

[1] S. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, p. 671, 2016.

[2] I. D. 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.

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

[4] S. Chiappa, "Path-specific counterfactual fairness," in AAAI Conference on Artificial Intelligence, 2019.

[5] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in Uncertainty in Artificial Intelligence, 2019.

[6] 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.

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

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

[9] 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, vol. 20, no. 75, pp. 1–42, 2019.

[10] 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.

[11] 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.

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

[13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in International Conference on Machine Learning, 2013.

[14] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in Advances in Neural Information Processing Systems, 2017.

[15] 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.

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

TABLE III: Feed Forward Single Layered Neural Network with sigmoidal non-linear activation functions. (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	UNFAIR	ERR	UNFAIR	ERR	UNFAIR	ERR	UNFAIR
		Sensitive feature not in the functional form of the model							
Same Tasks	School	10.30 ± 0.88	0.090 ± 0.008	11.02 ± 0.68	0.036 ± 0.002	7.21 ± 0.27	0.063 ± 0.003	7.93 ± 0.35	0.009 ± 0.001
	UNIV	13.50 ± 0.91	0.910 ± 0.042	14.45 ± 0.89	0.364 ± 0.022	9.45 ± 0.34	0.637 ± 0.032	10.39 ± 0.55	0.091 ± 0.005
	Movielens	15.30 ± 0.43	0.120 ± 0.006	16.37 ± 0.93	0.036 ± 0.001	7.65 ± 0.38	0.084 ± 0.004	8.72 ± 0.38	0.001 ± 0.001
		Sensitive feature in the functional form of the model							
	School	9.58 ± 0.53	0.096 ± 0.005	10.30 ± 0.48	0.042 ± 0.002	6.49 ± 0.32	0.069 ± 0.004	7.21 ± 0.39	0.015 ± 0.001
	UNIV	12.55 ± 0.71	0.974 ± 0.055	13.50 ± 0.77	0.428 ± 0.020	8.50 ± 0.46	0.701 ± 0.019	9.45 ± 0.36	0.155 ± 0.010
	Movielens	14.23 ± 0.77	0.128 ± 0.005	15.30 ± 0.89	0.044 ± 0.003	6.58 ± 0.35	0.092 ± 0.005	7.65 ± 0.29	0.008 ± 0.001
		Sensitive feature not in the functional form of the model							
New Tasks	School	12.36 ± 0.70	0.099 ± 0.006	13.08 ± 0.44	0.045 ± 0.002	9.27 ± 0.42	0.072 ± 0.002	9.99 ± 0.37	0.018 ± 0.001
	UNIV	14.85 ± 0.64	0.956 ± 0.051	15.79 ± 1.03	0.409 ± 0.017	10.80 ± 0.44	0.682 ± 0.031	11.74 ± 0.64	0.136 ± 0.007
	Movielens	16.83 ± 0.64	0.132 ± 0.009	17.90 ± 0.99	0.048 ± 0.003	9.18 ± 0.42	0.096 ± 0.003	10.25 ± 0.39	0.012 ± 0.001
		Sensitive feature in the functional form of the model							
	School	11.64 ± 0.68	0.105 ± 0.005	12.36 ± 0.65	0.051 ± 0.002	8.55 ± 0.24	0.078 ± 0.004	9.27 ± 0.38	0.024 ± 0.001
	UNIV	13.90 ± 0.56	1.019 ± 0.035	14.85 ± 0.69	0.473 ± 0.023	9.85 ± 0.44	0.746 ± 0.035	10.80 ± 0.42	0.200 ± 0.011
	Movielens	15.76 ± 1.17	0.140 ± 0.008	16.83 ± 0.89	0.056 ± 0.003	8.11 ± 0.42	0.104 ± 0.006	9.18 ± 0.44	0.020 ± 0.001

TABLE IV: Feed Forward Single Layered Neural Network with sigmoidal non-linear activation functions. Comparison of the following methods: (M1) Standard MTL with the fairness constraints on the outputs, (M2) FFNN with the fair shared representation method presented in [23], (M3) FFNN with the fair shared representation method presented in [21], (M4) Fair constrained MTL (Our Method).

