A Primal-Dual Subgradient Approach for Fair Meta Learning

Chen Zhao, Feng Chen, Zhuoyi Wang, Latifur Khan
Department of Computer Science
The University of Texas at Dallas
Richardson Texas, USA
{chen.zhao, feng.chen, zhuoyi.wang1, lkhan}@utdallas.edu

***Abstract*—The problem of learning to generalize on unseen classes during the training step, also 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 [2]. 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 novel 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 an example of bias controlling using decision boundary covariance (DBC) [3] as the fairness constraint for each task, and demonstrate the versatility of our proposed approach by applying it to classification on a variety of three real-world datasets. Our experiments show substantial improvements over the best prior work for this setting. Our code and datasets are available at https://github.com/charliezhaoyinpeng/PDFM.git.**

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

I. 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. To address this challenge, meta-learning (*a.k.a* learning to learn) leverages the transferable knowledge learned from previous tasks, then adapts on 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], 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 instantiates the transferable knowledge as latent representations, and gradient-based methods [1], [9]–[12] that aim to learn proper model initialization for all tasks, such that the summation query errors is 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.

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 [14]. 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 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

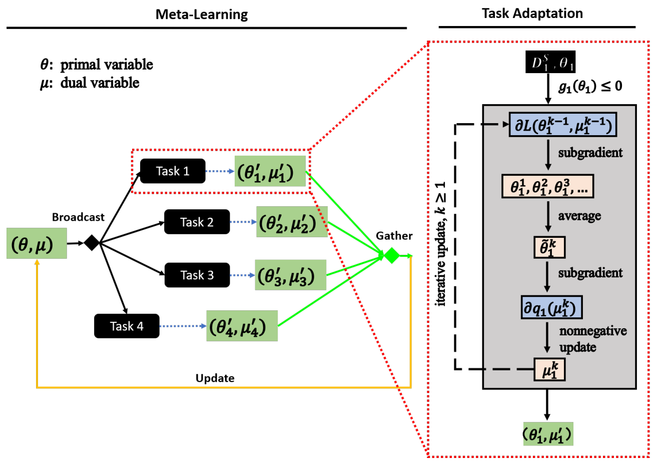


Fig. 1: Schematic of our proposed PDFM pipeline. (Left) The global meta-parameters (θ, μ) are sent to each task and each task optimizes in parallel to find a good task-specific primal-dual pair, *e.g.* (θ'_1, μ'_1), that is approximated by an averaging scheme dual subgradient algorithm presented on the 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. The inner loop dual subgradient algorithm ensures that the duality gap of each task is minimum.

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 [15], [16] 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 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 an example of decision boundary covariance (DBC) as the fairness constraint for justification, such constraint indicates the covariance between the protected variable and the signed distance from the feature vectors to the decision boundary [3]. We demonstrate the versatility of our proposed approach on a variety of three 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 Primal-Dual Fair Meta-learning framework, namely PDFM, 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 two 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 three 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.

II. Related Work

Meta-Learning based on few-shot studies that trained models to make it quickly adapt to new tasks, under a few labeled samples. Several recent approaches have made significant progress in meta-learning [17]–[20]. Previous algorithms majorly focus on the metric-based idea, which aim to learn an embedding space between query and support examples, where similar instances are closer and different ones are further apart [5], [6], [21], [22]. 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.

In addition, gradient descent based algorithms [1], [8], [9], [12], [23] aim to learn good model initialization so that the meta-loss is minimum. They tend 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, which could achieve good generalization over new tasks by encoding prior knowledge. Some existing work such as Franceschi et al. [24] also provide convergence guarantees for gradient-based meta-learning with strongly-convex functions. Despite methods in the area of meta-learning 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 adaption on 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 [25]. Based on the taxonomy by tasks, fairness learning can be typically categorized to classification [3], [26], [27], regression [25], [28], [29], clustering [30], and recommendation [31], [32] 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 [33]–[35]. These methods focus on studies of fairness generalization onto unseen tasks by adding an uniformed fairness regularizer to each task. In addition, Lagrange multipliers were consider as hyperparameters and they were manually tuned by grid search. However, such prior studies suffer from limitations that (1) the trade-off parameter is valued the same for each task, and (2) hence there is a big room for improvement on the generalization of both accuracy and fairness onto new tasks. In this paper, to overcome such limitations, we develop a novel fair meta-learning framework. Each task is underwent a task-specific soft fairness constraint. Besides, we consider Lagrange multipliers as dual variables and hence, instead of grid search, they are optimized to minimize the duality gap between the primal and dual functions.

