Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, Ed H. Chi

Google Research, USA

{yuyanw,xuezhiw,alexbeutel,fprost,jilinc,edchi}@google.com

ABSTRACT

As multi-task models gain popularity in a wider range of machine learning applications, it is becoming increasingly important for practitioners to understand the fairness implications associated with those models. Most existing fairness literature focuses on learning a single task more fairly, while how ML fairness interacts with multiple tasks in the joint learning setting is largely underexplored. In this paper, we are concerned with how group fairness (e.g., equal opportunity, equalized odds) as an ML fairness concept plays out in the multi-task scenario. In multi-task learning, several tasks are learned jointly to exploit task correlations for a more efficient inductive transfer. This presents a multi-dimensional Pareto frontier on (1) the trade-off between group fairness and accuracy with respect to each task, as well as (2) the trade-offs across multiple tasks. We aim to provide a deeper understanding on how group fairness interacts with accuracy in multi-task learning, and we show that traditional approaches that mainly focus on optimizing the Pareto frontier of multi-task accuracy might not perform well on fairness goals. We propose a new set of metrics to better capture the multi-dimensional Pareto frontier of fairness-accuracy tradeoffs uniquely presented in a multi-task learning setting. We further propose a Multi-Task-Aware Fairness (MTA-F) approach to improve fairness in multi-task learning. Experiments on several real-world datasets demonstrate the effectiveness of our proposed approach.

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**; **Machine learning algorithms**; **Multi-task learning**.

KEYWORDS

fairness, multi-task learning, Pareto frontier, multi-task-aware fairness treatment

1 INTRODUCTION

Over the past years, multi-task deep learning has gained popularity through its success in a wide range of applications, including natural language processing [12], computer vision [20, 40], and online recommendation systems [3, 30]. Despite the popularity of using multi-task learning in various real-world applications, the understanding of fairness under such a framework is largely underexplored. Most existing work on fairness focuses on a single-task learning setting, including work on fair representation learning [7, 31, 45, 54, 58], and fairness mitigation [1, 6, 13, 14, 53].

Here we focus on the widely used notion of group fairness, *equal opportunity* and *equalized odds* proposed by [22], which aims at closing the gap of true positive rates and false positive rates across different groups. In a single-task learning setting, the Pareto frontier can be characterized by trading-off some notions of accuracy and group fairness over a single task [35, 46].

Fairness in multi-task learning poses new challenges and the need of characterizing a multi-dimensional Pareto frontier. In a traditional multi-task learning setting where fairness is not taken into consideration, people focus on optimizing the Pareto frontier of multiple accuracies across tasks. Instead, our work aims at demystifying the multi-dimensional trade-off and improving fairness on top of accuracy objectives for multi-task learning problems.

First, we show analysis that traditional multi-task learning objectives might not correlate well with fairness goals, thus demonstrating that fairness needs to be better handled in multi-task learning settings. To better evaluate fairness in multi-task learning, we then propose a new set of metrics to capture the multi-dimensional nature of the Pareto frontier. This includes two sets of trade-offs, one that captures the fairness vs. accuracy trade-off with respect to each task, and the other captures the trade-offs across multiple tasks. We show that due to the intrinsic differences in various tasks, a simple aggregation of the accuracy or fairness metrics of each task might not be a fair assessment of the entire system. Specifically, when different tasks suffer from fairness issues to different extents (and thus have fairness metrics in different scales), we propose a set of normalized metrics (Average Relative Fairness Gap and Average Relative Error, Section 4) that better captures the overall trade-off between fairness and accuracy.

Finally, we propose a data-dependent mitigation method, *Multi-Task-Aware Fairness treatment (MTA-F)*, that improves the Pareto frontiers for fairer multi-task learning. The main idea is to decompose the remediation treatments for different components of the multi-task model architecture according to label distributions across all tasks. By doing so, MAT-F allows flexible mitigation over (1) the representation learning shared by all tasks, and (2) the non-shared sub-networks specific to each task. This effectively enables more efficient use of limited model capacity across multiple tasks

over both accuracy and fairness goals. Compared with a baseline method which is not data-dependent and the same fairness loss is applied to all parts of the multi-task models, MTA-F improves the Pareto efficiency when trading-off fairness vs. accuracy across multiple tasks. Experiment results on multiple real-world datasets demonstrate the effectiveness of MTA-F in multi-task applications.

To summarize, our contributions are:

- **Problem Framing:** We provide insights on how fairness interacts with accuracy in multi-task learning. Notably, traditional approaches that focus on optimizing the Pareto frontier of multi-task accuracy may not correlate well with equal opportunity goals.
- **New Metrics:** We propose a new set of metrics that captures the multi-dimensional fairness-accuracy trade-offs uniquely presented in a multi-task learning setting.
- **New Mitigation:** We provide a data-dependent multi-task fairness mitigation approach, MTA-F, which decomposes fairness losses for different model components by exploiting task relatedness and the shared architecture for multi-task models.

2 RELATED WORK

Fairness metrics. A number of fairness metrics have been proposed in existing works, with the focus mostly on fairness in classification. *Demographic parity* [22, 46] requires that all subgroups receive the same proportion of positive outcomes. Although being adopted as the definition of fairness in a series of works [9, 24, 25], demographic parity may be unrealistic to achieve in practice, especially in the common scenarios where the base rates differ across subgroups [60]. Another set of commonly used fairness metrics, *equal opportunity* and *equalized odds* [22], has also been widely adopted in measuring discrimination against protected attributes. Instead of requiring equal outcomes, equal opportunity and equalized odds require equal true positive rates and false positive rates across different subgroups, a somewhat more realistic fairness notion in a wide range of applications in practice. For our work, we focus on equal opportunity and equalized odds as our notion of fairness.

Fairer representation learning. One way to address fairness in machine learning is through fairer representation learning [45, 54, 58], with the goal being obfuscating information about the protected group membership in the learned representations. Fairer representation learning can be achieved adversarially [7, 31]. Moreover, Schumann et al. [43] propose transfer learning in the representation space to adapt fairness to a new domain.

Fairness mitigation. There is a large body of work on mitigating fairness issues in a single-task learning setting. Examples include pre-processing the data embeddings to help downstream models be trained more fairly [8], and post-processing model's predictions to improve fairness [28, 38]. In addition, intervening the model training process has also been popular, including adding fairness constraints [1, 13, 14] or regularization [6, 52, 53]. Different from existing works that add fairness constraints directly into the model's objective for single-task learning, our work exploits the multi-task model architecture with a decomposition and redistribution of the fairness constraints for fairer multi-task learning.

Multi-task learning. A common approach for multi-task learning is to design a parameterized model that shares a subset of parameters across different tasks [41]. The benefits of such shared architectures are numerous. First, it exploits task relatedness with inductive bias learning [4, 10]. Learning a shared representation across related tasks is beneficial especially for harder tasks or tasks with limited training examples. Second, by forcing tasks to share model capacity, it introduces a regularization effect and improves generalization. Third, it offers a compact and efficient form of modeling which enables training and serving multiple prediction quantities for large-scale systems. Learning multiple tasks together can improve the performance of some tasks but hurt others [44]. Existing literature in multi-task learning focus on reducing the task training conflicts [11, 26, 27, 30, 51] and improving the Pareto frontier.