		M1		M2		M3		M4 (OURS)	
	Dataset	ERR	UNFAIR	ERR	UNFAIR	ERR	UNFAIR	ERR	UNFAIR
		Sensitive feature not in the functional form of the model							
Same Tasks	School	9.97 ± 0.34	0.015 ± 0.001	10.12 ± 0.39	0.016 ± 0.001	9.93 ± 0.37	0.010 ± 0.001	7.93 ± 0.35	0.009 ± 0.001
	UNIV	10.89 ± 0.55	0.091 ± 0.005	11.12 ± 0.55	0.099 ± 0.006	12.43 ± 0.55	0.101 ± 0.007	10.39 ± 0.55	0.091 ± 0.005
	Movielens	9.12 ± 0.38	0.001 ± 0.001	9.38 ± 0.38	0.002 ± 0.002	9.98 ± 0.38	0.003 ± 0.002	8.72 ± 0.38	0.001 ± 0.001
		Sensitive feature in the functional form of the model							
	School	9.41 ± 0.31	0.021 ± 0.001	9.21 ± 0.49	0.019 ± 0.003	8.98 ± 0.44	0.019 ± 0.002	7.21 ± 0.39	0.015 ± 0.001
	UNIV	9.18 ± 0.38	0.120 ± 0.011	10.65 ± 0.41	0.154 ± 0.014	10.31 ± 0.44	0.161 ± 0.013	9.45 ± 0.36	0.155 ± 0.010
	Movielens	8.11 ± 0.34	0.009 ± 0.001	7.02 ± 0.21	0.007 ± 0.001	7.34 ± 0.23	0.008 ± 0.001	7.65 ± 0.29	0.008 ± 0.001
		Sensitive feature not in the functional form of the model							
New Tasks	School	13.46 ± 0.31	0.026 ± 0.001	12.97 ± 0.47	0.023 ± 0.001	13.41 ± 0.43	0.032 ± 0.002	9.99 ± 0.37	0.018 ± 0.001
	UNIV	12.14 ± 0.73	0.142 ± 0.007	12.98 ± 0.79	0.160 ± 0.011	11.02 ± 0.59	0.101 ± 0.003	11.74 ± 0.64	0.136 ± 0.007
	Movielens	10.81 ± 0.42	0.012 ± 0.001	11.43 ± 0.48	0.018 ± 0.002	11.51 ± 0.51	0.022 ± 0.003	10.25 ± 0.39	0.012 ± 0.001
		Sensitive feature in the functional form of the model							
	School	12.12 ± 0.37	0.032 ± 0.001	11.97 ± 0.48	0.034 ± 0.002	11.46 ± 0.47	0.025 ± 0.001	9.27 ± 0.38	0.024 ± 0.001
	UNIV	11.98 ± 0.65	0.241 ± 0.014	10.34 ± 0.37	0.191 ± 0.009	11.08 ± 0.61	0.221 ± 0.012	10.80 ± 0.42	0.200 ± 0.011
	Movielens	9.97 ± 0.52	0.022 ± 0.001	10.23 ± 0.67	0.028 ± 0.003	10.07 ± 0.63	0.024 ± 0.003	9.18 ± 0.44	0.020 ± 0.001

[17] 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.

[18] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in IEEE international conference on Data mining, 2009.

[19] S. Verma and J. Rubin, "Fairness definitions explained," in IEEE/ACM International Workshop on Software Fairness, 2018.

[20] 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.

[21] H. Edwards and A. Storkey, "Censoring representations with an adversary," in International Conference on Learning Representations, 2016.

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

[23] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in International Conference on Machine Learning, 2018.