III. Methodology

A. Problem Setting

Let $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ be the data space, where $\mathcal{X}\subset\mathbb{R}^n$ is the input space, $\mathcal{Y}=\{1,2,...,N\}$ means a sequence of discrete classes of 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 Θ will be a closed, convex, non-empty subset of an Euclidean space.

In this work, we consider a collection of supervised learning tasks $\mathcal{T}=\{(\mathcal{D}_t^S,\mathcal{D}_t^Q)\}_{t=1}^T$ which distributions over $\mathcal{Z}$ and T is denoted as the number of tasks. $\mathcal{T}$ 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 $\{\mathbf{x}_{t,i},y_{t,i}\}_{i=1}^m\in(\mathcal{X}\times\mathcal{Y})$ 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 [28], [36], [37]. For this purpose, in this paper, we assume that convexity of task constraints always holds.

B. 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_t:

$$\begin{aligned}\theta_t'=Alg(\mathcal{D}_t^S,\theta)=\arg\min_{\theta_t\in\Theta}\quad & f_t(\theta_t;\theta):=\mathcal{L}^{inner}(\mathcal{D}_t^S,\theta_t;\theta)\\ \text{subject to}\quad & g_t(\mathcal{D}_t^S,\theta_t)\leq 0 \qquad (1)\end{aligned}$$

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

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

The optimal value of the Eq.(1) is denoted as f_t^*, which is assume to be finite and is achieved at an optimal and feasible solution θ_t^*, *i.e.* $f_t^*=f_t(\theta_t^*)$. The goal of training a single task is to output local parameter θ_t given the meta-parameter θ such that it minimizes the task loss $f_t(\theta_t)$ subject to the task constraint $g_t(\theta_t)\leq 0$. Next, to update the meta-parameter, we minimize the generalization error $\mathcal{L}^{meta}$ 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

$$\begin{aligned}\min_{\theta\in\Theta}\quad & \mathcal{L}^{meta}=\sum_{t=1}^T f_t(\theta_t';\theta):=\sum_{t=1}^T\mathcal{L}^{inner}(\mathcal{D}_t^Q,Alg(\mathcal{D}_t^S,\theta))\\ \text{subject to}\quad & \sum_{t=1}^T g_t(\mathcal{D}_t^Q,Alg(\mathcal{D}_t^S,\theta))\leq 0 \qquad (2)\end{aligned}$$

where $\theta_t'=\arg\min_{\theta_t\in\Theta,g_t(\theta_t)\leq 0}f_t(\theta_t)$ 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 $g_t(\mathcal{D}_t^Q,Alg(\mathcal{D}_t^S,\theta))\leq 0$, 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.

C. 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 [33]–[35]. 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

$$L(\theta_t, \mu_t) = f_t(\theta_t) + \mu_t^T g_t(\theta_t)$$

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

$$q_t(\mu_t) = \inf_{\theta_t \in \Theta} L(\theta_t, \mu_t) = \inf_{\theta_t \in \Theta} \{f_t(\theta_t) + \mu_t^T g_t(\theta_t)\}$$

Since the dual function $q_t(\mu_t)$ is a pointwise affine function of μ_t, we thus can maximize the dual function to obtain a tightest lower bound of the optimal primal f_t^* and through out this paper, we assume f_t^* is finite. The goal is to obtain the dual optimal value q_t^* at μ_t^*, such that the duality gap, *i.e.* $f_t^* - q_t^*$, is as small as possible. Zero duality gap thus indicates that the optimal values of the primal and dual problems are equal, *i.e.* $f_t^* = q_t^*$. Due to space limit, the same idea is applied to solve Eq.(2). The Lagrangian function of the outer loop is hence parameterized by the meta-pair (θ, μ) and the goal is to find a good pair of initializations by optimizing a max-min problem.