Fairness in multi-task learning. Fairness has been mostly studied in single-task settings, and little work has been done in the context of multi-task learning. However, as multi-task learning becomes prevalent in state-of-the-art model designs [42, 55], it is important for practitioners to understand how fairness interacts with multi-task learning. D'Amour et al. [15] provide insights on how pure multi-task learning can have unintended effects on fairness. Zhao and Chen [56] study fairness in multi-task regression models and uses a rank-based non-parametric independence test to improve fairness in ranking. Oneto et al. [37] propose to use multi-task learning enhanced with fairness constraints to jointly learn classifiers that leverage information between sensitive groups. Zhao et al. [57] study fair meta-learning. Wang et al. [47] show that fairness might not compose in multi-component recommenders.

Fairness-accuracy trade-off and Pareto fairness. Finally, a great amount of work has shown that fairness usually comes with a trade-off over accuracy, e.g., [33, 35, 46, 59], again mostly under a single-task setting. In single-task learning, the Pareto frontier can be characterized by trading-off some notions of the accuracy and group fairness over a single task. For example, the objective is generally in the form of $\mathcal{L}(t) + \lambda\mathcal{F}(t)$, where $\mathcal{L}(t)$ represents the loss over accuracy for task t, $\mathcal{F}(t)$ is the fairness loss, and λ is a parameter trading-off accuracy and fairness. Balashankar et al. [2] explore the pareto-efficient fairness and show that it achieves Pareto levels in accuracy for all subgroups for a single-task. Martinez et al. [34] and Diana et al. [17] propose to find Pareto optimal solutions for minimax group fairness in which fairness is measured by worst-case single-task accuracies across groups.

In a traditional multi-task learning setting where fairness is not taken into consideration, people focus on optimizing the Pareto frontier of multiple accuracies across tasks, where the objective can be usually written as $\alpha\mathcal{L}(t_1) + (1 - \alpha)\mathcal{L}(t_2)$ assuming two tasks t_1, t_2 and $\alpha \in [0, 1]$ is the weight on task t_1. Fairness in multi-task learning poses the need of characterizing a multi-dimensional Pareto frontier: For each single-task, there is a trade-off between fairness and accuracy, while at the same time this trade-off needs to be further balanced across all the tasks. For example, in the case with two tasks, one can extend the fairness losses from single-task learning to multi-task learning, by formulating the objective as:

$$\alpha[\mathcal{L}(t_1) + \lambda_1\mathcal{F}(t_1)] + (1 - \alpha)[\mathcal{L}(t_2) + \lambda_2\mathcal{F}(t_2)].$$

The number of objectives and metrics of interest is doubled compared with the case with single-task learning or multi-task learning

without fairness treatments, hence introducing challenges in understanding and optimizing the multi-dimensional Pareto frontier.

In this work, we focus on studying this fairness-accuracy trade-off under the *multi-task* learning setting which has been largely under-explored before. In particular, we aim to provide a better understanding on the following key questions: 1) How does inductive transfer [4] in multi-task learning implicitly impacts fairness? 2) How do we efficiently measure the fairness-accuracy trade-off for multi-task learning, given the multi-dimensional nature of the trade-off surface? 3) Are we able to achieve a better Pareto efficiency in fairness across multiple tasks, by exploiting task relatedness and shared architecture that's specific to multi-task learning?

3 PROBLEM DEFINITION AND EXAMPLES

3.1 Multi-Task Learning

Suppose there are T tasks sharing an input space $\mathcal{X}$. Each task has its own task space $\{\mathcal{Y}^t\}_{i=1}^T$. A dataset of n i.i.d. examples from the input and task spaces is given by $\{(x_i, y_i^1, ..., y_i^T)\}_{i=1}^n$, where y_i^t is the label of the t-th task for example i. We assume a multi-task model parameterized by $\theta \in \Theta$. Here $\theta = (\theta_{sh}, \theta_1, ..., \theta_T)$ includes shared-parameters θ_{sh} and task-specific parameters $\theta_1, ..., \theta_T$. Let $f_t(\cdot, \cdot) : \mathcal{X} \times \Theta \to \mathcal{Y}^t$ be the model function and $\mathcal{L}_t(\cdot, \cdot) : \mathcal{Y}^t \times \mathcal{Y}^t \to \mathbb{R}^+$ be the loss function for the t-th task. This generalizes easily to the more general multi-task learning setting where different tasks have different inputs. Figure 1 is an illustration of the typical shared-bottom architecture for a multi-task model, as used in a wide range of applications [10, 12, 30, 42].
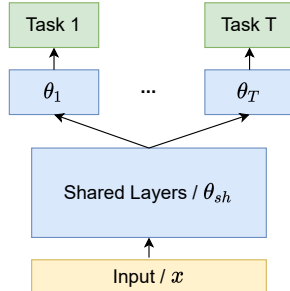

Figure 1: Shared-bottom architecture for a multi-task model.

Let $\hat{\mathcal{L}}_t(\theta) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}_t(f_t(x_i; \theta_{sh}, \theta_t), y_i^t)$ be the empirical loss for the t-th task, where we drop the dependency on x and y for ease of notation. The optimization for multi-task learning can then be formulated as a joint optimization of a vector-valued loss function:

$$\min_{\theta}(\hat{\mathcal{L}}_1(\theta), ..., \hat{\mathcal{L}}_T(\theta))^\top. \tag{1}$$

It is unlikely that a single θ optimizes all objectives simultaneously. The solution to (1) is therefore a set of points which represent different trade-off preferences. Formally, solution θ_a is said to *Pareto dominate* solution θ_b if $\hat{\mathcal{L}}_t(\theta_a) \le \hat{\mathcal{L}}_t(\theta_b), \forall t$ and there exist at least one task j such that the inequality is strict. A solution θ is called *Pareto optimal* or *Pareto efficient* if there is no solution $\theta' \ne \theta$ such that θ' dominates θ. The *Pareto frontier* is the set of all Pareto optimal solutions.

The most popular choice for the minimization objective is a scalarization of the empirical loss vector [32]:

$$\hat{\mathcal{L}}(\theta) := \sum_{t=1}^T w_t \hat{\mathcal{L}}_t(\theta), \tag{2}$$

where $\{w_t\}_{t \in \{1, ..., T\}} \ge 0$ are weights for individual tasks, which controls the trade-off among different tasks.

3.2 Fairness Definition and Metrics

To measure fairness, we assume each example belongs to a subgroup defined by the sensitive attribute A (e.g. gender). The sensitive attribute information could be available only for a fraction of the training data. For ease of notation we assume the sensitive attribute is a binary variable (e.g. $A = 1$ for female, $A = 0$ for male).

We focus on two widely used notions of group fairness, equal opportunity and equalized odds [22]. Equal opportunity is defined as equality of true positive rates (TPR) or false positive rates (FPR) across groups[1]. Equalized odds is defined as the predicted outcome $\hat{Y}$ being independent of the group membership conditional on the true outcome. In the case of classification, it entails equal true positive and false positive rates across different groups. The discrepancies in FPR and TPR are defined as FPR gap and TPR gap:

$$FPRGap = |P(\hat{Y} = 1|Y = 0, A = 0) - P(\hat{Y} = 1|Y = 0, A = 1)|,$$
$$TPRGap = |P(\hat{Y} = 1|Y = 1, A = 0) - P(\hat{Y} = 1|Y = 1, A = 1)|. \tag{3}$$

We use FPR Gap to measure the deviation from equal opportunity, and FPR gap and TPR gap for equalized odds.

3.3 Measuring Fairness in Multi-Task Learning

In this section, we look into how training multiple tasks together can have implications on the fairness of individual tasks.

