PDFM: A Primal-Dual Fairness-Aware Framework for Meta-learning

Chen Zhao, Feng Chen Member, IEEE, Zhuoyi Wang, and Latifur Khan Senior Member, IEEE

Abstract—The problem of learning to generalize to unseen classes during training, known as few-shot classification, has attracted considerable attention. Initialization based methods, such as the gradient-based model agnostic meta-learning (MAML) [1], tackle the few-shot learning problem by “learning to fine-tune”. The goal of these approaches is to learn proper model initialization, so that the classifiers for new classes can be learned from a few labeled examples with a small number of gradient update steps. Few shot meta-learning is well-known with its fast-adapted capability and accuracy generalization onto unseen tasks. Learning fairly with unbiased outcomes is another significant hallmark of human intelligence, which is rarely touched in few-shot meta-learning. In this work, we propose a Primal-Dual Fair Meta-learning framework, namely PDFM, which learns to train fair machine learning models using only a few examples based on data from related tasks. The key idea is to learn a good initialization of a fair model’s primal and dual parameters so that it can adapt to a new fair learning task via a few gradient update steps. Instead of manually tuning the dual parameters as hyperparameters via a grid search, PDFM optimizes the initialization of the primal and dual parameters jointly for fair meta-learning via a subgradient primal-dual approach. We further instantiate examples of bias controlling using mean difference and decision boundary covariance [2] as fairness constraints to each task for supervised regression and classification, respectively. We demonstrate the versatility of our proposed approach by applying our approach to various real-world datasets. Our experiments show substantial improvements over the best prior work for this setting.

Index Terms—dual subgradient, dual decomposition, meta-learning, fairness, few shot.

1 INTRODUCTION

In contrast to the conventional machine learning systems, the ability to learn from a handful of examples is one of the critical characteristics of human intelligence. Learning quickly yet remains a daunting challenge for artificial intelligence, which receives significant attention from the machine learning community, especially when it needs to transfer knowledge from a given distribution of tasks onto unseen ones. Overcoming this limitation can have a broad impact on artificial intelligence, as it can save the expensive preparation of large training samples, often humanly annotated, needed by current machine learning methods [3]. To address the challenge of fast adaptation, meta-learning (a.k.a learning to learn) leverages the transferable knowledge learned from previous tasks, then adapts to new environments rapidly with a few training examples. The goal of a few-shot meta-learning problem is to minimize generalization error across a distribution of tasks with few training examples (i.e. few-shot). This technique has demonstrated success in both supervised learning, such as few-shot regression [1], [4] and classification [5], [6], and reinforcement learning [7] settings.

There are several lines of meta-learning algorithms for base learners, nearest neighbors based methods [5], [6] which address the problem by “learning to compare”; recurrent network based methods [8] that instantiate the transferable knowledge as latent representations, and gradient-based methods [1], [9], [10], [11], [12] that aim to learn proper model initialization for all tasks, such that the summation query errors are minimized and further the meta-parameter is adapted to novel tasks using a few optimization steps.

Despite their early success in the few-shot application, to the best of our knowledge, most of the existing meta-learning algorithms ignore to mitigate the notion of fairness in tasks and thus lack the capability of fairness generalization on new tasks.

Learning with fairness can be defined as follows: (1) people that are similar in terms of non-sensitive characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by non-sensitive characteristics [13]. The first condition is related to direct discrimination and the second ensures that there is no indirect discrimination. The Equality Act [14] calls the sensitive characters as protected characteristics.

Machine learning models trained to output prediction based on historical data will naturally inherit the past biases. With the biased input, the main goal of training an unbiased model is to make the output fair. In other words, the predictions are statistically independent of protected variables (e.g. race and gender) [13]. These models may be enhanced by attempting to mask some attributes to the decision-maker, however, as many attributes may be correlated with the protected one [15]. Moreover, techniques in the area of fairness learning are incapable of adapting deep learning models on fairness to new tasks. The motivation of this paper is: can we develop meta-learning methods that adapt deep learning models on both generalization accuracy and fairness to unseen tasks?

In this paper, we bridge areas of few-shot meta-learning and unfairness prevention, and formulate this problem by enhancing the meta-learning model with fairness constraints. More concretely, for each task during the training stage, it is constrained with a task-specific fair inequality, which
we propose a pair of primal-dual meta-parameters, which can be decomposed into some sub-problems, and proportion of the population as a whole. To this end, we process, the overall proportion of members in a protected prevention approach is shown. In the meta-training stage, learning approach. Besides, we instantiate examples of mean and theoretically grounded analysis for the proposed meta-algorithm (meta-level) plays a key role in our work. The condition on task predictions. In the support set during the training process, the overall proportion of members in a protected group would receive predictions, which are identical to the proportion of the population as a whole. To this end, we resort to a dual subgradient algorithm with an averaging scheme dual subgradient algorithm presented on the right. Query losses and fairness are gathered and utilized to update the meta-initialization pair. (Right) A few-shot unfairness prevention approach is shown. In the meta-training stage, in each task, support loss is optimized under a fairness constraint which performs a trade-off between losses and fairness.

ensures the independent effect of the protected variable on task predictions. In the support set during the training process, the overall proportion of members in a protected group would receive predictions, which are identical to the proportion of the population as a whole. To this end, we resort to a dual subgradient algorithm with an averaging scheme for each task. It approximately optimizes a pair of task-specific primal and dual parameters, which minimizes the summation of query losses and fairness constraints are satisfied simultaneously. In contrast to the grid search technique, we consider Lagrange multipliers as dual variables that they are optimized to minimize the duality gap between the primal and dual functions.

Furthermore, instead of updating the meta-parameter from the outer loop (such as MAML [1]), in our work, inspired by the concept of resource allocation from economics, we propose a pair of primal-dual meta-parameters, which could be optimized iteratively through a dual decomposition and divided into broadcast and gather steps. We apply such decomposition to leverage the observation that problems can be decomposed into some sub-problems, and then introduce fairness constraints to enforce the notion of agreement between solutions to the different issues. The agreement constraints are incorporated using Lagrange multipliers, and an iterative algorithm is used to minimize the resulting dual. As shown in Figure 1, the interplay between the inner-algorithm (task-level) and the meta-algorithm (meta-level) plays a key role in our work. The former one is used to compute a good approximation of the meta-subgradient, and supplied to the latter. Finally, another key merit of this paper is that we derive an efficient and theoretically grounded analysis for the proposed meta-learning approach. Besides, we instantiate examples of mean difference (MD) and decision boundary covariance (DBC) as fairness constraints for justification of supervised regression and classification tasks, respectively. We demonstrate the versatility of our proposed approach on a variety of real-world datasets and extensive experiments to show substantial improvements over the best prior work.

In summary, the main contributions of this paper is threefold:

- We propose a novel fair meta-learning framework, in which a good pair of meta-parameters is approximately optimized. Our framework efficiently controls biases for each task, and ensures the generalization capability of both accuracy and fairness onto unseen tasks.
- We further implement optimized strategies for inner loop and meta-subgradient update. Specific and theoretically grounded analysis for the proposed strategies justifies the efficiency and effectiveness of them.
- Finally, we validate the performance of our approach with state-of-the-art techniques on real-world datasets. Our results demonstrate the proposed approach is capable of mitigating biases, generalizing accuracy and fairness to unseen tasks with the minimized input training data.

This paper is organized as follows. In Section 2, some related works are referred. Section 3 provides details of our proposed PDFM framework. In Section 4, we discuss the theoretically grounded analysis for the learning approach. In Section 5, we instantiate two examples to justify our approach on few-shot fair regression and classification problems, respectively. Section 6 describes the evaluation settings. In Section 7, we conduct experiments on real-world benchmarks compared with cutting-edge techniques. Finally, the paper is concluded in Section 8.

### 2 RELATED WORK

Meta-Learning based on few-shot studies that trained models and make them able to quickly adapt to new tasks under a few labeled samples, several recent approaches have made significant progress in meta-learning [18], [19], [20], [21], which could be broadly classified into three categories: metric-based, network-based, and gradient-based. The goal of metric-based approaches aim to learn an embedding space between the query and support examples, where similar instances are closer and different ones are further apart. For example, the Matching-Net [5] employed ideas from k-nearest neighbors and metric learning based on a feature encoder to extract embedding in the context of the support set, and Prototypical networks [6] learn a metric space in which classification is able to be performed by computing Euclidean distances to prototype representations of each class. The network-based approaches execute fast adaptation into network architecture by generating input-conditioned weights, or adding an external memory.

In this work, we focus on the third type, the gradient descent based algorithms [1], [9], [12], [27], [28], which aim to meta-learn an initial set of weights for neural networks, and quickly adapted to new task with just a few steps.
of gradient descent. Such approach could achieve good generalization over new tasks by encoding prior knowledge, existing work such as Franceschi et al. [29] also provides convergence guarantees for gradient-based meta-learning with strongly-convex functions. Despite methods in this area that have been shown effective for adaption of deep learning models on generalization accuracy to new tasks, our experiments show such state-of-the-arts have difficulties in adapting to fairness.