D. 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 $\alpha \succ 0$ to update the dual solution iteratively:

$$\mu_t^k = [\mu_t^{k-1} + \alpha^T g_k]^+ \tag{3}$$

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 $\mu_t^0 \succ 0$ is an initial dual point. The subgradient iterate g_k is a subgradient of the dual function q_t at a given $\mu_t^k \succeq 0$:

$$g_k = g_t(\tilde{\theta}_t^k) \in \partial q_t(\mu_t^k) = \text{conv}(\{g_t(\tilde{\theta}_t^k) | \tilde{\theta}_t^k \in \Theta_{\mu_t^k}\}) \tag{4}$$

where $\Theta_{\mu_t^k} = \{\tilde{\theta}_t^k \in \Theta | q_t(\mu_t^k) = f_t(\tilde{\theta}_t^k) + (\mu_t^k)^T g_t(\tilde{\theta}_t^k)\}$ 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 $\{\theta_t^k\}$ to approximate primal optimal solutions. In particular, the sequence $\{\tilde{\theta}_t^k\}$ is defined as the averages of the previous vectors through θ_t^0 to θ_t^{k-1},

$$\tilde{\theta}_t^k = \frac{1}{k} \sum_{i=1}^{k-1} \theta_t^i, \quad \forall k \geq 1 \tag{5}$$

where the corresponding primal feasible iterate θ^k is given by any solution of the set.

$$\theta_t^k \in \arg\min_{\theta_t \in \Theta} \{f_t(\theta_t^{k-1}) + (\mu_t^{k-1})^T g_t(\theta_t^{k-1})\} \tag{6}$$

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

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: $\alpha \succ 0, \gamma \succ 0$: learning rate
Require: $q > 0$: a small number of subgradient update steps
$\mu_t^0 \leftarrow \mu$, $\theta_t^0 \leftarrow \theta$
Initialize an empty array $a = \emptyset$
for $k = 1, 2, ...$ do
    for $q = 1, 2, ...$ do
        Evaluate the primal feasible subgradient $\bar{\nabla} \in \nabla_{\theta_t^{k-1}}\{f_t(\theta_t^{k-1}) + (\mu_t^{k-1})^T g_t(\theta_t^{k-1})\}$
        $\theta_t^k \leftarrow \theta_t^{k-1} - \gamma^T \bar{\nabla}$
    end for
    Add $\theta_t^k$ in $a$
    Evaluate $\tilde{\theta}_t^k$ by taking the average of previous vectors in $a$: $\tilde{\theta}_t^k = \frac{1}{k}\sum_{i=0}^{k-1} \theta_t^i$
    Calculate the subgradient iterate $g_k = g_t(\tilde{\theta}_t^k)$
    Update the dual solution $\mu_t^k = [\mu_t^{k-1} + \alpha^T g_k]^+$
end for
return $(\theta_t', \mu_t')$, where $\theta_t' = \theta_t^k, \mu_t' = \mu_t^k$
```

Moreover, the dual subgradient schemes can be applied efficiently to approximate a solution to Eq.(1). Specifically, it returns a good pair of task-level primal and dual parameters (θ'_t, μ'_t). In the following section, due to the decomposable structure of the meta-learning framework for few-shot learning, meta-parameters (θ, μ) are updated by minimizing the summation of query losses across all training tasks.

E. Update Meta-parameters via Dual Decomposition

In this work, inspired by the concept of resource allocation from economics [15], [16], our model's goal is to estimate a good pair of primal-dual weight initialization (θ, μ), 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 [38]. This method is typically simple and efficient, which can be divided into two steps for each iterate, *i.e. broadcast* and *gather*. In the *broadcast* step, the meta-dual parameter μ is sent to each of tasks $\mathcal{T}_t$. Through Algorithm 1, local primal, and dual parameters θ_t and μ_t of a single task are iteratively optimized using few-shot support data. Query loss $f_t(\mathcal{D}_t^Q, \theta'_t)$ and fairness estimate $g_t(\mathcal{D}_t^Q, \theta'_t)$, 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,

$$\theta^{s+1} \in \arg\min_{\theta\in\Theta} \sum_{t=1}^{T} f_t(\theta'_t; \theta^s) + \mu^s \sum_{t=1}^{T} g_t(\theta'_t; \theta^s) \qquad (7)$$

$$\mu^{s+1} = [\mu^s + \beta \sum_{t=1}^{T} g_t(\theta'_t)]^+ \qquad (8)$$

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

IV. 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 $M = \{\mu_t | \mu_t \succeq 0, -\infty < q_t(\mu_t) < \infty\}$, and for every fixed $\mu_t \in M$, we have the solution set $\mathcal{C} \subset \Theta$ for $q_t(\mu_t)$.

Assumption 2. *(Slater Condition and Bounded Subgradients) The convex set Θ is compact (i.e. closed and bounded). There exists a Slater point $\bar{\theta}_t \in \Theta$, such that $g_j(\bar{\theta}_t) < 0, \forall j = 1, 2, ..., m$, 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 $L = \max_{\tilde{\theta}_t \in \Theta} ||g_t(\tilde{\theta}_t)||$.

Algorithm 2 The Primal-Dual Fair Meta-learning (PDFM) Algorithm