We look at two datasets with known fairness concerns. The CelebA dataset [29] contains public domain images of public figures. It has more than 200K images and each has 40 binary attribute annotations. We pick 2 attributes, *Attractive* and *Smiling* which are known to be biased in gender [16], as the tasks for multi-task learning. The UCI-Adult dataset[2] [18] contains census information of over 48,000 adults from the 1994 Census, where we define *Income* > \$50,000 or ≤ \$50,000 as Task 1, and *Capital Gain* > 0 or not as Task 2. Details of the model can be found in Appendix A.2. For both datasets, we focus on gender as the sensitive attribute.

To understand how training multiple tasks together affects the fairness and accuracy of individual tasks, we compare the FPR gap and error rate on the test data under single-task learning (STL) and multi-task learning (MTL). For STL of each task, we use the same model architecture as in MTL but without the head layers for other tasks in Fig. 1. Therefore, each task has the same effective model capacity under STL and MTL. For MTL, we set $w_1 = w_2 = 0.5$ as the task loss weights in Eq. (2).

Table 1 summarizes the numerical results. On the CelebA dataset (Table 1a), multi-task learning hurts the fairness of Task 1 (*Attractive*) but improves the fairness of Task 2 (*Smiling*). On the UCI-Adult dataset (Table 1b), while multi-task learning helps the accuracy of

---

[1]The original paper [22] frames equal opportunity as equality of TPR, but equality of FPR is also widely used in many literature and applications.
[2]https://archive.ics.uci.edu/ml/datasets/adult

Task 2 (*capital gain* > 0), it significantly degrades its fairness by increasing its FPR gap by 11.6%.

To further understand this, we perform another 10k runs of the multi-task model on the UCI-Adult dataset, with the task weights w_1 and $w_2 = 1 - w_1$ varying in $[0, 1]$. Figure 2 shows the accuracy Pareto frontier and fairness Pareto frontier over those runs. We pick the model runs that are Pareto optimal in accuracies (Fig. 2a) and plot their fairness metrics together with the fairness Pareto frontier in Fig. 2b. We see that the model runs that are Pareto optimal in accuracies are far from the fairness Pareto frontier, meaning that they are *not* Pareto optimal in terms of fairness.

The comparisons suggest that training multiple tasks together by simply pooling the accuracy objectives using a shared model architecture may have serious fairness implications—a fact that could be easily overlooked in multi-task applications. Even when only optimizing for accuracy objectives, multi-task learning may have larger impacts on fairness goals than on accuracy goals (Table 1a), or hurting the fairness of some tasks while benefiting from accuracy gains (Table 1b). More generally, optimizing for accuracy among multiple tasks may lead to suboptimal trade-offs in fairness (Fig. 2). Ignoring these implications in practice may lead to unwanted fairness consequences in multi-task applications.

	T1 Error	T1 FPR Gap	T2 Error	T2 FPR Gap
STL-T1	0.2030	0.2716	-	-
STL-T2	-	-	0.0784	0.0145
MTL	0.2035	0.2846	0.0783	0.0137
Difference	+0.24%	**+4.78%**	-0.08%	**-5.39%**

(a) **CelebA**: MTL hurts Task 1 fairness but improves Task 2 fairness.

	T1 Error	T1 FPR Gap	T2 Error	T2 FPR Gap
STL-T1	0.1659	0.1200	-	-
STL-T2	-	-	0.1313	0.0661
MTL	0.1656	0.1205	0.1299	0.0738
Difference	-0.20%	+0.34%	**-1.10%**	**+11.60%**

(b) **UCI-Adult**: MTL improves Task 2 accuracy but hurts its fairness.

Table 1: Fairness and accuracy metrics for single-task learning and multi-task learning. **STL-T1**: single-task learning for Task 1; **STL-T2**: single-task learning for Task 2; **MTL**: multi-task learning; **Difference**: percentage difference of MTL and STL metrics. Results are averaged over 100 runs. **Bold** font means the difference is statistically significant at 95% confidence level.
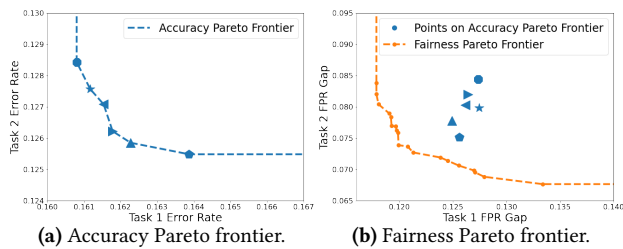

(a) Accuracy Pareto frontier.　　(b) Fairness Pareto frontier.

Figure 2: Pareto frontiers on UCI-Adult dataset, where points on lower-left means better accuracy/fairness metrics. Blue points in (b) correspond to the Pareto optimal runs in (a).

4 METRICS FOR FAIRNESS-ACCURACY PARETO FRONTIER

Here we propose metrics to measure the multi-dimensional trade-off between fairness and accuracy across multiple tasks, and then in the next section propose treatments to improve the trade-off for multi-task models. As shown above, the usual multi-task learning objectives such as accuracy often do not align with fairness goals; A good accuracy trade-off among tasks may have bad fairness implications. Therefore, fairness trade-off must be considered together with accuracy trade-off when evaluating multi-task models.

To this end, we propose metrics that characterize the Pareto frontier of fairness-accuracy trade-off for multi-task learning. Consider, for example, equal opportunity as the fairness objective and FPR gap as the fairness metric. Let $Err^{(1)}, ..., Err^{(T)}$ be the error rates for the T tasks where $Err = P(\hat{Y} \neq Y)$, and $FPRGap^{(1)}, ..., FPRGap^{(T)}$ be the corresponding FPR gaps as defined in Eq. (3). This leaves us with a $2T$-dimensional Pareto frontier, which is hard to visualize and compare among multiple models.

A straight-forward approach for measuring fairness-accuracy trade-off in multi-task learning is to project the $2T$-dimensional metric into a 2-dimensional (accuracy, fairness) metric by averaging accuracy metrics and fairness metrics across all tasks:

$$\overline{Err} = \frac{1}{T}\sum_{t=1}^{T} Err^{(t)}, \quad \overline{FPRGap} = \frac{1}{T}\sum_{t=1}^{T} FPRGap^{(t)}. \quad (4)$$

However, naive averaging ignores the relative metric scale and baseline values for each task. We define a set of normalized average metric, **average relative fairness gap** (*ARFG*) and **average relative error** (*ARE*), which consolidates metric scales by comparing each multi-task metric against its single-task counterparts:

DEFINITION 1. *(Average relative fairness gap and average relative error). Let $Err_S^{(t)}$ and $FPRGap_S^{(t)}$ be the error rate and FPR gap for task t using single-task learning with the same model architecture as in multi-task learning (t = 1, ..., T) and without any fairness remediation. We define Average Relative Fairness Gap (ARFG) and Average Relative Error (ARE) for a multi-task learning model as the average of relative FPR Gap and error rates against single-task baselines:*

$$ARFG := \frac{1}{T}\sum_{t=1}^{T} FPRGap^{(t)}/FPRGap_S^{(t)},$$
$$ARE := \frac{1}{T}\sum_{t=1}^{T} Err^{(t)}/Err_S^{(t)}. \quad (5)$$

Note that the single-task learning baselines $FPRGap_S^{(t)}$ and $Err_S^{(t)}$ are obtained using the same model architecture as in the multi-task learning case. Similar to the individual *FPRGap* and error rates, *low* values of *ARFG* and *ARE* indicate good overall fairness and accuracy. Table 2 shows an example that compares these new definitions with naive averaging as in Eq. (4). We see that between the two hypothetical models with the same average error $\overline{Err}$, Model A has much lower *ARE* than Model B. This is because *ARE* considers relative changes rather than absolute values. Compared with single-task learning, Model A reduces Task 1 error by 5% (from 0.40 to 0.38), Model B reduces Task 2 error by 50% (from

0.04 to 0.02) while keeping the other task's error unchanged. When the two tasks are equally important, intuitively Model B is better than Model A as it achieves better relative error reduction. This is reflected in *ARE* but not $\overline{Err}$.