Fairness researchers develop machine learning algorithms that would produce predictive models, ensuring that those models are free from biases. Standard predictive models, induced by machine learning and data mining algorithms, may discriminate groups of entities because (1) data bias comes from data being collected from different sources, or (2) dependence on sensitive attributes was identified in the data mining community [30]. Based on the taxonomy by tasks, fairness learning can be typically categorized to classification [2], [31], [32], regression [30], [33], [34], clustering [35], and recommendation [36] works. Even though techniques for unfairness prevention on classification were well developed, to the best of our knowledge, the majority of existing fairness-aware machine learning algorithms are under the assumption of giving abundant training examples. Learning quickly, however, is another significant hallmark of human intelligence.

Several recent approaches have been developed in fair meta-learning [37], [38], [39]. These methods focus on studies of fairness generalization onto unseen tasks by adding an uniformed fairness regularizer to each task. Concretely, Lagrange multipliers were considered as dual variables and hence, instead of grid search, they are optimized to minimize the duality gap between the primal and dual functions.

### 3 Methodology

In this section, the problem of controlling unfairness is formulated, which underlies the training of a constrained Model-Agnostic Meta-Learning (MAML) [1]. Different from [1], in the proposed framework, each task is fed with a task-specific fairness constraint. The vector of Lagrangian multipliers are considered as dual variables. We then introduce a task-level dual subgradient algorithm using an averaging scheme which provides an approximate solution to the inner problem. In the outer loop, the dual decomposition strategy is therefore applied to update the pair of primal-dual meta-parameters.

#### 3.1 Problem Setting

Let \( Z = \mathcal{X} \times \mathcal{Y} \) be the data space, where \( \mathcal{X} \subset \mathbb{R}^n \) is the input space, \( \mathcal{Y} \) means the output space, and \( N \) is the number of classes. Meta-learning for few-shot learning aims to train a meta-learner which is able to learn on a large number of various tasks from a small amount of data. Gradient based meta-learning frameworks, such as Model-Agnostic Meta-Learning (MAML) [1], lead to state-of-the-art performance and fast adaptation to unseen tasks. More precisely, the goal of MAML is to estimate a good meta-parameter \( \theta \in \Theta \) such that the summation of empirical risks for each task is minimized. Throughout this work, the \( \Theta \) will be a closed, convex, non-empty subset of an Euclidean space.

In this work, we consider a collection of supervised learning tasks \( T = \{ (\mathcal{D}_t^S, \mathcal{D}_t^Q) \}_{t=1}^T \) which distributions over \( Z \) and \( T \) is denoted as the number of tasks. \( T \) is often referred to as a meta-training set as well as an episode \( (\mathcal{D}_t^S, \mathcal{D}_t^Q) \) explicitly contains a pair of a support \( (i.e., \mathcal{D}_t^S) \) and a query \( (i.e., \mathcal{D}_t^Q) \) data sets. For each task \( t \in \{1, 2, ..., T\} \), we let \( \{x_{t,i}, y_{t,i}\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y}) \) be the corresponding task data, and \( m \) is the number of datapoints in the support set. For example, standard few-shot learning benchmarks evaluate model in \( N \)-way \( K \)-shot classification tasks and thus \( m = N \times K \) indicates, in the support set of the \( t \)-th task, it contains \( N \) categories and each consists of \( K \) datapoints. We emphasize that we need to sample without replacement, \( i.e., \mathcal{D}_t^S \cap \mathcal{D}_t^Q = \emptyset \).

To study fairness generalization problem under meta-learning frameworks, a fairness constraint, \( g_t(\theta_t) \leq 0 \), is considered in each task, where \( t \) indicates task index. In researches of bias prevention, convexity of the constraint receives increasing attention in the machine learning fields [35], [40], [41]. For this purpose, in this paper, we assume that convexity of task constraints always holds.

#### 3.2 Model-Agnostic Meta-Learning with constraints

Meta-learning approaches for few-shot learning aim to minimize the generalization error across a distribution of tasks sampled from a task distribution. It is often assume that the support and query sets of a task are sampled from the same distribution. In our work, for each single task, the
objective is to minimize the predictive error $\mathcal{L}^{inner}$ such that it is constrained by $g_i$:

$$\theta'_i = \arg \min_{\theta_i \in \Theta} f_i(\theta_i) := \mathcal{L}^{inner}(D_i^S, \theta'_i; \theta)$$

$$\text{subject to } g_i(D_i^S, \theta'_i) \leq 0$$

(1)

where $\mathcal{L}^{inner} : \mathbb{R}^n \to \mathbb{R}$ is a loss function, such as cross-
entropy loss for classification problems and $\theta_i$ is the model parameter at task $i$, which is initialized with $\theta$. $\mathcal{L}(\mathcal{D}_i, \theta)$ corresponds to one or multiple steps of gradient descent initialized at $\theta$. $g : \mathbb{R}^n \to \mathbb{R}$ is an appropriate complexity function ensuring the existence and the uniqueness of the above minimizer. A point $\theta_i$ in the domain of the problem is feasible if it satisfies the constraint $g_i(\theta_i) \leq 0$.

**Assumption 1. (Task Loss and Constraint).** Let $f_i(\theta_i)$ be a convex-real-valued function for any $\theta_i \in \Theta$. Let $\Gamma(\Theta) = \{ \gamma \in \Gamma(\Theta) \}$ be a set of proper, closed and convex function over $\Theta$ and $g_i \in \Gamma(\Theta)$ be such that, for any $\theta_i \in \Theta$, $g_i(\theta_i)$ is convex over $\mathbb{R}^n$, $\inf_{\theta_i \in \Theta} g_i(\theta_i) = 0$ and, for any $\theta_i \notin \Theta$, $\operatorname{dom}(g_i(\theta_i)) = \emptyset$.

The optimal value of the Eq. (1) is denoted as $f^*_i$, which is assume to be finite and is achieved at an optimal and feasible solution $\theta^*_i$, i.e. $f^*_i = f_i(\theta^*_i)$. The goal of training a single task is to output local parameter $\theta^*_i$ given the meta-parameter $\gamma$ such that it minimizes the task loss $f_i(\theta^*_i)$ subject to the task constraint $g_i(\theta^*_i) \leq 0$. Next, to update the meta-parameter, we minimize the generalization error $L^{meta}$ using query sets across every tasks in the batch such that query constraints for all tasks are satisfied. Formally, the learning objective across all tasks is

$$\min_{\theta \in \Theta} \mathcal{L}^{meta} = \sum_{t=1}^{T} f_t(\theta'_t; \gamma) := \sum_{t=1}^{T} \mathcal{L}^{inner}(D_t^Q, \mathcal{L}(D_t^S, \theta_t))$$

$$\text{subject to } \sum_{t=1}^{T} g_t(D_t^Q, \mathcal{L}(D_t^S, \theta'_t)) \leq 0$$

(2)

where $\theta'_t = \arg \min_{\theta'_t \in \Theta} g_t(\theta'_t) \in D_t^S$ is a local optimum of each task $t$. Here, for the purpose of optimization with simplicity, the constraint of Eq. (2) is approximated, which originally takes the form of a sequence $g_t(D_t^Q, \mathcal{L}(D_t^S, \theta_t)) \leq 0$, where $t = 1, ..., T$. In this setting, the meta-objectives and the consequently their subgradients used by the meta-algorithm are dependent on the properties of the inner algorithm. We will show the algorithm details and analysis in the following sections.

### 3.3 Primal and Dual Formulation

Our approach aims to optimize a pair of meta-parameters (i.e. primal and dual variables) as model initialization, instead of using the conventional grid search technique. It consists of two nested primal-dual algorithms, one operating within each task and another across all tasks. In this section, we briefly recall from the primal-dual interpretation of the algorithm framework and such interpretation will be used in the subsequent analysis for both inner and meta problems.

To recover the primal optimal solution of Eq. (1), we use the Lagrange duality theory to relax the primal problem by its constraints, and the Lagrangian function is

$$L(\theta, \mu) = f(\theta) + \mu^T g(\theta)$$

where $\mu_\iota \in \mathbb{R}^m$ is the Lagrange multiplier (or dual variable). The dual function hence is defined as

$$q(\mu) = \inf_{\theta \in \Theta} L(\theta, \mu) = \inf_{\theta \in \Theta} \{ f(\theta) + \mu^T g(\theta) \}$$

Since the dual function $q(\mu)$ is a pointwise affine function of $\mu$, we can thus maximize the dual function to obtain a tightest lower bound of the optimal primal $f^*_i$ and through out this paper, we assume $f^*_i$ is finite. The goal is to obtain the dual optimal value $q^*_i$ at $\mu^*_i$, such that the duality gap, i.e. $f^*_i = g^*_i$, is as small as possible. Zero duality gap thus indicates that the optimal values of the primal and dual problems are equal, i.e. $f^*_i = g^*_i$. Due to space limit, the same idea is applied to solve Eq. (2).