```
Require: $p(\mathcal{T})$: distribution over tasks
Require: $\eta \succ 0, \beta \succ 0$: learning rate
randomly initialize primal and dual meta-parameter, i.e. $\theta \in \Theta$ and $\mu \in \mathbb{R}_+^m$
while not done do
    sample batch of tasks $\mathcal{T}_t \sim p(\mathcal{T}), t = 1, 2, ..., T$
    for all $\mathcal{T}_t = \{\mathcal{D}_t^S, \mathcal{D}_t^Q\}$ do
        Sample datapoints $\mathcal{D}_t^S = \{\mathbf{x}_t, \mathbf{y}_t\}$ from $\mathcal{T}_t$
        Compute adapted primal-dual parameters $\theta'_t$ and $\mu'_t$ using $\mathcal{D}_t^S$ by applying Algorithm 1
        Sample datapoints $\mathcal{D}_t^Q = \{\mathbf{x}_t, \mathbf{y}_t\}$ from $\mathcal{T}_t$ for the meta-update, where $\mathcal{D}_t^S \cap \mathcal{D}_t^Q = \emptyset$
        Evaluate query loss $f_t(\theta'_t)$ and query constraint $g_t(\theta'_t)$ using $\mathcal{D}_t^Q$
    end for
    Update $\theta$ and $\mu$ using Eq.(7).    ▷ Update Meta-parameters.
end while
```

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

Due to space limit, Lemma 1 is easily proved using the *Weierstrass Theorem* proposed in [39]. Next, for the averaged primal sequence $\{\tilde{\theta}_t^k\}$, we show that it always converges when Θ is compact [40].

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

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

$$\begin{aligned} ||\tilde{\theta}^{k'} - \tilde{\theta}^k|| &= ||\tilde{\theta}^{k'} - \tilde{\theta}^{k'-1} + \cdots + \tilde{\theta}^{k+1} - \tilde{\theta}^k|| \\ &= ||\frac{\theta^{k'-1} - \tilde{\theta}^{k'-1}}{k'} + \cdots + \frac{\theta^k - \tilde{\theta}^k}{k+1}|| \\ &\leq \frac{||\theta^{k'-1}|| + ||\tilde{\theta}^{k'-1}||}{k'} + \cdots + \frac{||\theta^k|| + ||\tilde{\theta}^k||}{k+1} \\ &\leq \frac{2M(k'-k)}{k+1} \end{aligned}$$

Therefore, for any arbitrary $\epsilon > 0$, we let $\frac{2M(k'-k)}{k+1} < \epsilon$ and we have $||\tilde{\theta}^{k'} - \tilde{\theta}^k|| < \epsilon, \forall k', k \geq K$. Thus, $\{\tilde{\theta}^k\}$ is a Cauchy sequence. Furthermore, since a Cauchy sequence is bounded, there is a subsequence b_n 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 $||b_{m_k} - L|| < \frac{\epsilon}{2}$.

$$\begin{aligned} ||\tilde{\theta}^n - L|| &= ||\tilde{\theta}^n - b_k + b_k - L|| \\ &\leq ||\tilde{\theta}^n - b_k|| + ||b_k - L|| \\ &< ||\tilde{\theta}^n - \tilde{\theta}^m|| + \frac{\epsilon}{2} < \epsilon \end{aligned}$$

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

Besides, since the proposed Algorithm 2 is considered as an extended and modified version of [1], convergence of Algorithm 2 is guaranteed and detailed analysis is stated in [41]. 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 $||\nabla\theta|| \leq O(\tilde{\sigma}/\sqrt{NK})$, where $\theta = \mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})} l_{\mathcal{T}}(f_\theta)$, $l_{\mathcal{T}}$ is the query loss of task $\mathcal{T}$, σ is a bound on the standard deviation of $\nabla L_t(\theta_t, \mu_t)$ from its mean $\nabla L(\theta, \mu)$, and $\tilde{\sigma}$ is a bound on the standard deviation of estimating $\nabla L_t(\theta_t, \mu_t)$ using a single data point.

V. A Classification Example in Unfairness Prevention