	T1 Error	T2 Error	$\overline{Err}$	*ARE*
STL-T1	0.40	-	-	-
STL-T2	-	0.04	-	-
MTL Model A	0.40	0.02	**0.21**	**0.75**
MTL Model B	0.38	0.04	**0.21**	**0.975**

Table 2: A hypothetical example comparing *ARE* with $\overline{Err}$. **STL-T1/T2**: single-task learning for Task 1/Task 2, used to get the baseline error $Err_S^{(t)}$ for Task $t = 1, 2$; **MTL Model A/B**: Two hypothetical multi-task models for comparison.

Note that these average relative metrics *ARFG* and *ARE* are always positive, and can be either smaller or greater than 1 as multi-task learning could either improve or hurt accuracy or fairness of individual tasks as shown in Section 3.3. *ARFG* < 1 (*ARE* < 1) suggests that multi-task learning reduces relative FPR gap (error) on average, and vice versa. The trade-off between *ARFG* and *ARE* depicts the Pareto frontier of accuracy and fairness while accounting for different metric scales for different tasks. We call it *ARFG-ARE Pareto frontier*.

5 METHOD

We first discuss the baseline method for adding fairness treatment to a multi-task learning model, which is a naive generalization of the single-task learning case. We focus on equal opportunity in this section, but the formulation and discussions can be generalized to other group fairness objectives such as equalized odds.

5.1 Baseline: Per-Task Fairness Treatment

Group Loss definitions. For single-task learning, equal opportunity (i.e. equalized FPR) entail matching the predictions over negative examples across groups, and can be accomplished through:

- minimizing the correlation between group membership and the predictions over negative examples [5, 6] (Eq. 6a);
- kernel-based distribution matching through Maximum Mean Discrepancy (MMD) over negative examples [39] (Eq. 6b);
- minimizing FPR gap directly in the loss [19, 35, 52] (Eq. 6c).

The way these methods work is intuitive: By closing the gap on the prediction differences on negative examples, the model is able to achieve similar FPR across all groups, thus optimizing towards equal opportunity. Therefore, we can define losses as essentially different measures of distance between the predictions on sensitive subgroups defined by A over the *subpopulation with negative examples* (i.e. $Y = 0$), represented by N. To be precise, the *population* version of these fairness losses are defined as:

$$\mathcal{F}_{Corr}(f|N) := Corr(f(x), A \mid Y = 0), \tag{6a}$$

$$\mathcal{F}_{MMD}(f|N) := MMD(\{f(x)|A = 0\}, \{f(x)|A = 1\} \mid Y = 0), \tag{6b}$$

$$\mathcal{F}_{FPR}(f|N) := |P(\hat{Y} = 1|Y = 0, A = 0) - P(\hat{Y} = 1|Y = 0, A = 1)|. \tag{6c}$$

Empirical Loss definitions. Given the examples $\{(x_i, y_i, a_i)\}_{i=1}^n$ where a_i is the value of the sensitive attribute A on the i-th example, the *empirical* equal opportunity fairness losses are computed as:

$$\hat{\mathcal{F}}_{Corr}(f|N) = \hat{Corr}(\{f(x_i)\}, \{a_i\} \mid y_i = 0), \tag{7a}$$

$$\hat{\mathcal{F}}_{MMD}(f|N) = MMD(\{f(x_i) : a_i = 0\}, \{f(x_i) : a_i = 1\} \mid y_i = 0), \tag{7b}$$

$$\hat{\mathcal{F}}_{FPR}(f|N) = |\frac{\sum_i \mathbb{1}[\hat{y}_i = 1, y_i = 0, a_i = 0]}{\sum_i \mathbb{1}[y_i = 0, a_i = 0]} - \frac{\sum_i \mathbb{1}[\hat{y}_i = 1, y_i = 0, a_i = 1]}{\sum_i \mathbb{1}[y_i = 0, a_i = 1]}|, \tag{7c}$$

where MMD distance [21] measures the distance between two groups of values based on kernel methods.

Baseline Remediation. For remediation in single-task learning, the empirical loss function is a linear combination of the primary accuracy loss and the fairness loss:

$$\hat{\mathcal{L}}_{STL}(f) = \hat{\mathcal{L}}(f) + \lambda \hat{\mathcal{F}}(f|N), \tag{8}$$

where $\hat{\mathcal{L}}(f) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x_i), y_i)$ is the empirical accuracy loss and $\hat{\mathcal{F}}(f|N)$ is the fairness loss of choice as defined in Eq. (7); The subscript STL stands for single-task learning, and λ is a hyperparameter controlling the trade-off between accuracy and fairness.

Note that while the primary loss $\hat{\mathcal{L}}(f)$ is computed for every example, the computation for fairness loss depends on the fairness notion of choice. For equal opportunity which minimizes FPR gap, only negative examples are included in the fairness loss computation (Eq. (6) and (7)); For equalized odds which minimized both FPR gap and TPR gap, there will be two terms in the fairness loss $\hat{\mathcal{F}}(f|\cdot)$, with one computed on negative examples corresponding to minimizing FPR gap, and the other computed on positive examples corresponding to minimizing TPR gap.

A naive generalization of fairness treatments to multi-task learning is to apply the single-task fairness treatment to each task separately. For equal opportunity as an example, for every task we take its negative examples, compute the fairness loss, and add to the primary losses for the accuracy objectives in Eq. (2). This yields a *per-task* fairness treatment for multi-task learning: Using the notations in Section 3.1 where the task prediction functions are parameterized by $\theta = (\theta_{sh}, \theta_1, ..., \theta_T)$ and dropping the dependency on (x_i, y_i^t, a_i) for ease of notation, the total empirical loss equals:

$$\hat{\mathcal{L}}_{MTL}(\theta) = \sum_{t=1}^T w_t [\hat{\mathcal{L}}_t(\theta) + \lambda_t \hat{\mathcal{F}}_t(\theta)], \tag{9}$$

where $\hat{\mathcal{L}}_t(\theta)$ as in Eq. (2) is the empirical loss for the accuracy objective (we call it "accuracy los" in the remainder of this paper) for task t, and $\hat{\mathcal{F}}_t(\theta)$ is the empirical loss for the fairness objective (we call it "fairness loss" in the remainder of this paper), and λ_t is the fairness weight which controls the relative trade-off between accuracy and fairness for task t.

For a multi-task model as in Fig. 1, the shared layers (parameterized by θ_{sh}) receive gradient updates from all task losses during training, while the task-specific head layers (parameterized by $\theta_1, ..., \theta_T$) effectively receive gradient updates only from its own task losses, as $\nabla_{\theta_t} \mathcal{L}(f(x; \theta_{sh}, \theta_{t'}), y) \equiv 0$ for any $t \neq t'$. Figure 3a illustrates this fact for the case of $T = 2$ tasks.
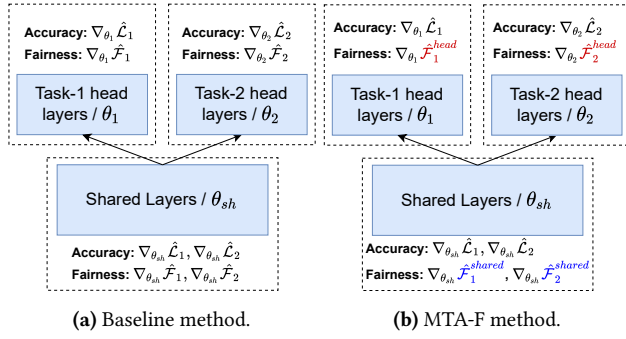
(a) Baseline method.   **(b)** MTA-F method.