**Algorithm 1 Update Model-parameters of Task $t$ using Dual Subgradient Method**

**Require:** $\theta \in \Theta, \mu \in \mathbb{R}^m$ prime and dual initializations

**Require:** $\alpha > 0, \gamma > 0$: learning rate

**Require:** $q > 0$: a small number of subgradient update steps

1: $\mu_t^i \leftarrow \mu_t^i, \theta_t^i \leftarrow \theta$
2: Initialize an empty array $a = \emptyset$
3: for $k = 1, 2, \ldots$ do
4:  for $q = 1, 2, \ldots$ do
5:  Evaluate the primal feasible subgradient $\nabla \in \nabla_{\theta_t^i} \{ f_t(\theta_t^{i-1}) + \mu_t^{i-1} g_t(\theta_t^{i-1}) \}$
6:  $\theta_t^k \leftarrow \theta_t^{i-1} - \gamma T \nabla$
7:  end for
8: end for
9: Evaluate $\theta_t^k$ by taking the average of previous vectors in $a$: $\theta_t^k = \frac{1}{a} \sum_{i=1}^{a} \theta_t^k$
10: Calculate the subgradient iterate $g_k = g_t(\theta_t^k)$
11: Update the dual solution $\mu_t^k = \mu_t^{k-1} + \alpha T g_k$
12: end for
13: return $\theta_t^k, \mu_t^k$, where $\theta_t^k = \theta_t^k, \mu_t^k = \mu_t^k$

### 3.4 Update Task-Specific Model-Parameters via Dual Subgradient

In order to find a good pair of meta-parameters $(\theta, \mu) \in \Theta \times \mathbb{R}^m$, such that constraints of all tasks can be satisfied and generalization error is minimized. To this end, in this section, we provide an approximate solution to the inner task of Eq. (1) by proposing a task-level dual subgradient algorithm. This method takes in the meta-parameter pair from the previous outer (or meta) loop and the task-specific (or local) primal and dual parameters are then iterative updated using the support data of the single task.

In the subsequent development, to solve the dual problem of Eq. (1) for a single task, we consider a subgradient algorithm with a constant step size $\alpha > 0$ to update the dual solution iteratively:
where \([u]^+\) denotes the projection of \([u]\) on the nonnegative orthant in \(\mathbb{R}^m_+\), namely \([u]^+ = (\max\{0, u_1\}, ..., \max\{0, u_m\})\), \(k = 1, 2, ...\) is the index of iterations, subscript \(t\) is the task index number, and \(t^0 > 0\) is an initial dual point. The subgradient iterate \(g_k\) is a subgradient of the dual function \(q_t\) at a given \(t^k \geq 0\):

\[
g_k = g_t(\hat{\theta}^k_t) \in \partial q_t(\hat{\theta}^k_t) = \text{conv}\{(g_t(\hat{\theta}^k_i) | \hat{\theta}^k_i \in \Theta^k_t)\}
\]

where \(\Theta^k_t = \{\hat{\theta}^k_i \in \Theta | q_t(\hat{\theta}^k_i) = f_t(\hat{\theta}^k_i) + (\mu^k)^T g_t(\hat{\theta}^k_i)\}\) and conv\(\{Y\}\) denotes the convex hull of a set \(Y\). Although a general dual subgradient method can generate near-optimal dual solutions with a sufficiently small step size and a large number of iterations, it does not directly provide primal solutions which are of our interest. But even worse, it may fail to produce any useful information. Motivated by this reason, we apply an averaging scheme to the primal sequence \(\{\hat{\theta}^k_t\}\) to approximate primal optimal solutions. In particular, the sequence \(\{\hat{\theta}^k_t\}\) is defined as the averages of the previous vectors through \(\hat{\theta}^0_t\) to \(\hat{\theta}^{k-1}_t\):

\[
\hat{\theta}^k_t = \frac{1}{k} \sum_{i=1}^{k-1} \hat{\theta}^i_t, \quad \forall k \geq 1
\]

where the corresponding primal feasible iterate \(\theta^k_t\) is given by any solution of the set.

\[
\theta^k_t^* \in \arg\min_{\theta^1_t \in \Theta} \{f_t(\theta^k_t-1) + (\mu^k)^T g_t(\theta^k_t-1)\}
\]

As the subgradient method can usually generate a reasonably estimation of the dual optimal solutions within several iterations, approximate primal solutions are obtained accordingly. The constant stepsize \(\alpha\) is a simple hyperparameter for controlling, then through choosing an appropriate value of \(\alpha\), the proposed Algorithm 1 is able to approach the optimal value arbitrarily close within a small finite number of steps.

Moreover, the dual subgradient schemes can be applied efficiently to approximate a solution to Eq. 7. Specifically, it returns a good pair of task-level primal and dual parameters \(\theta_t^*, \mu_t^*\). In the following section, due to the decomposable structure of the meta-learning framework for few-shot learning, meta-parameters \((\theta, \mu)\) are updated by minimizing the summation of query losses across all training tasks.

### 3.5 Update Meta-parameters via Dual Decomposition

Dual decomposition has been used since [42], and has been applied in engineering, such as in rate control for communication networks [43], in networking problems for simultaneous routing and resource allocation [16]. The dual decomposition method was also used to solve large computationally intractable problems with multiple cooperative agents, which resulted in an easily implemented algorithm, in a decentralized manner [17].

In this work, inspired by the concept of resource allocation from economics [16, 17], our model’s goal is to estimate a good pair of primal-dual weight initialization

\[\mu^k_t = [\mu^k_{t-1} + \alpha^T g_t]\]

where \([u]^-\) denotes the projection of \([u]\) on the nonnegative orthant in \(\mathbb{R}^m_+\), namely \([u]^- = (\max\{0, u_1\}, ..., \max\{0, u_m\})\), \(k = 1, 2, ...\) is the index of iterations, subscript \(t\) is the task index number, and \(\mu^0_t > 0\) is an initial dual point. The subgradient iterate \(g_k\) is a subgradient of the dual function \(q_t\) at a given \(\mu^k_t \geq 0\):

\[
g_k = g_t(\hat{\theta}^k_t) \in \partial q_t(\hat{\theta}^k_t) = \text{conv}\{(g_t(\hat{\theta}^k_i) | \hat{\theta}^k_i \in \Theta^k_t)\}
\]

where \(\Theta^k_t = \{\hat{\theta}^k_i \in \Theta | q_t(\hat{\theta}^k_i) = f_t(\hat{\theta}^k_i) + (\mu^k)^T g_t(\hat{\theta}^k_i)\}\) and conv\(\{Y\}\) denotes the convex hull of a set \(Y\). Although a general dual subgradient method can generate near-optimal dual solutions with a sufficiently small step size and a large number of iterations, it does not directly provide primal solutions which are of our interest. Even worse, it may fail to produce any useful information. Motivated by this reason, we apply an averaging scheme to the primal sequence \(\{\hat{\theta}^k_t\}\) to approximate primal optimal solutions. In particular, the sequence \(\{\hat{\theta}^k_t\}\) is defined as the averages of the previous vectors through \(\hat{\theta}^0_t\) to \(\hat{\theta}^{k-1}_t\):

\[
\hat{\theta}^k_t = \frac{1}{k} \sum_{i=1}^{k-1} \hat{\theta}^i_t, \quad \forall k \geq 1
\]

where the corresponding primal feasible iterate \(\theta^k_t\) is given by any solution of the set.

\[
\theta^k_t^* \in \arg\min_{\theta^1_t \in \Theta} \{f_t(\theta^k_t-1) + (\mu^k)^T g_t(\theta^k_t-1)\}
\]

As the subgradient method can usually generate a reasonably estimation of the dual optimal solutions within several iterations, approximate primal solutions are obtained accordingly. The constant stepsize \(\alpha\) is a simple hyperparameter for controlling, then through choosing an appropriate value of \(\alpha\), the proposed Algorithm 1 is able to approach the optimal value arbitrarily close within a small finite number of steps.

Moreover, the dual subgradient schemes can be applied efficiently to approximate a solution to Eq. 7. Specifically, it returns a good pair of task-level primal and dual parameters \(\theta_t^*, \mu_t^*\). In the following section, due to the decomposable structure of the meta-learning framework for few-shot learning, meta-parameters \((\theta, \mu)\) are updated by minimizing the summation of query losses across all training tasks.