In the previous section, we derived a theoretically principled algorithm under the assumption that the convexity always holds for both $f_t(\cdot)$ and $g_t(\cdot)$. 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 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 classification problems, and empirically evaluate the performance in Section VII.

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 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be the data space, where $\mathcal{X} = \mathcal{E} \cup \mathcal{S}$. Here $\mathcal{E} \subset \mathbb{R}^n$ is an input space, $\mathcal{S} = \{0, 1\}$ is a protected space, and $\mathcal{Y} = \{0, 1\}$ is an output space for binary classification. For each task $t \in \{1, 2, ..., T\}$, we let $\{\mathbf{e}_{t,i}, y_{t,i}, s_{t,i}\}_{i=1}^m \in (\mathcal{E} \times \mathcal{Y} \times \mathcal{S})$ be the corresponding task data and m is the number of datapoints in the support set. 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 referencing K-shot 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) [3].

Definition 1 (Decision Boundary Covariance [3])**.** *The covariance between the protected variables* $\mathbf{s} = \{s_i\}_{i=1}^h$ *and the signed distance from the feature vectors to the decision boundary,* $d_\theta(\mathbf{e}) = \{d_\theta(\mathbf{e}_i)\}_{i=1}^h$*,*

$$\begin{aligned} DBC(\mathbf{s}, d_\theta(\mathbf{e})) &= \mathbb{E}[(\mathbf{s} - \bar{\mathbf{s}})d_\theta(\mathbf{e})] - \mathbb{E}[\mathbf{s} - \bar{\mathbf{s}}]\bar{d}_\theta(\mathbf{e}) \\ &\approx \frac{1}{h}\sum_{i=1}^{h}(\mathbf{s}_i - \bar{\mathbf{s}})d_\theta(\mathbf{e}) \end{aligned} \tag{9}$$

where $\mathbb{E}[\mathbf{s} - \bar{\mathbf{s}}]\bar{d}_\theta(\mathbf{e})$ is cancels out since $\mathbb{E}[\mathbf{s} - \bar{\mathbf{s}}] = 0$ 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 $\theta^T\mathbf{e} = 0$. A point θ_t in the domain of a task is feasible if it satisfies the constraint $g_t(\theta_t) \leq 0$. More concretely, $g_t(\theta_t)$ is defined by the definition of DBC in Eq.(9), *i.e.*

$$g_t(\theta_t) = \left| \frac{1}{2K} \sum_{\mathbf{s}_i, \mathbf{e}_i \sim \mathcal{T}_t} (\mathbf{s}_i - \bar{\mathbf{s}}) d_{\theta_t}(\mathbf{e}_i) \right| - c \tag{10}$$

where c is a small positive fairness relaxation. To formalize the supervised classification problem in the context of meta-learning definitions, a cross-entropy loss function is used to describe the adapted loss over a support set for each task. Integrated with DBC fairness constraint, the classification problem of a single task is formulated as follow

$$\begin{aligned} \min_{\theta_t \in \Theta} \quad & f_t(\theta_t) = \sum_{(\mathbf{e}^i, y^i)\sim\mathcal{T}_t} y^i \log \hat{y}(\mathbf{e}^i, \theta_t) \qquad (11) \\ & \qquad + (1 - y^i)\log(1 - \hat{y}(\mathbf{e}^i, \theta_t)) \\ \text{subject to} \quad & \left| \frac{1}{2K} \sum_{\mathbf{s}_i, \mathbf{e}_i \sim \mathcal{T}_t} (\mathbf{s}_i - \bar{\mathbf{s}}) d_{\theta_t}(\mathbf{e}_i) \right| \leq c \end{aligned}$$

where $(\mathbf{e}^i, y^i)$ are an input/output pair sampled from task $\mathcal{T}_t$ and $\hat{y}$ is a predicted outcome. The goal of a single task optimization is to approximate a good parameter pair (θ_t', μ_t') 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 (θ, μ) are iteratively updated using the proposed dual decomposition approach outlined in Algorithm 2.

VI. Experimental Settings

To validate our approach of unfairness prevention in few-shot meta-learning models, we conduct experiments with three real-world datasets which are available from the UCI ML-repository.