Figure 3: Effective gradients on different model components.

5.2 Our Method: Multi-Task-Aware Fairness Treatment (MTA-F)

We now present our multi-task-aware fairness treatment framework for improving fairness-accuracy trade-off for multi-task learning. We start with equal opportunity as the fairness objective and generalize to equalized odds later.

The fairness loss is computed on the negative examples of each task in order to minimize FPR gap towards equal opportunity. Let $N_1, ..., N_T$ be the sets of negative examples for each task, which may or may not overlap. In the example with 2 tasks, the fairness loss acting on the shared bottom layer is computed on the union of negative examples for either Task 1 or Task 2 ($N_1 \cup N_2$). However, the examples that are negative on Task 1 but positive on Task 2 ($N_1 \cap N_2^c$, the orange region in Fig. 4a) should only be relevant to the fairness goal of Task 1; likewise the examples that are negative on Task 2 but positive on Task 1 ($N_2 \cap N_1^c$, the blue region in Fig. 4a) should only be relevant to the fairness goal of Task 2. In the baseline method in Section 5.1, the fairness loss does not distinguish those examples and include all of them in the fairness loss computation for both shared layers and head layers, which intuitively is a suboptimal use of the model capacity. Inspired by this observation, we propose a decomposition and redistribution of fairness treatments for multi-task learning: For every task t, obtain its fairness loss $\hat{\mathcal{F}}_t$ computed on the negatives N_t as in the baseline method. Then compute the fairness loss on the exclusive negatives $N_t \cap (\cap_{k \neq t} N_k^c)$, i.e. the examples that are negative only on task t but not the rest, and apply them to the head layers; And leave the remaining fairness loss to the shared layers.

Formally, define $N_t := \{(x_i, y_i^1, ..., y_i^T, a_i) : y_i^t = 0\}$ as subsets of examples that have negative labels on task t for $t = 1, ..., T$. Decompose the total fairness loss $\hat{\mathcal{F}}(\theta|N_t)$ computed on N_t as task-specific fairness loss $\hat{\mathcal{F}}_t^{head}$ and shared fairness loss $\hat{\mathcal{F}}_t^{shared}$:

$$\hat{\mathcal{F}}_t^{head}(\theta) := \hat{\mathcal{F}}(\theta \mid N_t \cap (\cap_{k \neq t} N_k^c)),$$
$$\hat{\mathcal{F}}_t^{shared}(\theta) := \hat{\mathcal{F}}(\theta|N_t) - \hat{\mathcal{F}}_t^{head}(\theta),$$

(10)

for every task $t = 1, ..., T$, where $\hat{\mathcal{F}}_t$ is any fairness loss function as defined in Eq. (7). During training, we backpropagate task-specific fairness losses $\hat{\mathcal{F}}_t^{head}$ to the head layers (θ_t) and the remaining fairness losses $\hat{\mathcal{F}}_t^{shared}$ to the shared layers (θ_{sh}).

Figure 4b illustrates the proposed idea on a 2-task model and how the gradient flows during backpropagation. Note that $\hat{\mathcal{F}}_t^{shared}$ is
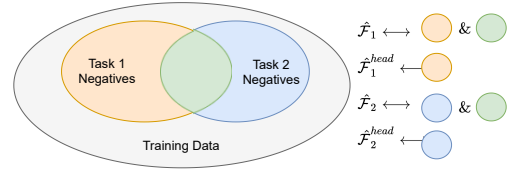
(a) Fairness loss decomposition with MTA-F.
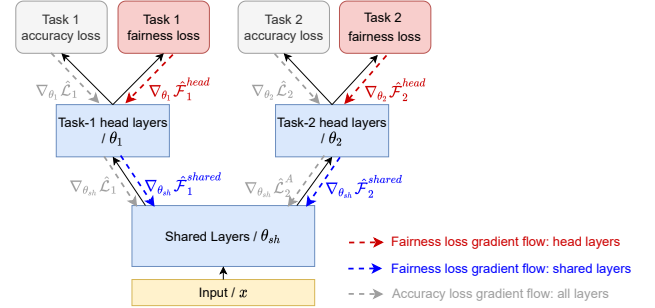
(b) Backpropagation with MTA-F: We backpropagate task-specific fairness losses $\hat{\mathcal{F}}_t^{head}$ to head layers, and the remaining fairness loss $\hat{\mathcal{F}}_t^{shared}$ to shared layers ($t = 1, 2$).

Figure 4: Multi-task-aware fairness treatment (MTA-F).

not equivalent to the fairness loss computed on the shared negative examples (green region in Fig 4a) as most fairness losses are not additive. The number of fairness losses with this decomposition scales linearly with T. Figure 3 illustrates the difference in fairness gradients between the proposed method and the baseline method. Because it is considering the labels of other tasks when designing the fairness loss for each task, we call this method **multi-task-aware fairness treatment (MTA-F)**.

The multi-task learning update step described in Algorithm 1 is applicable to any problem that uses gradient-based optimization. For task t, in addition to the accuracy loss weight w_t and fairness loss weight λ_t as in the baseline algorithm in Section 5.1, we introduce an additional hyperparameter $r_t > 0$ which controls the relative ratio of the fairness loss weight between the head layers and the shared layers.

Section 6 shows the effectiveness of MTA-F in achieving a better fairness-accuracy trade-off on real-world datasets. The fact that a simple decomposition and redistribution of the fairness losses improves the Pareto efficiency in multi-task learning is not surprising to us. MTA-F enables different parts of the model to address the fairness concerns across multiple tasks in a targeted way. Head layers address fairness issues that are *specific* to the task itself, while shared layers address fairness issues that are *common* to multiple tasks. The mediation leads to a more efficient allocation of model capacity for fairness objectives, which effectively "saves" the model more capacity for accuracy objectives, thus enabling a better fairness-accuracy trade-off. On a related note, existing studies show that multi-task models of the same capacity can have different generalization performance for accuracy objectives, by implicitly manipulating the information flow in different layers [36, 48].

We point out that although we describe the formulation of MTA-F for optimizing equal opportunity, it can be easily generalized

---

Algorithm 1: MTA-F Update Rule

Input: Mini-batch $\{(x_i, y_i^1, ..., y_i^T)\}_{i=1}^n$, model parameters $\theta = (\theta_{sh}, \theta_1, ..., \theta_T)$, task weights $\{w_t\}_{t=1}^T$, fairness weights $\{\lambda_t\}_{t=1}^T$, head-to-shared ratio $\{r_t\}_{t=1}^T$, and learning rate η

1 **for** $t = 1, ..., T$ **do**

2     $\hat{\mathcal{L}}_t(\theta_{sh}, \theta_t) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_t(f_t(x_i; \theta_{sh}, \theta_t), y_i^t)$
        ▷ *Compute accuracy losses*

3     $\hat{\mathcal{F}}_t^{head}(\theta_{sh}, \theta_t)$ and $\hat{\mathcal{F}}_t^{shared}(\theta_{sh}, \theta_t)$ as in Eq. (10) / (11)
        ▷ *Compute fairness losses*

4     $\theta_t = \theta_t - \eta w_t [\nabla_{\theta_t} \hat{\mathcal{L}}_t(\theta_{sh}, \theta_t) + \lambda_t r_t \nabla_{\theta_t} \hat{\mathcal{F}}_t^{head}(\theta_{sh}, \theta_t)]$
        ▷ *Gradient descent on head parameters*