Algorithm 2 The Duality-MAML Algorithm

Require: \(p(T)\): distribution over tasks
Require: \(\eta > 0, \beta > 0\): learning rate

1: randomly initialize primal and dual meta-parameter, i.e. \(\theta \in \Theta\) and \(\mu \in \mathbb{R}^m_+\)
2: while not done do
3: sample batch of tasks \(T_t \sim p(T), t = 1, 2, ..., T\)
4: for all \(T_t = \{D^C_t, D^Q_t\}\) do
5: Sample datapoints \(D^C_t = \{x_t, y_t\}\) from \(T_t\)
6: Compute adapted primal-dual parameters \(\theta_t^*\) and \(\mu_t^*\) using \(D^C_t\) by applying Algorithm 1
7: Sample datapoints \(D^Q_t = \{x_t, y_t\}\) from \(T_t\) for the meta-update, where \(D^C_t \cap D^Q_t = \emptyset\)
8: Evaluate query loss \(f_t(\theta_t^*)\) and query constraint \(g_t(\mu_t^*)\) using \(D^Q_t\)
end for
10: Update \(\theta\) and \(\mu\) using Eq. (7) \(\triangleright Update\ Meta-parameters.
11: end while

\((\theta, \mu)\), such that both the meta-loss across tasks is minimum and constraints of all tasks are also satisfied. To this end, we update the pair of primal-dual initialization iteratively using a dual decomposition method that is normally considered as a special case of Lagrangian relaxation [44]. This method is typically simple and efficient, which can be divided into two steps for each iteration, i.e., broadcast and gather. In the broadcast step, the meta-dual parameter \(\mu_t\) is sent to each of tasks \(T_t\). Through Algorithm 1 local primal, and dual parameters \(\theta_t\) and \(\mu_t\) of a single task are iteratively optimized using few-shot support data. Query loss \(f_t(D^Q_t, \theta_t^*)\) and fairness estimate \(g_t(D^Q_t, \theta_t^*)\), therefore, are evaluated using query data set. In the gather step, both query losses and fairness estimates collected across all tasks are applied to update primal and dual meta-parameters,

\[
\theta^{s+1}_t \in \arg\min_{\theta^s_t} \sum_{t=1}^{T} f_t(\theta^s_t; \theta^*) + \mu^s \sum_{t=1}^{T} g_t(\theta^s_t; \theta^*)
\]

\[
\mu^{s+1} = [\mu^s + \beta \sum_{t=1}^{T} g_t(\theta^s_t)]^+
\]

where \(s = 1, 2, ..., \) is the index of the outer iteration and \(\beta > 0\) is the stepsize. The full algorithm of the proposed approach is outlined in Algorithm 2.

### 4 Theoretical Analysis

Recall that the proposed averaging scheme used to approximate the task-specific primal-dual parameter pair is built upon the dual subgradient method with a constant stepsize. We denote the dual feasible set as \(M = \{\mu_t | \mu_t \geq 0, \infty < g_t(\mu_t) < \infty\}\), and for every fixed \(\mu_t \in M\), we have the solution set \(C \subset \Theta\) for \(g_t(\mu_t)\).

**Assumption 2.** (Slater Condition and Bounded Subgradients) The convex set \(\Theta\) is compact (i.e. closed and bounded). There exists a Slater point \(\bar{\theta} \in \Theta\), such that \(g_t(\bar{\theta}) < 0, \forall j = 1, 2, ..., m,\) and exists \(L > 0, L \in \mathbb{R}\), such that \(|g_k| < L, \forall k \geq 0\).

When \(f_t^*\) is finite, the Slater condition is sufficient for the existence of a dual optimal solution, and therefore the
proposed task adaptation approach efficiently reduces the amount of feasibility violation at the approximate primal solutions. Furthermore, intuitively, bounded subgradients in Assumption 2 is satisfied when $L = \max_{\hat{\theta}_i \in \Theta} ||g_i(\hat{\theta}_i)||$.

**Lemma 1.** If Assumption 1 and the continuity of $f_i(\theta_i)$ and $g_i(\theta_i)$ hold, there exists at least one optimal solution $\theta_{\mu} \in \mathcal{C}$. Furthermore, $\theta_{\mu}$ is unique if $f_i(\theta_i)$ is strictly convex, otherwise there may be multiple solutions.

Lemma 1 is easily proved using the Weierstrass Theorem proposed in [45]. Next, for the averaged primal sequence $\{\hat{\theta}_k^T\}$, we show that it always converges when $\Theta$ is compact.

**Proposition 1.** Under Assumption 2 when the convex set $\Theta$ is compact, let the approximate primal sequence $\{\hat{\theta}_k^T\}$ be the running averages of the primal iterates given in Eq.(5). Then $\{\hat{\theta}_k^T\}$ can converge to its limit $\hat{\theta}_*^T$.

The proof of Proposition 1 is in Appendix A. The following proposition [46] provides bounds on the feasibility violation and the primal cost of the running averages $f(\hat{\theta}_k^T)$, where the bounds are given per iteration $k$. Its proof is given in Appendix B.

**Proposition 2.** Let the dual sequence $\{\mu_k^T\}$ is generated through Eq.(5) and the vectors $\hat{\theta}_k^T$ for $k \geq 1$ be the averages in Eq.(5). Under the assumptions that $f_i(\theta_i)$ and $g_i(\theta_i)$ are convex, $\Theta$ is a convex compact set. For all $k \geq 1$, we have

1. An upper bound on the amount of constraint violation of the vector $\hat{\theta}_k^T$, i.e. $||g_i(\hat{\theta}_k^T)|| \leq \frac{\mu_k^T}{\kappa_n}$,
2. If the bounded subgradients in Assumption 2 holds, then $\forall k \geq 0$, we have an upper bound of $f_i(\hat{\theta}_k^T)$ such that $f_i(\hat{\theta}_k^T) \leq f_i^* + \frac{(||\mu_k^T||^2 + \alpha L^2)}{2}\kappa_n$,
3. and a lower bound $f_i(\hat{\theta}_k^T) \geq f_i^* - ||\mu_k^T|| \cdot ||g_i(\hat{\theta}_k^T)||$.

where $f_i^*$ and $\mu_k^*$ are the primal and dual optimal solutions of task $t$, respectively.

Furthermore, since the proposed Algorithm 2 is considered as an extended and modified version of Algorithm 2, convergence of Algorithm 2 is guaranteed and detailed analysis is stated in [47]. Accessing to sufficient samples, the running time of the proposed approach is $O(s \cdot k \cdot q)$, where $s$, $k$ are respectively the number of outer and inner iterations, and $q$ is gradient steps of inner loop. For a $N$-way-$K$-shot learning, the best accuracy is achieved when $||\nabla \theta|| \leq O(\sigma / \sqrt{NK})$, where $\theta = \mathbb{E}_{T \sim \mathcal{P}(T)}(T, \theta, T')$ is the query loss of task $T$, $\sigma$ is a bound on the standard deviation of $\nabla L_t(\theta, \mu)$ from its mean $\nabla L_t(\theta, \mu)$, and $\sigma$ is a bound on the standard deviation of estimating $\nabla L_t(\theta, \mu)$ using a single data point.

5 Supervised Examples in Unfairness Prevention

In the previous section, we derived a theoretically principled algorithm under the assumption that the convexity always holds for both $f_i(\cdot)$ and $g_i(\cdot)$. However, many problems of interest in machine learning and deep learning have a non-convex landscape due to the non-linearity of neural networks, where theoretical analysis is challenging. Nevertheless, algorithms originally developed for convex optimization problems like gradient descent and Adam [48] have shown promising results in practical non-convex settings. Taking inspiration from these successes, in this section, we respectively describe practical instantiations of our unfairness prevention for both regression and classification problems, and empirically evaluate the performance in Section 7.

Intuitively, an attribute affects the target variable if one depends on the other. Strong dependency indicates strong effects. Currently, most fairness criteria used for evaluating and designing machine learning models focus on the relationships between the protected attribute and the system output. For simplicity, we consider one binary protected attribute (e.g. white and black) in this work. However, our ideas can be easily extended to many protected attributes with multiple levels. We thus modify the introduced setting by letting $\mathcal{E} = \mathcal{E} \times \mathcal{Y}$, where $\mathcal{E} \subseteq \mathbb{R}^n$ is an input space, $\mathcal{S} = \{0, 1\}$ is a protected space, and $\mathcal{Y} = \mathbb{R}$ is an output space for regression and $\mathcal{Y} = \{0, 1\}$ for binary classification. For each task $t \in \{1, 2, ..., T\}$, we let $\{e_{t,i}, y_{t,i}, s_{t,i}\}_{i=1}^m \in (\mathcal{E} \times \mathcal{Y} \times \mathcal{S})$ be the corresponding task data and $m$ is the number of data points in the support set.

5.1 Regression Example

Since the protected attribute is binary valued while target variable is continuous, it is not possible to use typical same-type measures of dependency like correlation coefficient and point-wise mutual information to quantify the statistical dependency between them [30]. To quantify the effect of protected attribute $s$ on its predicted target $\hat{y}$ for $\mathcal{D}$, we apply Mean Difference (MD) for evaluating biases in regression problems, where such constraint is wildly used in studies [30], [44], [49] of fairness learning.