TABLE I: Key characteristics and statistics of real dataset.

Data	Adult	Communities and Crime	Bank
s	{M, F}	{Black, non-Black}	{Married, non-Married}
y	income $\{\geq$ or $< 50K\}$	crime rate $\{\geq$ or $< 50\%\}$	deposit {Yes, No}
# of instance	48,842	2,216	41,188
tasks	countries	states	months and dates
# of total tasks	34	46	50
# of input features	12	98	17
tasks for training	22	30	40
tasks for validation	6	8	5
tasks for testing	6	8	5
DBC	0.043	0.052	0.067
Discrimination	0.195	0.214	0.028
Consistency	0.485	0.222	0.377

A. Data

The **Adult** income dataset [42] 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.

Communities and Crime dataset [43] 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 [33], 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.

Bank Marketing dataset [44] 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 dataset contains 50 tasks.

B. Evaluation Metrics

To evaluate the proposed techniques for fairness learning, we introduced two classic evaluation metrics 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.

Discrimination measures the bias with respect to the protected attribute S in the classification:

$$\text{Disc} = \left| \frac{\sum_{i:s_i=1} \hat{y}_i}{\sum_{i:s_i=1} 1} - \frac{\sum_{i:s_i=0} \hat{y}_i}{\sum_{i:s_i=0} 1} \right|$$

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 [14] compares a model's classification prediction of a given data item to its k-nearest neighbors:

$$\text{Cons} = 1 - \frac{1}{|D|k} \sum_{i=1}^{|D|} \left| \hat{y}_i - \sum_{j \in kNN(\mathbf{e}_i)} \hat{y}_j \right|$$

where $|D|$ is the sample size, k is the number of nearest neighbors, and a nearest neighbor is defined based on a similarity measure (*i.e.* euclidean distance) of unprotected attributes $\mathbf{e}$. As demonstrated in [14], 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.

C. Baseline Methods

We evaluate all datasets – the proposed approach against various baselines – by comparing the results of generalization on both classification 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 removing the protected attributes.
3) **pretrain**: In computer vision, models pre-trained on large-scale image classification have been shown to learn effective features [46]. In this paper, the pre-train baseline trains a single network on all tasks and in each task an unified fairness constraint is added to ensure DBC is satisfied.
4) **fair-MAML**: [35] controls unfairness for each task and tunes a shared Lagrangian multiplier across tasks by simply applying grid search.
5) **F-MAML**$_{dp}$: is a fair meta-learning approach proposed in [33]. 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 $\mathcal{R}_{dp} = 1 - p(\hat{y} = 1 | s = 0)$.
6) **F-MAML**$_{eop}$: is another fair meta-learning approach proposed in [33], in which the demographic parity regularizer is replaced with the one aimed at improving equal opportunity, where $\mathcal{R}_{eop} = 1 - p(\hat{y} = 1 | s = 0, y = 1)$.
7) **LAFTR** [45]: 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.

D. 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

TABLE II: Consolidated overall result for few-shot classification.

		Adult Data				Communities and Crime				Bank Marketing			
K	**Approach**	**Acc**	**DBC**	**Disc**	**Cons**	**Acc**	**DBC**	**Disc**	**Cons**	**Acc**	**DBC**	**Disc**	**Cons**