5 **end**

6 $\theta_{sh} = \theta_{sh} - \eta \{\sum_{t=1}^T w_t [\nabla_{\theta_{sh}} \hat{\mathcal{L}}_t(\theta_{sh}, \theta_t)$
        $+ \lambda_t \nabla_{\theta_{sh}} \hat{\mathcal{F}}_t^{shared}(\theta_{sh}, \theta_t)]\}$
        ▷ *Gradient descent on shared parameters*

Output: Updated model parameters $\theta = (\theta_{sh}, \theta_1, ..., \theta_T)$

---

to other fairness metrics such as TPR gap for equal opportunity (instead of FPR gap), or both FPR gap and TPR gap for equalized odds. When TPR gap is the fairness goal, the fairness losses are computed on the positive examples (instead of negative examples) of each task. In the case for equalized odds where both FPR gap and TPR gap are of interest, we just need to add two addition fairness loss terms in $\hat{\mathcal{F}}_t^{head}$ and $\hat{\mathcal{F}}_t^{share}$ in Eq. (10) as:

$$\hat{\mathcal{F}}_t^{head}(\theta) = \hat{\mathcal{F}}(\theta \mid N_t \cap (\underset{k \neq t}{\cap} N_k^c)) + \hat{\mathcal{F}}(\theta \mid P_t \cap (\underset{k \neq t}{\cap} P_k^c)),$$
$$\hat{\mathcal{F}}_t^{shared}(\theta) = \hat{\mathcal{F}}(\theta | N_t) + \hat{\mathcal{F}}(\theta | P_t) - \hat{\mathcal{F}}_t^{head}(\theta), \quad (11)$$

where $P_t \coloneqq \{(x_i, y_i^1, ..., y_i^T, a_i) : y_i^t = 1\}$ is the set for *positive* examples of task t ($1 \leq t \leq T$), and $\hat{\mathcal{F}}$ is the fairness loss of choice. Everything else works exactly the same as in Algorithm 1. Therefore MTA-F is a general framework for improving Pareto efficiency for multi-task learning toward fairness goals including equal opportunity and equalized odds, and works with different fairness loss functions. MTA-F also works when the sensitive attribute is available only for a fraction of the training data, for which the fairness loss computation and decomposition in Eq. (10) is done on the subset of training data where the sensitive attribution information is available.

6 EXPERIMENTS

In this section, we present experiment results of MTA-F on three real-world datasets to demonstrate its effect on improving fairness-accuracy trade-off for multi-task learning problems. We first introduce the datasets and the experiment setup, then we present the results on the Pareto frontier of ARFG and ARE metrics defined in Section 4, along with the fairness and accuracy metrics per task.

6.1 Experiment Setup

Datasets: We test on three widely used datasets with known fairness concerns: the UCI-Adult dataset, the German Credit Data[3]

[3]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

and the LSAC Law School dataset[49]. We formulate the multi-task learning problems by picking two binary tasks for each of them: *Income* > \$50, 000 and *Capital Gain* > 0 for UCI-Adult dataset, *Good loans* and *High credit* for the German Credit Data, and *Pass bar exam* and *High first year GPA* for the LSAC Law School dataset. Details on the datasets and task constructions can be found in Appendix A.1. Gender is treated as the sensitive attribute for all three datasets.

Treatments: The methods we compare in the experiments are:

- **Vanilla MTL**: The naive multi-task learning without any fairness remediation, i.e. only minimizing combination of accuracy losses as in Eq. (2);
- **Baseline**: The baseline treatment described in Section 5.1, with the minimization objective as in Eq. (9);
- **MTA-F**: Our proposed multi-task-aware fairness treatment described in Section 5.2 and Algorithm 1.

Fairness Loss: In order to test our method on different fairness loss functions, we adopt MMD loss $\mathcal{F}_{MMD}$, FPR Gap loss $\mathcal{F}_{FPR}$ and correlation loss $\mathcal{F}_{Corr}$ as defined in Eq. (6) as the fairness loss on the three datasets respectively. The losses are computed on negative examples, optimizing toward equal opportunity as the fairness goal which is measured by FPR gap. Cross-entropy loss is used for all accuracy losses.

Architecture and tuning: For all three datasets, we adopt the standard multi-task architecture with shared representation learning and task sub-networks as in Fig. 1. For each method, the same set of hyperparameter values is used across different fairness and accuracy weights. Details on the model architecture, tuning and the code can be found in Appendix A.2. Once the hyperparameters are selected, we perform 20k runs for each method with varying task fairness weights and loss weights, evaluate each of them on test dataset, and report the Pareto frontier of the test metrics.

6.2 Results

In Table 3, we first present the numerical results by reporting one point on the *ARFG-ARE* Pareto frontier and read the corresponding *ARFG* and *ARE* metrics. For each method, the point is chosen to be such that its *ARE* value is closest to the midpoint of the interval of *ARE* values from all runs of that method. The corresponding per-task fairness and accuracy metrics are reported in Table 4.

Dataset	UCI-Adult		German Credit		LSAC Law School	
Metric	*ARFG*	*ARE*	*ARFG*	*ARE*	*ARFG*	*ARE*
Vanilla MTL	0.3444	1.1040	0.1336	0.8367	0.3497	0.9778
Baseline	0.0871	1.1032	0.0999	0.8356	0.1126	0.9864
MTA-F	**0.0437**	**1.0820**	**0.0364**	**0.8264**	**0.0310**	**0.9731**

Table 3: Average relative fairness gap (*ARFG*) and average relative error (*ARE*) on **UCI-Adult**, **German Credit Data** and **LSAC Law School** datasets, as defined in Section 4. Lower metric values indicate better overall fairness / accuracy across all tasks.

6.2.1 UCI-Adult Results. Here we adopt the MMD fairness loss as defined in Eq. (6b) and (7b) with Gaussian kernel.

Figure 5a-5b show the fairness-accuracy Pareto frontiers of each task separately, and that MTA-F is able to improve the per-task fairness-accuracy trade-off for both tasks. Note that these Pareto

		T_1 Err	T_1 FPRGap	T_2 Err	T_2 FPRGap
UCI-Adult	Vanilla MTL	0.1911	0.0715	0.1359	0.0091
	Baseline	0.1938	0.0186	0.1336	0.0020
	MTA-F	**0.1891**	**0.0083**	**0.1319**	**0.0016**
German Credit	Vanilla MTL	0.205	0.0150	0.220	0.0084
	Baseline	0.255	0.0879	**0.180**	0.0069
	MTA-F	**0.200**	**0.0033**	0.220	**0.0034**
LSAC Law School	Vanilla MTL	0.1555	0.0503	**0.1565**	0.0004
	Baseline	0.1568	0.0119	0.1580	0.0006
	MTA-F	**0.1540**	**0.0015**	0.1565	0.0004

Table 4: Per-task metrics for **UCI-Adult, German Credit Data** and **LSAC Law School** datasets.

(a) Fairness-accuracy Pareto frontier for **Task 1**.