**Definition 1 (Mean Difference [30]).** The mean difference (MD) of numeric predicted target $\hat{y}$ in data set $\mathcal{D}$ partitioned by a binary protected variable $s$ is given by:

$$MD(\hat{y}, s) = |\mathbb{E}[\hat{y}|s = 1] - \mathbb{E}[\hat{y}|s = 0]|$$

where $\mathbb{E}$ indicates the sample mean and the mean difference is a positive number with a value of zero indicating no dependency of target on the protected variable. We say that $\hat{Y}$ achieves $c$-MD fairness if $MD(\hat{y}, s) \leq c$, where $c$ is a small positive fairness relaxation.

To formalize in the context of meta-learning in the form of the inner problem of Eq.(4), a mean-squared error is used to describe the adapted loss over a support set for each task. Let the task-specific constraint $g_i(\theta)$ satisfies $c$-MD fairness. The regression problem of a single task takes the form:

$$\min_{\theta_i \in \Theta} \sum_{(e', y') \sim T_t} ||\hat{y}(e', \theta_i) - y'||^2$$

subject to $|\mathbb{E}[\hat{y}(e', \theta_i)|s = 1] - \mathbb{E}[\hat{y}(e', \theta_i)|s = 0]| \leq c \leq 0$

where $(e', y')$ are an input/output pair sampled from task $T_t$ and $\hat{y}$ is a predicted outcome.

5.2 Classification Example

In a $N$-way-$K$-shot classification problem, since we assume all the tasks to be binary labeled, in this example, all of our tasks are 2-way (i.e. $N = 2$). In this classification
example of referencing K-fairness, we mean that we are using K training examples irrespective of class label, with the assumption that all tasks are 2-way. A fine-grained measurement to ensure fairness in class label prediction is to design fair classifiers by controlling the decision boundary covariance (DBC) [2].

Definition 2 (Decision Boundary Covariance [2]). The covariance between the protected variables s = \{s_i\}_{i=1}^K and the signed distance from the feature vectors to the decision boundary, \(d_\theta(e) = \{d_\theta(e_i)\}_{i=1}^h\),

\[
DBC(s, d_\theta(e)) = E[(s - \bar{s})d_\theta(e)] - E[s - \bar{s}]\bar{d}_\theta(e) 
\approx \frac{1}{h} \sum_{i=1}^{h} (s_i - \bar{s})d_\theta(e)(10)
\]

where \(E[s - \bar{s}]d_\theta(e)\) is cancels out since \(E[s - \bar{s}] = 0\) and \(h = N \times K\) is the sample size of a support set of a single task. In a linear model for classification, such as logistic regression, the decision boundary is simply the hyperplane defined by \(\theta^T e = 0\). A point \(\theta_i\) in the domain of a task is feasible if it satisfies the constraint \(DBC(s, d_\theta(e_i)) - c \leq 0\). Similar to the regression example, \(c\) is a user-defined fairness relaxation, and for discrete classification tasks with a cross-entropy loss, the problem of a single task takes the form:

\[
\min_{\theta_i \in \Theta} \sum_{(e_i, y_i) \sim T_i} y_i \log \tilde{y}(e_i, \theta_i) + (1 - y_i) \log (1 - \tilde{y}(e_i, \theta_i))
\]

subject to \(\frac{1}{N \times K} \sum_{s_i \sim S, \epsilon_i \sim T_i} (s_i - \bar{s})d_\theta(e_i) - c \leq 0(11)\)

The goal of a single task optimization is to approximate a good parameter pair \((\theta_i', \mu_i')\) by applying the proposed dual subgradient method and further pass the pair to evaluate accuracy and fairness (i.e., DBC) over the query data. As the original meta-learning problem in Eq. (2) is decomposed into a batch of single tasks, meta-parameters \((\theta, \mu)\) are iteratively updated using the proposed dual decomposition approach outlined in Algorithm [2]

### 6 Evaluation Settings

To evaluate the effectiveness of the proposed Duality-MAML, we conduct experiments of few-shot learning in the domain of supervised tasks. Formulation of the regression and classification problems are introduced in Section [3]. In this section, for both supervised settings, the datasets, evaluation metrics, the comparison methods, and tuning of hyperparameters are introduced in turn.

#### 6.1 Datasets

##### 6.1.1 Regression Datasets

The Chicago Crime dataset [34] includes information relevant to crime (e.g., household, unemployment) as well as demographic information (such as race and gender) in different communities across the Chicago city in 2015. These information were separately collected from American FactFinder (AFF) which is an online and self-service database provided by the U.S. Census Bureau and then aggregated various sources to the final data prepared for experiments. The Chicago Crime dataset is divided into 801 sub-tasks according to different communities in the Chicago city. All tasks were further split into 501 for training, 100 for validation and 200 for testing. Each task contains 52 crime records. Since the feature in the original data that described the percentage of African American population is numeric, in this experiment, we convert it into binary values based on the majority (> 70%) population of Black and non-Black. Thus, each record represents a weekly information including 13 numeric explanatory variables and one binary protected variable. In this data set, crime count is the target attribute that we need to detect discrimination from.

The Law School Admissions Council (LSAC) dataset [4] is a survey among law students taking the bar exam and attending law schools in the U.S. in 1991, which is widely applied in cutting-edge fair regression techniques [33], [49], [50]. The goal is to predict the score of each student based on features such as LSAT score and undergraduate GPA. The protected attribute is race. This dataset consists of 20,798 instances with 14 features and each group is considered as a task.

The COMPAS recidivism dataset [51] consists of a collection of data from over 10,000 criminal defendants from Broward County, Florida during 2013-2014. It includes attributes such as the sex, age, race, and priors for the

---

1. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
2. http://www2.law.ucla.edu/sander/Systemic/Data.htm
defendants. We create a binary sensitive column for whether the defendant is African American. We predict the ProPublica collected label of whether the defendant was rearrested within two years. The COMPAS dataset contains 60,843 records and is split to 50 tasks according to defendants’ dates of birth. Since the original dataset has a small mean rating difference between the protected groups, for our experiments, we randomly select instances of African Americans and increase the target values of 25%.

6.1.2 Classification Datasets
The Adult income dataset [52] contains a total of 34 tasks according to different countries and regions, totally 48,842 instances with 14 features (e.g., age, educational level) and a binary label, which indicates whether a subject’s incomes is above or below 50K dollars. We consider gender, i.e., male and female, as the protected attribute.

The Communities and Crime dataset [53] includes information relevant to crime (e.g., police per population, income) as well as demographic information (such as race and sex) in different communities across the U.S. We convert this dataset to a few-shot fairness setting by using each state as a different task. Following the same setting in [37], since the violent crime rate is a continuous value, we convert it into a binary label based on whether the community is in the top 50% violent crime rate within a state. Additionally, we add a binary sensitive column that receives a protected label if African-Americans are the highest or second highest population in a community in terms of percentage racial makeup.

The Bank Marketing dataset [54] contains a total 41,188 subjects, each with 20 attributes (e.g. loan, housing, etc.) and a binary label, which indicates whether the client has subscribed or not to a term deposit. In this case, we consider the marital status as the binary protected attribute, which is discretized to indicate whether the client is married or not. Since the dataset contains information of different months (i.e. January to December) and dates (i.e. Monday to Friday), we combine them as task labels and thus the Bank Marketing data contains 50 tasks.

6.2 Evaluation Metrics
To evaluate the proposed techniques for fairness learning, we introduced classic evaluation metrics, two for each supervised setting, to measure data biases. These measurements came into play that allows quantifying the extent of bias taking into account the protected attribute and were designed for indicating indirect discrimination.

6.2.1 Metrics for Regression
The area under the ROC curve (AUC) [30] point-wisely quantifies the statistical dependency between the protected variable and predictions.

\[
AUC = \frac{\sum_{(s_i, y_i) \in \mathcal{D}_+} \sum_{(s_j, y_j) \in \mathcal{D}_-} I(y_i > y_j)}{|\mathcal{D}_+| \times |\mathcal{D}_-|}
\]  

(12)

where \(I(\cdot)\) is an indicator function which returns 1 if its argument is true, 0 otherwise. \(AUC = 0.5\) represents random predictability, thus \(S\) is independent on \(Y\).

Impact Ratio (IR) [55] is defined as the ratio of mean over the protected and unprotected group in data \(\mathcal{D}\).

\[
IR = \frac{\sum_{y_i \in \mathcal{D}_+} y_i}{|\mathcal{D}_+|} / \frac{\sum_{y_j \in \mathcal{D}_-} y_j}{|\mathcal{D}_-|}
\]

(13)

The decisions are deemed to be discriminatory if the ratio of positive outcomes for the protected attribute is below 80% [56]. \(IR = 1\) indicates that there is no bias of data \(\mathcal{D}\).