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

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

[26] Y. Wang, T. Koike-Akino, and D. Erdogmus, "Invariant represen-

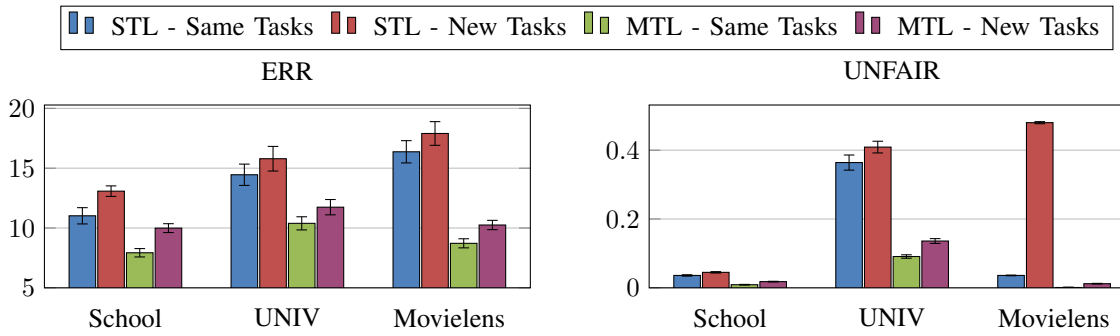

Fig. 3: Graphical representation of the results in Table III, when the sensitive feature is not included in the functional form of the model and the fairness constraint is active.

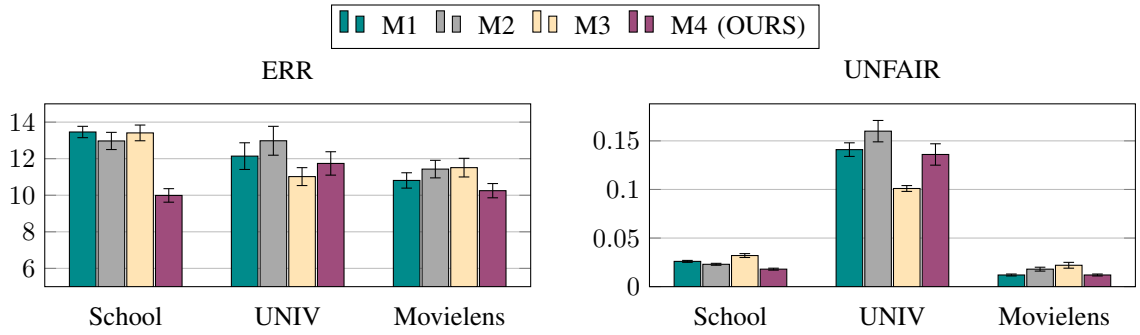

Fig. 4: Graphical representation of the results in Table IV for new tasks when the sensitive feature is not included in the functional form of the model.

tations from adversarially censored autoencoders," arXiv preprint arXiv:1805.08097, 2018.
[27] A. Maurer, "Transfer bounds for linear feature learning," Machine learning, vol. 75, no. 3, pp. 327–350, 2009.
[28] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in International conference on machine learning, 2016.
[29] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Dataset shift in machine learning. The MIT Press, 2009.
[30] N. Srebro, "Learning with matrix factorizations," PhD thesis, Massachusetts Institute of Technology, 2004.
[31] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," Machine Learning, vol. 73, no. 3, pp. 243–272, 2008.
[32] C. Ciliberto, D. Stamos, and M. Pontil, "Reexamining low rank matrix factorization for trace norm regularization," arXiv preprint arXiv:1706.08934, 2017.
[33] P. Gajane and M. Pechenizkiy, "On formalizing fairness in prediction with machine learning," arXiv preprint arXiv:1710.03184, 2017.
[34] L. Oneto, M. Donini, and M. Pontil, "General fair empirical risk minimization," arXiv preprint arXiv:1901.10080, 2019.
[35] J. Baxter, "A model of inductive bias learning," Journal of Artificial Intelligence Research, vol. 12, no. 149-198, p. 3, 2000.
[36] H. Goldstein, "Multilevel modelling of survey data," Journal of the Royal Statistical Society. Series D (The Statistician), vol. 40, no. 2, pp. 235–244, 1991.
[37] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," Acm transactions on interactive intelligent systems (tiis), vol. 5, no. 4, p. 19, 2016.
[38] M. Pontil and A. Maurer, "Excess risk bounds for multitask learning with trace norm regularization," in Conference on Learning Theory, 2013.
[39] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," Foundations of computational mathematics, vol. 12, no. 4, pp. 389–434, 2012.
[40] S. Boucheron, G. Lugosi, and P. Massart, Concentration inequalities: A nonasymptotic theory of independence. Oxford university press, 2013.
[41] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," Journal of Machine Learning Research, vol. 3, no. Nov, pp. 463–482, 2002.