(b) Fairness-accuracy Pareto frontier for **Task 2**.
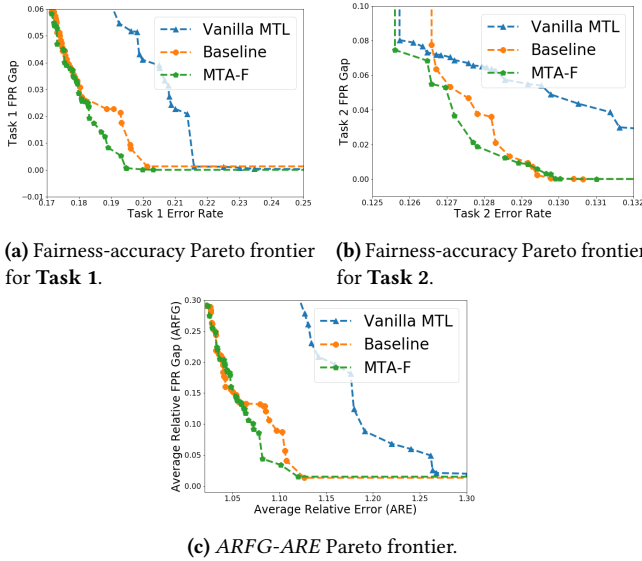
(c) *ARFG-ARE* Pareto frontier.

Figure 5: Pareto frontiers on UCI-Adult dataset. Lower-left indicates better Pareto optimality, i.e. better fairness-accuracy trade-off.

frontiers are only *marginal* frontiers in the sense that model runs that are Pareto optimal for Task 1 (i.e. points on Fig. 5a) may not correspond to Pareto optimal runs for Task 2 (i.e. points on Fig. 5b). Therefore, better per-task fairness-accuracy trade-offs do not guarantee overall better fairness-accuracy trade-off across all tasks.

On the contrary, the *ARFG-ARE* Pareto frontier proposed in Section 4, which consists of Pareto optimal runs for *ARFG* and *ARE*, is able to capture the trade-off between overall fairness and overall accuracy for multi-task learning problems. Figure 5c shows the *ARFG-ARE* Pareto frontier. We see that MTA-F is also able to improve the overall fairness-accuracy trade-off across all tasks, compared with Baseline fairness treatment and Vanilla MTL.

As we can see from Table 3, both Baseline and MTA-F are able to significantly reduce the *ARFG* metric compared with Vanilla MTL, confirming the effectiveness of fairness treatments using MMD loss. MTA-F achieves the lowest *ARFG* value. Note that the *ARE* metrics for all methods are slightly greater than 1, meaning that training both tasks together via multi-task learning generate slightly worse accuracy metrics than single-task learning, which suggests potential task training conflicts due to limited model capacities. However, MTA-F *Pareto dominates* Vanilla MTL and Baseline in that it has
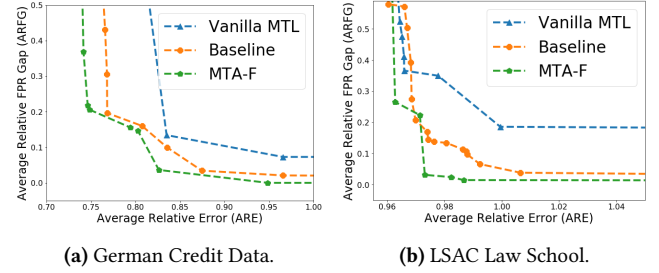

(a) German Credit Data.

(b) LSAC Law School.

Figure 6: *ARFG-ARE* Pareto frontiers.

lower values in both *ARFG* and *ARE* metrics, indicating a better overall fairness-accuracy trade-off across all tasks.

6.2.2 German Credit Data Results. We adopt FPR gap as the fairness loss $\mathcal{F}_{FPR}$ as defined in Eq. (6c) for equal opportunity on German Credit Data. Figure 6a and Table 3 and 4 show the *ARFG-ARE* Pareto frontier and numerical values, respectively, where the numerical results correspond to one of the points on the Pareto frontier chosen in the same way as described in the beginning of Section 6.2. We see that different from the results on UCI-Adult dataset, the *ARE* metrics are smaller than 1 on German Credit Data for all three methods, which means that the two tasks benefit from each other when trained together on a share model architecture. Wu et al. [50] call this *positive transfer*. The results show that with a different fairness loss function, MTA-F is still able to improve both *ARFG* and *ARE* over Baseline and Vanilla MTL, therefore improving the overall fairness-accuracy trade-off.

6.2.3 LSAC Law School Dataset Results. Here we test another choice for fairness loss: the correlation loss $\mathcal{F}_{Corr}$ defined in Eq. (6a). Figure 6b shows the *ARFG-ARE* Pareto frontier and Table 3 and 4 show the numerical results. On LSAC Law School dataset, training both tasks together with multi-task learning achieves on-par accuracy results as the single-task learning case (*ARE* metrics are close to 1). Our MTA-F method is able to reduce *ARFG* significantly while keeping ARE relatively the same, suggesting improved Pareto fairness compared with other methods.

7 CONCLUSION

In this paper, we present insights and improvements for the understudied trade-off between fairness and accuracy in multi-task learning. Notably, the accuracy goals across multiple tasks, which current multi-task learning algorithms optimize for, may not align well with the fairness goals. To understand and measure the overall fairness-accuracy trade-off as well as the trade-offs across tasks in a multi-task learning setting, we propose a new set of metrics, average relative fairness gap (*ARFG*) and average relative error (*ARE*), which quantifies the multi-dimensional trade-off in multi-task settings where fairness is a concern. We then propose a data-dependent *multi-task-aware fairness treatment (MTA-F)*, which adaptively decouples the fairness treatments in standard multi-task model architectures based on inter-task label distributions. MTA-F works with a number of fairness loss functions toward equal opportunity and equalized odds. Experimental results on three real-world datasets demonstrate the effectiveness of MTA-F in improving the fairness-accuracy trade-off for multi-task applications.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML*.

[2] Ananth Balashankar, Alyssa Lees, Chris Welty, and Lakshminarayanan Subramanian. 2019. What is Fair? Exploring Pareto-Efficiency for Fairness Constrained Classifiers. *ArXiv* abs/1910.14120 (2019).

[3] Trapit Bansal, David Belanger, and Andrew McCallum. 2016. Ask the gru: Multitask learning for deep text recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 107–114.

[4] Jonathan Baxter. 2000. A model of inductive bias learning. *Journal of artificial intelligence research* 12 (2000), 149–198.

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *KDD' 19*. 2212–2220.

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[7] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In *2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *NIPS 2016*.

[9] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[10] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[11] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. 2020. Just Pick a Sign: Optimizing Deep Multitask Models with Gradient Sign Dropout. *arXiv preprint arXiv:2010.06808* (2020).

[12] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*.

[13] Andrew Cotter, Michael Friedlander, Gabriel Goh, and Maya Gupta. 2016. Satisfying Real-world Goals with Dataset Constraints. (2016).

[14] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *ICML*.

[15] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[16] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. 2019. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439* (2019).

[17] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2020. Convergent Algorithms for (Relaxed) Minimax Fairness. *arXiv preprint arXiv:2011.03108* (2020).

[18] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[19] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *KDD' 15*. 259–268.

[20] Ross Girshick. 2015. Fast r-cnn. In *ICCV*. 1440–1448.

[21] Arthur Gretton, Karsten Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. 2008. A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368* (2008).

[22] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[24] James E Johndrow, Kristian Lum, et al. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13, 1 (2019), 189–220.

[25] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*.

[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*.

[27] Krishnateja Killamsetty, Changbin Li, Chen Zhao, Rishabh Iyer, and Feng Chen. 2020. A Reweighted Meta Learning Framework for Robust Few Shot Learning. *arXiv preprint arXiv:2011.06782* (2020).

[28] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *AIES '19*.

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.

[30] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *KDD' 18*. 1930–1939.

[31] David Madras, Elliot Creager, T. Pitassi, and R. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*.

[32] R Timothy Marler and Jasbir S Arora. 2004. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization* 26, 6 (2004), 369–395.