6.2.2 Metrics for Classification
Discrimination [15] measures the bias with respect to the protected attribute \(S\) in the classification:

\[
Disc = \left| \frac{\sum_{i: s_i = 1} y_i - \sum_{i: s_i = 0} y_i}{\sum_{i: s_i = 1} 1 - \sum_{i: s_i = 0} 1} \right|
\]

This is a form of statistical parity that is applied to the binary classification decisions. It measures the difference in the proportion of positive classifications of individuals in the protected and unprotected groups. \(Disc = 0\) indicates there is no discrimination.

Consistency [15] compares a model’s classification prediction of a given data item to its \(k\)-nearest neighbors:

\[
Cons = 1 - \frac{1}{|D|v} \sum_{i=1}^{|D|} \left| \hat{y}_i - \sum_{j \in NN(\mathbf{x}_i)} \hat{y}_j \right|
\]

where \(|D|\) is the sample size, \(v\) is the number of nearest neighbors, and a nearest neighbor is defined based on a similarity measure (i.e. euclidean distance) of unprotected attributes \(e\). As demonstrated in [15], we applied the kNN function to the full set of examples to obtain the most accurate estimate of each point’s nearest neighbors. The consistency is a real number with a value of one signifying a fair prediction.

6.3 Competing Methods
We evaluate all datasets – the proposed approach against various baselines – by comparing the results of generalization on both loss/accuracy and fairness applied to:

1) MAML: The model-agnostic meta-learning model with no fairness constraints proposed by Finn et al., [1].
2) Masked MAML: Similar to MAML, this approach is applied to modified datasets by simply removing the protected attributes.
3) pretrain: In computer vision, models pre-trained on large-scale image classification have been shown to learn effective features [57]. In this paper, the pretrain baseline trains a single network on all tasks and in each task an unified fairness constraint is added to ensure the fairness condition is satisfied.
4) UP-MAML (regression only): Unfairness Prevention MAML [38] efficiently mitigates the dependency effect between the binary protected attribute and the numeric predictions by reducing the mean difference of sub-groups. Statistical parity in each task is balanced by setting a fixed trade-off parameter.
5) fair-MAML (classification only): [59] controls unfairness for each task and tunes a shared Lagrangian
multiplier across tasks by simply applying grid search.

6) **F-MAML$_{dp}$ (classification only)**: is a fair meta-learning approach proposed in [37]. In this baseline, Slack et al., proposed a simple regularization term aimed at achieving demographic parity for each task. All tasks share an unified regularization term in which the fairness hyperparameter is tuned through grid search, where the demographic parity regularizer $R_{dp} = 1 - p(y = 1|s = 0)$.

7) **F-MAML$_{eop}$ (classification only)**: is another fair meta-learning approach proposed in [37], in which the demographic parity regularizer is replaced with the one aimed at improving equal opportunity, where $R_{eop} = 1 - p(y = 1|s = 0, y = 1)$.

8) **LAFTR [58]**: is a transferring fair machine learning approach across domains that uses an adversarial approach to create an encoder that can be used to generate fair representations of datasets and demonstrate the utility of the encoder for fair transfer learning.

6.4 **Experiment Setup and Parameter Tuning**

Our neural network trained follows the same architecture used by [1], which contains 2 hidden layers of size of 40 with ReLU activation functions. When training, we use only one step gradient update (i.e. $q = 1$) and $k = 10$ inner primal-dual updates with $2NK$ samples of query set ($N = 1$ for regression setting and $N = 2$ for classification), and a fixed primal and dual learning rate of $\gamma = 0.01$ and $\alpha = 0$. We use Adam as the meta-optimizer. Similarly, we set meta-learning rates of $\eta = 0.001$ and $\beta = 0.01$ used to update the meta-loss in the outer loop. For all datasets, all the unprotected attributes are standardized to zero mean and unit variance and prepared for experiments. Besides, taking few-shot learning into account, we set a meta batch-size of 8 tasks and 4000 meta-iterations for all datasets. Some key characteristics for all real data are listed in Table 2.

All baseline models used to compare with our proposed approach share the same neural network architecture and parameter settings. Hyperparameters are selected by a held-out validation procedure. All experiments are repeated 10 times with the same settings. Results shown with these
methods in this paper are mean of experimental outputs.

7 Experimental Results

This section evaluates the effectiveness of the proposed approach and its competitors on a regression and classification task, respectively. We first focus on generalization of statistical parity on unseen tasks evaluated using fairness metrics. Moreover, taking the classification setting as an example, ablation studies reveal the trade-off relationship between validation loss and fairness that the proposed dual subgradient method alleviates when used to train classifiers, as well as demonstrate the contributions of the proposed model components. For all baseline methods, wherever applicable, hyper-parameters were tuned via grid search. Specifically, we chose the models that were Pareto-optimal with regard to MD or DBC and all other evaluation metrics.

7.1 Performance on Regression

Consolidated and detailed performance of the different techniques over real-world data are listed in Table 1 of the supplemental materials. We evaluate performance for regression examples by fine-tuning the model learned by all methods on \(K\)-shot of \(\{5, 10, 15, 20\}\) datapoints of each task for each dataset. Best performance of the tables in each experimental unit are labeled in bold. We first observe that there is a considerable amount of unfairness in the original datasets, which are reflected in the results of Data in the table. Experiment results shown in Table 1 demonstrate our proposed approach outperforms than other baseline methods in terms of controlling biases. It efficiently reduces MD limited to close zero signify a fair prediction. In addition, fairness results based on evaluation metrics are visualized in Figure 2. Each trail was repeated 10 times and results shown in the figure are mean of experimental outputs followed by error bars representing one standard deviation of uncertainty.

\(\text{MAML}\) became a famous meta-learning algorithm because of its fast adaptation and good generalization performance on losses \([1]\). However, our results shows it fails to control biases nor performs success in fairness generalization in a few-shot meta-learning, although \(\text{MAML}\) is stably able to produce high generalization accuracy. \(\text{Masked MAML}\) shows an improvement in fairness; however, there is still substantial
unfairness hidden in the data in the form of correlated attributes. In contrast to MAML, a simple pre-trained model is to find a meta-parameter $\theta$ such that the summation of losses over all tasks is minimum. In other words, it lacks gradient steps within each task. In our experiments, each task in the pretrain model undergoes a fairness constraint which is uniformly regulated and optimized by grid search. Although the pretrain method controls unfairness in a little bit, since the fairness controller is shared over all tasks and it is not updated either within a task nor in the outer loop, there is a lot of room for improvement in the performance of using the pretrain technique. To this end, UP-MAML [38] is designed for regression to boost generalization accuracy in few-shot fairness learning. Our results, which is consistent with [38], demonstrate that by simply setting a fixed parameter to control fairness for all meta-learning tasks is able to mitigate fairness onto new tasks, but the key to improve capability of fairness generalization is to separately learn such an optimal parameter for each task. Furthermore, though LAFT performs a way to transfer machine learning models between tasks, consistent with [37], we observe it is unsuccessful in very data light situations.

The proposed method, PDFM, circumvents the shortcomings of the competing methods mentioned above and outperforms such state-of-the-art in bias controlling with better results. Its effectiveness on unfairness prevention is performed and displayed by applying evaluated metrics, such as AUC and IR in regression. Specifically, AUC, which relates to rank tests, point-wisely quantifies the statistical dependency between the protected variable and predictions and hence is used for measuring discrimination between two groups. The more its value is close to 0.5, the less dependency of the outcome on the protected attributes and therefore the fairer the model we train. Results shown in Figure 2 (d)-(f) are evaluated using AUC metric, which reflects the performance of fairness generalization on unseen tasks based on regression datasets. In contrast to baseline methods, PDFM efficiently mitigates AUC approaching to a fair level. Moreover, Impact Ratio (IR), also known as slift, is used for quantifying discrimination, where the decisions are deemed to be discriminatory if the ratio of positive outcomes for the protected group is below 80%. Results in Figure 2 (g)-(i) demonstrate that IR performed via our method is not only above 80% level, but also higher than those using other techniques.

7.2 Performance on Classification

Similar to the Regression experiments, detailed performance of the different methods over datasets are listed in Table 2 of the supplemental materials. Specifically in classification experiments, different from the regression setting, K-shot tasks use K input/output pairs from each class, for a total of $N \times K$ datapoints for N-way classification.

Besides, another two state-of-the-art techniques designed for classification, F-MAML$_{dp}$ and F-MAML$_{eop}$, proposed by Slack et al., in [37] intuitively control unfairness by taking advantage of demographic parity and equal opportunity, respectively. Our results in Figure 3 demonstrate that these two baseline methods fail to show fairness generalization onto unseen tasks in contrast to the proposed approach, in terms of reducing $Disc$ and promoting Cons.