[33] Natalia Martínez, Martín Bertrán, and Guillermo Sapiro. 2019. Fairness With Minimal Harm: A Pareto-Optimal Approach For Healthcare. *CoRR* abs/1911.06935 (2019). arXiv:1911.06935 http://arxiv.org/abs/1911.06935

[34] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax Pareto fairness: A multi objective perspective. In *ICML*.

[35] A. Menon and R. Williamson. 2018. The cost of fairness in binary classification. In *FAT*.

[36] Elliot Meyerson and Risto Miikkulainen. 2018. Pseudo-task augmentation: From deep multitask learning to intratask sharing—and back. In *ICML*.

[37] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. 2019. Taking Advantage of Multitask Learning for Fair Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*.

[38] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *NIPS 2017*.

[39] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. 2019. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779* (2019).

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[41] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).

[42] Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. *ArXiv* abs/1706.05098 (2017).

[43] Candice Schumann, Xuezhi Wang, Alex Beutel, J. Chen, Hai Qian, and Ed Huai hsin Chi. 2019. Transfer of Machine Learning Fairness across Domains. *ArXiv* abs/1906.09688 (2019).

[44] Trevor Standley, Amir R Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2019. Which Tasks Should Be Learned Together in Multi-task Learning? *arXiv preprint arXiv:1905.07553* (2019).

[45] Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. 2020. Learning Fair Representations for Kernel Models, Vol. 108. PMLR, 155–166.

[46] Indrė Žliobaitė. 2015. On the relation between accuracy and fairness in binary classification. *ArXiv* abs/1505.05723 (2015).

[47] Xuezhi Wang, Nithum Thain, Anu Aradhana Sinha, Flavien Prost, Ed H. Chi, Jilin Chen, and Alex Beutel. 2021. Practical Compositional Fairness: Understanding Fairness in Multi-Component Recommender Systems. In *WSDM 2021*.

[48] Yuyan Wang, Zhe Zhao, Bo Dai, Christopher Fifty, Dong Lin, Lichan Hong, and Ed H Chi. 2020. Small Towers Make Big Differences. *arXiv preprint arXiv:2008.05808* (2020).

[49] Linda F Wightman. 1998. *LSAC national longitudinal bar passage study*. Law School Admission Council.

[50] Sen Wu, Hongyang R Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944* (2020).

[51] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782* (2020).

[52] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20, 75 (2019), 1–42.

[53] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*. PMLR, Fort Lauderdale, FL, USA.

[54] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *ICML*.

[55] Y. Zhang and Qiang Yang. 2017. A Survey on Multi-Task Learning. *ArXiv* abs/1707.08114 (2017).

[56] C. Zhao and F. Chen. 2019. Rank-Based Multi-task Learning for Fair Regression. In *2019 IEEE International Conference on Data Mining (ICDM)*. 916–925.

[57] Chen Zhao, Feng Chen, Zhuoyi Wang, and Latifur Khan. 2020. A Primal-Dual Subgradient Approach for Fair Meta Learning. *arXiv preprint arXiv:2009.12675* (2020).

[58] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey Gordon. 2020. Conditional Learning of Fair Representations. In *ICLR*.

[59] H. Zhao and G. Gordon. 2019. Inherent Tradeoffs in Learning Fair Representation. *ArXiv* abs/1906.08386 (2019).

[60] Han Zhao and Geoffrey J Gordon. 2019. Inherent tradeoffs in learning fair representations. *arXiv preprint arXiv:1906.08386* (2019).

A APPENDIX

A.1 Datasets

CelebA Dataset: The CelebA dataset[4] [29] contains public domain images of public figures. It has about 200,000 64×64 face images of celebrities, each with 5 landmark locations, and 40 binary attributes including gender. We pick two out of the 40 attributes, *Attractive* and *Smiling* as the two tasks for the multi-task problem. The dataset is randomly split into training data of size 162,770 and test data of size 19,867.

UCI-Adult Dataset: The UCI-Adult dataset[5] contains census information with the training / test data containing information of 32,561 / 16,281 individuals. We pick the default binary label of the dataset, *Income* > \$50,000 or ≤ \$50,000 as Task 1; and another quantity *Capital Gain* from the dataset, transform it into a binary label *Capital Gain* > 0 or = 0 as Task 2, to formulate a multi-task problem. The rest of the attributes, including age, working hours per week, education, occupation and race, are used as input features to the model. We use the original train/test split with the dataset.

German Credit Data: The German Credit Data[6] provides a set of attributes for 1000 individuals, including credit history, credit amount, gender, age, etc., and the corresponding credit risk for that individual. We construct two binary tasks from the dataset: *Good loans vs. Bad loans* as Task 1, and *Credit amount* > 2000 or ≤ 2000 as Task 2. There are 16 attributes that we use as features. We randomly split the dataset into training data of size 800 and test data of size 200.

LSAC Law School: The LSAC dataset[7] [49] is generated from the survey conducted by the Law School Admission Council in the United States. It contains information on 21,790 law students such as their entrance exam scores (LSAT), their undergrad grade-point average (GPA) collected prior to law school, gender and race etc. For each student, we predict whether they passed the bar exam as Task 1, and whether they have a high first year average grade (defined by z-score > 0) as Task 2. The dataset is randomly split into training data of size 18585 and test data of size 7966.

A.2 Experiment Setup

A.2.1 Architecture. For all datasets used in the paper, we adopt the standard multi-task architecture with shared representation learning and task sub-networks as in Fig. 1. For the CelebA dataset in Section 3.3, we use ResNet-18 [23] without the final layers as a shared representation, and two fully connected layers of size [1000, 500] as head layers for each task. For UCI-Adult dataset, the model has a single shared hidden layer with 64 shared hidden units and task-specific fully-connected hidden layers of size 32 with ReLU activation for task-specific networks, where categorical features are encoded as 40-dimensional embeddings. For German Credit Data and LSAC dataset, we pick a smaller architecture due to their smaller sample sizes: a single shared hidden layer of size 32 and single task-specific hidden layer of size 16 for each task, both with ReLU activations. Categorical features are encoded as 10-dimensional embeddings.

A.2.2 Tuning. The same set of hyperparameter values is used for each method, across different fairness and accuracy weights. Adagrad optimizer is used with learning rate lr and number of epochs as hyperparameters.

We use batch size of 512 for CelebA, UCI-Adult and LSAC dataset, and 128 for German Credit Data. For MTA-F, there is one more set of hyperparameters, $\{r_1, r_2\}$ which controls the ratio between task-specific fairness losses $\mathcal{F}_t^{head}$ and shared fairness losses $\mathcal{F}_t^{share}$ for Task $t = 1, 2$. To ensure we are not over-tuning MTA-F over other methods due to its additional hyperparameters, for all methods, hyperparameters (i.e. $\{lr, epochs\}$ for Vanilla MTL and Baseline, and $\{lr, epoch, r_1, r_2\}$ for MTA-F) are tuned over the *same* number of runs (50k). During the 50k tuning runs, we vary the Task 1 fairness loss weights λ_1, λ_2 in $[0, 5]$ and accuracy loss weights w_1 and $w_2 = 1 - w_1$ in $[0, 1]$.

Once the hyperparameters are selected, we perform another 20k runs for each method with task fairness weights and loss weights varying as above, evaluate each of them on test dataset, and report the Pareto frontier of the test metrics (i.e. the Pareto optimal solutions from the 20k runs) including per-task fairness (FPR gap) and accuracy (classification error rate), as well as the ARFG and ARE metrics defined in Section 4.

A.2.3 Code. The code for the experiments in Section 6 is available at: https://github.com/kdd-2021-yuyan/pareto-fairness-mtl.