Our results indicate, for both supervised settings, by training a batch of tasks where each contains a handful biased data, our method is able to generalize not only accuracy but also, more importantly, fairness onto unseen data, and hence mitigates bias more effectively by alleviating the dependency effect of the protected variable on predictions. Besides, when $K = 5$, results of both validation loss and fairness are repeated with larger standard deviations; but such instability tends to level off as the number of samples considered in each task increases. Furthermore, it is notable that the more instances of a task used for training, the better fairness generalization onto new tasks.

7.3 Loss–Fairness Tradeoffs

Although our proposed approach returns a bit smaller predictive accuracies (see Table 2 in the supplemental material), this is due to the trade-off between losses and fairness. To this end, we train each method and sweep over a range of seven dual initial values: [0.001, 0.01, 0.1, 1, 10, 100, 1000]. Taking 10-shot as an example, results presented in Figure 4 is the mean across 10 runs on each set of dual variable using randomly selected hold out validation tasks. The fairness, i.e. $DBC$, presented is the ratio between the protected and unprotected groups. Smaller validation loss and fairness values closer to zero (i.e. bottom left in each sub-figure) indicate more successful outcomes. Here, as MAML does not have hyper-parameters to control the loss/fairness trade-off, its outcomes across three datasets are presented with very low validation losses but high fairness values.

In the proposed problem setting, the pretrain neural network shows some ability to learn the new task using little data and
TABLE 3: Exploration of number of inner primal-dual update (denoted $k$) for 10-shot fair classification using PDFM.

| $k$ | Adult Data | Communities and Crime | Bank Marketing |
|-----|------------|----------------------|---------------|
|     | Acc        | DBC                  | Disc          | Cons          |
|     |            |                      |               |               |
| 1   | 78.5%      | 0.011                | 0.121         | 0.947         |
| 2   | 77.2%      | 0.017                | 0.121         | 0.944         |
| 5   | 79.8%      | 0.020                | 0.114         | 0.947         |
| 10  | 83.8%      | 0.011                | 0.123         | 0.943         |
| 20  | 78.5%      | 0.016                | 0.121         | 0.947         |

TABLE 4: Exploration of number of inner steps (denoted $q$) for 10-shot fair classification using PDFM.

| $q$ | Adult Data | Communities and Crime | Bank Marketing |
|-----|------------|----------------------|---------------|
|     | Acc        | DBC                  | Disc          | Cons          |
|     |            |                      |               |               |
| 1   | 83.8%      | 0.011                | 0.123         | 0.943         |
| 5   | 91.2%      | 0.017                | 0.149         | 0.938         |
| 10  | 86.0%      | 0.014                | 0.135         | 0.927         |
| 20  | 92.5%      | 0.019                | 0.170         | 0.903         |

8 CONCLUSION AND FUTURE WORK

Techniques in meta-learning have been shown effectiveness for adaption of deep learning models on accuracy generalization to new tasks. These methods, however, are unable to ensure fairness adaption. In this paper, for the first time a novel fair meta-learning framework is proposed, in which a good pair of primal-dual meta-parameters is optimally learned. To be specific, the meta-parameter pair is trained over a variety of learning tasks with a small amount of training samples. To produce the best performance, we implement two optimization strategies for both inner and meta-subgradient update. Theoretical analysis justifies the efficiency and effectiveness of the proposed algorithms to support existence of solutions and algorithmic convergence guarantee. Results from extensive experiments demonstrate substantial improvements over the best prior work and our proposed framework is capable of generalization both accuracy and fairness onto new tasks. Further research in this area can make multitask parameters a standard ingredient in explainable fairness transfer learning.

ACKNOWLEDGEMENT

This work is supported by NSF awards IIS-1815696, IIS-1750911, DMS-1737978, DGE-2039542, and MRI-1828467.

REFERENCES

[1] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” IML, 2017.
[2] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, “Fairness constraints: Mechanisms for fair classification,” AISTATS, 2017.
[3] G. Denevi, D. Stamos, C. Ciliberto, and M. Pontili, “Online-within-online meta-learning,” NeurIPS, 2019.
[4] J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn, “Bayesian model-agnostic meta-learning,” in NeurIPS, 2018, pp. 7332–7342.
[5] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” NeurIPS, 2016.
[6] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” NeurIPS, 2017.
[7] Z. Xu, H. P. van Hasselt, and D. Silver, “Meta-gradient reinforcement learning,” in NeurIPS, 2018, pp. 2396–2407.
[8] S. Ravi and H. Larochelle, “Optimization as a model for few-shot learning,” ICLR, 2017.
[9] C. Finn, K. Xu, and S. Levine, “Probabilistic model-agnostic meta-learning,” in NeurIPS, 2018, pp. 9516–9527.
[10] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, “Online meta-learning.” ICML, 2019.

[11] A. Nichol and J. Schulman, “Reptile: a scalable metalearning algorithm,” arXiv preprint arXiv:1803.02999, 2018.

[12] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, “Meta-learning with latent embedding optimization,” ICLR, 2019.

[13] I. Zliobaite, “A survey on measuring indirect discrimination in machine learning,” arXiv preprint arXiv:1511.00148, 2015.

[14] The U.S. Civil Rights Act, July 2, 1964.

[15] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” ICML, 2013.

[16] D. P. Palomar and M. Chiang, “Alternative distributed algorithms for network utility maximization: framework and applications.” IEEE Transactions on Automatic Control, vol. 52(12), pp. 2254–2269, 2007.

[17] R. L. Rafford, C. J. Tomlin, , and S. P. Boyd, “Distributed optimization for cooperative agents: Application to formation flight.” In Proceedings of the IEEE Conference on Decision and Control, pp. 7543-7548, 2003.

[18] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pa, “A meta-transfer objective for learning to disentangle causal mechanisms,” ICLR, 2020.

[19] H. Yao, Y. Wei, J. Huang, and Z. Li, “Hierarchically structured meta-learning,” ICML, 2019.

[20] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, “Cross-domain few-shot classification via learned feature-wise transformation,” ICLR, 2020.

[21] D. Lian, Y. Zheng, Y. Xu, Y. Lu, L. Lin, P. Zhao, J. Huang, and S. Gao, “Towards fast adaptation of neural architectures with meta-learning,” ICLR, 2020.

[22] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in CVPR, 2018, pp. 1199–1208.

[23] Z. Wang, Z. Kong, S. Changra, H. Tao, and L. Khan, “Robust high dimensional stream classification with novel class detection,” in ICDE. IEEE, 2019, pp. 1418–1429.

[24] B. Oreshkin, P. R. López, and A. Lacoste, “Tadam: Task dependent adaptive metric for improved few-shot learning,” in NeurIPS, 2018, pp. 721–731.

[25] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in ICML, 2016, pp. 1842–1850.

[26] Z. Wang, Y. Wang, Y. Lin, E. Delord, and K. Latifur, “Few-shot and adversarial representation learning for continual stream mining,” in Proceedings of The Web Conference 2020, 2020, pp. 718–728.

[27] A. Antoniou, H. Edwards, and A. Storkey, “How to train your model,” ICLR, 2019.

[28] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, “Meta-transfer learning for few-shot learning,” in CVPR, 2019, pp. 403-412.

[29] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, “Bilevel programming for hyperparameter optimization and meta-learning,” ICML, 2018.

[30] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, “Controlling attribute effect in linear regression,” in ICDM. ACM, 2015.

[31] M. Feldman, S. Friedler, J. Meller, C. Scheidegger, and S. Venkatasubramanian, “Certiﬁying and removing disparate impact.” KDD, 2015.

[32] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning.” NeurIPS, 2016.

[33] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, “A convex framework for fair regression.” FATML, 2018.

[34] C. Zhao and F. Chen, “Rank-based multi-task learning for fair regression,” IEEE International Conference on Data Mining (ICDM), 2019.

[35] D. Gondek and T. Hofman, “Non-redundant clustering with conditional ensembles.” KDD, 2005.

[36] A. Singh and T. Joachims, “Fairness of exposure in rankings,” in KDD, 2018, pp. 2219–2228.

[37] D. Slack, S. Friedler, and E. Givental, “Fairness warnings and fair-maml: Learning fairly with minimal data,” Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT), 2020.

APPENDIX A

Proof of Proposition 1

Proof: For simplicity, the subscript t is hidden. To prove the convergence, we first show that \( \hat{θ}^k \) is a Cauchy sequence, i.e. \( \forall ε > 0, \) there is a \( K \in \mathbb{N} \) such that \( ||\hat{θ}^k - \hat{θ}^j|| < ε, \forall k, j \geq K. \) Given Eq. (5), we can derive \( \hat{θ}^k = \frac{\hat{θ}^k + k \hat{θ}^k}{k + 1} \). And hence \( \hat{θ}^k - \hat{θ}^j = \frac{\hat{θ}^k - \hat{θ}^j}{k + 1} \). Since \( Θ \) is a compact convex set and we assume \( k' > k \), we have \( \hat{θ}^j, \hat{θ}^k \in Θ \) and \( ||\hat{θ}^k||, ||\hat{θ}^j|| \leq M \), where \( M \geq 0 \). Iteratively, we have

\[
||\hat{θ}^{k'} - \hat{θ}^j|| = ||\hat{θ}^{k' - 1} - \hat{θ}^{k' - 2} + \ldots + \hat{θ}^k - \hat{θ}^j|| = ||\hat{θ}^{k' - 1} - \hat{θ}^{k' - 2} + \ldots + \hat{θ}^k - \hat{θ}^j|| \\
\leq ||\hat{θ}^{k' - 1}|| + ||\hat{θ}^{k' - 2}|| + \ldots + ||\hat{θ}^{k - 1}|| + ||\hat{θ}^j|| \\
\leq 2M \left( \frac{k' - k}{k + 1} \right) \]

Therefore, for any arbitrary \( ε > 0 \), we let \( \frac{2M(k' - k)}{k + 1} < ε \) and we have \( ||\hat{θ}^{k'} - \hat{θ}^j|| < ε, \forall k', j \geq K. \) Thus, \( \{\hat{θ}^k\} \) is a Cauchy sequence. Furthermore, since a Cauchy sequence
is bounded, there is a subsequence $b_k$ converging to the limit $L$ of it. For any $\epsilon > 0$, there exists $n, m \geq K$ satisfying $|\tilde{\theta}^n - \tilde{\theta}^m| < \frac{\epsilon}{2}$. Thus, there is a $b_k = \tilde{\theta}^{m_k}$, such that $m_k \geq K$ and $|b_{m_k} - L| < \frac{\epsilon}{2}$.

$$
|\tilde{\theta}^n - L| = |\tilde{\theta}^n - b_k + b_k - L| \\
\leq |\tilde{\theta}^n - b_k| + |b_k - L|
$$

$$
< |\tilde{\theta}^n - L| + \frac{\epsilon}{2} < \epsilon
$$

Since $\epsilon$ is arbitrarily small, we proof that the sequence $\{\tilde{\theta}^k\}$ converges to its limit $L = \tilde{\theta}^*$ asymptotically.

**APPENDIX B**

**Proof of Proposition 2**

*Proof: Similar to the proof of Proposition 1, for simplicity, the subscript $i$ is hidden.*

(1) According to Eq. (3), we have $\mu^{k+1} = [\mu^k + \alpha g(\theta^i)]^{\top} \geq \mu^k + \alpha T g_i$. Since the constraint function $g(\theta)$ is convex, we have $g(\tilde{\theta}^k) \leq \frac{1}{k} \sum_{i=0}^{k-1} g(\theta_i) = \frac{1}{k} \sum_{i=0}^{k-1} \alpha g(\theta_i) \leq \frac{1}{k} \alpha (\mu^k - \mu^0) \leq \frac{\mu^0}{k}$. Thus we have $|g(\tilde{\theta}^k) + ||| \mu^0 ||| \leq \frac{1}{k} \alpha \mu^0 \leq \frac{1}{k} \alpha \mu^0$. Since all the $\mu_i$ is nonnegative, we have $g(\tilde{\theta}^k) \leq \frac{\mu^0}{k}$. Thus we have $|f(\tilde{\theta}^k) + ||| \mu^0 ||| \leq \frac{1}{k} \alpha \mu^0 \leq \frac{1}{k} \alpha \mu^0$. Under the assumption of convexity and Slater condition, $q^* = f^*$. Together with the condition that $f(\theta)$ is convex and $\tilde{\theta} \in \Theta$, we have

$$
f(\tilde{\theta}^k) \leq \frac{1}{k} \sum_{i=0}^{k-1} f(\theta^i)
$$

$$
= \frac{1}{k} \sum_{i=0}^{k-1} (f(\theta^i) + (\mu^i)\alpha T g_i) - (\mu^i)\alpha T g(\theta^i)
$$

$$
= \frac{1}{k} \sum_{i=0}^{k-1} (f(\theta^i) + (\mu^i)\alpha T g_i) - \frac{1}{k} \sum_{i=0}^{k-1} (\mu^i)\alpha T g(\theta^i)
$$

$$
= \frac{1}{k} \sum_{i=0}^{k-1} q(\mu^i) - \frac{1}{k} \sum_{i=0}^{k-1} (\mu^i)\alpha T g(\theta^i)
$$

$$
\leq q^* - \frac{1}{k} \sum_{i=0}^{k-1} (\mu^i)\alpha T g(\theta^i)
$$

From Eq. (3), we have

$$
||| \mu^{i+1} |||^2 = ||| \mu^i + \alpha T g_i |||^2 \\
\leq ||| \mu^i + \alpha T g(\theta^i) |||^2 \\
\leq ||| \mu^i |||^2 + 2\alpha (\mu^i)^\top g(\theta^i) + \alpha^2 ||| g(\theta^i) |||^2
$$

$$
- (\mu^i)^\top g(\theta^i) \leq ||| \mu^i |||^2 - ||| \mu^i |||^2 + \alpha^2 ||| g(\theta^i) |||^2
$$

By plugging in $- (\mu^i)^\top g(\theta^i)$, we thus obtain an upper bound for $f(\tilde{\theta}^k)$ that

$$
f(\tilde{\theta}^k) \leq q^* + \frac{1}{k} \sum_{i=0}^{k-1} ||| \mu_i |||^2 - ||| \mu^i |||^2 + \alpha^2 ||| g(\theta^i) |||^2
$$

$$
= q^* + \frac{1}{k} \sum_{i=0}^{k-1} ||| \mu_i |||^2 - ||| \mu^i |||^2 + \frac{1}{k} \sum_{i=0}^{k-1} \alpha^2 ||| g(\theta^i) |||^2
$$

$$
= q^* + \frac{||| \mu_i |||^2 - ||| \mu^i |||^2 + \alpha \sum_{i=0}^{k-1} ||| g(\theta^i) |||^2}{2\alpha}
$$

$$
\leq f^* + \frac{||| \mu^i |||^2}{2\alpha} + \frac{\alpha L^2}{2}
$$

(3) By definition, for any $\theta \in \Theta$, we have

$$
f(\theta) + (\mu^*)^\top g(\theta) \geq f(\theta^*) + (\mu^*)^\top g(\theta^*) = q(\mu^*)
$$

As $\tilde{\theta} \in \Theta$ and for all $k \geq 1$, we obtain an lower bound for $f(\tilde{\theta}^k)$ that

$$
f(\tilde{\theta}^k) = f(\tilde{\theta}^k) + (\mu^k)^\top g(\tilde{\theta}^k) - (\mu^k)^\top g(\tilde{\theta}^k)
$$

$$
\geq q(\mu^k) - (\mu^k)^\top g(\tilde{\theta}^k) + ||| \mu^k |||^2
$$

$$
\geq f^* - ||| \mu^* |||^2 ||| g(\tilde{\theta}^k) |||^2
$$

---

**Chen Zhao** is a Ph.D student in Computer Science at the University of Texas at Dallas. Before that, he received his Master's degree from the State University of New York at Albany. His work focuses on studying fairness in machine learning. He is interested in applying computational and statistical methods to reduce predictive bias and in building fairer and more interpretable algorithms in Multi-task learning (MTL), metric learning, and meta learning.

**Feng Chen** is an Associate Professor in the Department of Computer Science at the University of Texas at Dallas. He was an assistant professor at the University at Albany-SUNY from 2014 to July 2019 and a Postdoctoral Fellow at Carnegie Mellon University in 2013. Dr. Chen's research interests include large-scale data mining, network mining, and machine learning, focusing on event and pattern detection in massive, complex networks. He was awarded an NSF CAREER award in 2018 and a member of the IEEE and the ACM.

**Zhuoyi Wang** is currently a Ph.D student at the Big Data Analytics and Management Lab in the Computer Science department, at the University of Texas at Dallas. His research focuses on the area of stream mining, semi-supervised learning and cyber-security, especially the learning from limited labeled data in the non-stationary stream settings.

**Latifur Khan** is a full Professor in the Computer Science department at the University of Texas at Dallas. He received his Ph.D in Computer Science from the University of Southern California in 2000. He is an ACM Distinguished Scientist and received Fellow of SIRI (Society of Information Reuse and Integration) award in 2018. He received prestigious awards including the IEEE Technical Achievement Award for Intelligence and Security Informatics and IBM Faculty Award (research) in 2016. Dr. Khan's research focuses on big data management and analytics, data mining and its application on cyber security, complex data management including geo-spatial data and multimedia data. His research has been supported by grants from NSF, the Air Force Office of Scientific Research (AFOSR), DOE, NSA, IBM and HPE.