-	Data	-	0.043	0.195	0.485	-	0.052	0.214	0.222	-	0.067	0.028	0.377
5-shot	MAML [1]	82.1%	0.046	0.227	0.883	**98.4%**	0.039	0.450	0.726	61.1%	0.026	0.122	0.884
	Masked MAML	79.9%	-	0.157	0.916	85.8%	-	0.322	0.846	57.6%	-	0.083	0.926
	pretrain	76.5%	0.024	0.239	0.907	84.5%	0.030	0.337	0.815	57.1%	0.018	0.106	0.923
	fair-MAML [35]	59.7%	0.028	0.146	0.909	77.2%	0.026	0.358	0.758	56.2%	0.012	0.057	**0.952**
	F-MAML$_{dp}$ [33]	**82.8%**	0.030	0.159	0.913	95.1%	0.039	0.442	0.757	59.3%	0.017	0.081	0.929
	F-MAML$_{eop}$ [33]	79.5%	0.029	0.153	0.916	95.0%	0.041	0.387	0.775	57.0%	0.017	0.083	0.927
	LAFTR [45]	72.0%	0.035	0.188	0.891	89.2%	0.050	0.440	0.787	**62.1%**	0.030	0.100	0.865
	Ours	78.2%	**0.003**	**0.026**	**0.937**	79.0%	**0.013**	**0.200**	**0.893**	55.9%	**0.005**	**0.026**	0.950
10-shot	MAML [1]	81.9%	0.045	0.211	0.900	**99.6%**	0.038	0.463	0.760	59.7%	0.020	0.089	0.898
	Masked MAML	80.0%	-	0.143	0.930	86.5%	-	0.275	0.864	57.7%	-	0.059	0.941
	pretrain	78.8%	0.023	0.125	0.923	83.2%	0.030	0.293	0.849	58.3%	**0.008**	0.039	0.969
	fair-MAML [35]	70.0%	0.030	0.146	0.925	83.6%	0.035	0.356	0.797	60.2%	0.016	0.061	0.942
	F-MAML$_{dp}$ [33]	78.2%	0.025	**0.114**	0.943	97.6%	0.036	0.432	0.781	59.3%	0.013	0.058	0.945
	F-MAML$_{eop}$ [33]	71.9%	0.028	0.134	0.927	94.1%	0.032	0.253	0.901	57.3%	0.012	0.054	0.941
	LAFTR [45]	72.3%	0.030	0.179	0.912	90.1%	0.050	0.401	0.790	**62.3%**	0.025	0.098	0.877
	Ours	**83.8%**	**0.011**	0.123	**0.943**	90.1%	**0.016**	**0.215**	**0.927**	61.3%	0.010	**0.027**	**0.973**
15-shot	MAML [1]	**82.7%**	0.039	0.179	0.909	**99.1%**	0.047	0.380	0.788	60.4%	0.016	0.068	0.903
	Masked MAML	80.2%	-	0.141	0.934	86.2%	-	0.246	0.870	58.1%	-	0.049	0.947
	pretrain	80.6%	0.024	0.117	0.927	84.8%	0.029	0.264	0.859	57.6%	0.014	0.063	0.939
	fair-MAML [35]	65.4%	0.022	0.103	0.924	83.4%	0.021	0.221	0.895	56.2%	0.010	0.044	0.960
	F-MAML$_{dp}$ [33]	81.0%	0.030	0.141	0.935	94.8%	0.039	0.313	0.812	57.9%	0.011	0.046	0.946
	F-MAML$_{eop}$ [33]	80.8%	0.028	0.129	0.938	95.3%	0.040	0.320	0.815	58.4%	0.011	0.050	0.946
	LAFTR [45]	75.5%	0.029	0.159	0.915	91.2%	0.030	0.299	0.825	**61.1%**	0.012	0.089	0.892
	Ours	80.4%	**0.005**	**0.011**	**0.985**	80.6%	**0.009**	**0.093**	**0.959**	57.0%	**0.005**	**0.010**	**0.989**
20-shot	MAML [1]	82.5%	0.044	0.185	0.914	**99.8%**	0.048	0.380	0.774	60.8%	0.014	0.062	0.912
	Masked MAML	80.8%	-	0.137	0.938	84.8%	-	0.242	0.876	57.8%	-	0.042	0.952
	pretrain	80.4%	0.021	0.100	0.935	84.9%	0.027	0.229	0.869	57.5%	0.012	0.053	0.942
	fair-MAML [35]	69.7%	0.018	0.083	0.931	86.0%	0.018	0.229	0.891	55.2%	**0.005**	0.044	0.964
	F-MAML$_{dp}$ [33]	80.6%	0.028	0.132	0.939	98.0%	0.042	0.314	0.816	**67.4**%	0.010	0.042	0.951
	F-MAML$_{eop}$ [33]	**83.3%**	0.029	0.135	0.936	95.7%	0.038	0.318	0.817	58.1%	0.010	0.041	0.948
	LAFTR [45]	76.2%	0.032	0.175	0.911	89.8%	0.029	0.353	0.810	62.1%	0.015	0.095	0.875
	Ours	79.2%	**0.001**	**0.018**	**0.988**	85.7%	**0.008**	**0.076**	**0.965**	57.5%	0.006	**0.006**	**0.991**

step gradient update (*i.e.* $q = 1$) and $k = 10$ inner primal-dual updates with $2NK$ samples of query set, and a fixed primal and dual learning rate of $\gamma = 0.01$ and $\alpha = 0.01$. We use Adam as the meta-optimizer. Because we only consider a binary classification problem, all of tasks are 2-way, *i.e.* $N = 2$. 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 three 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 I.

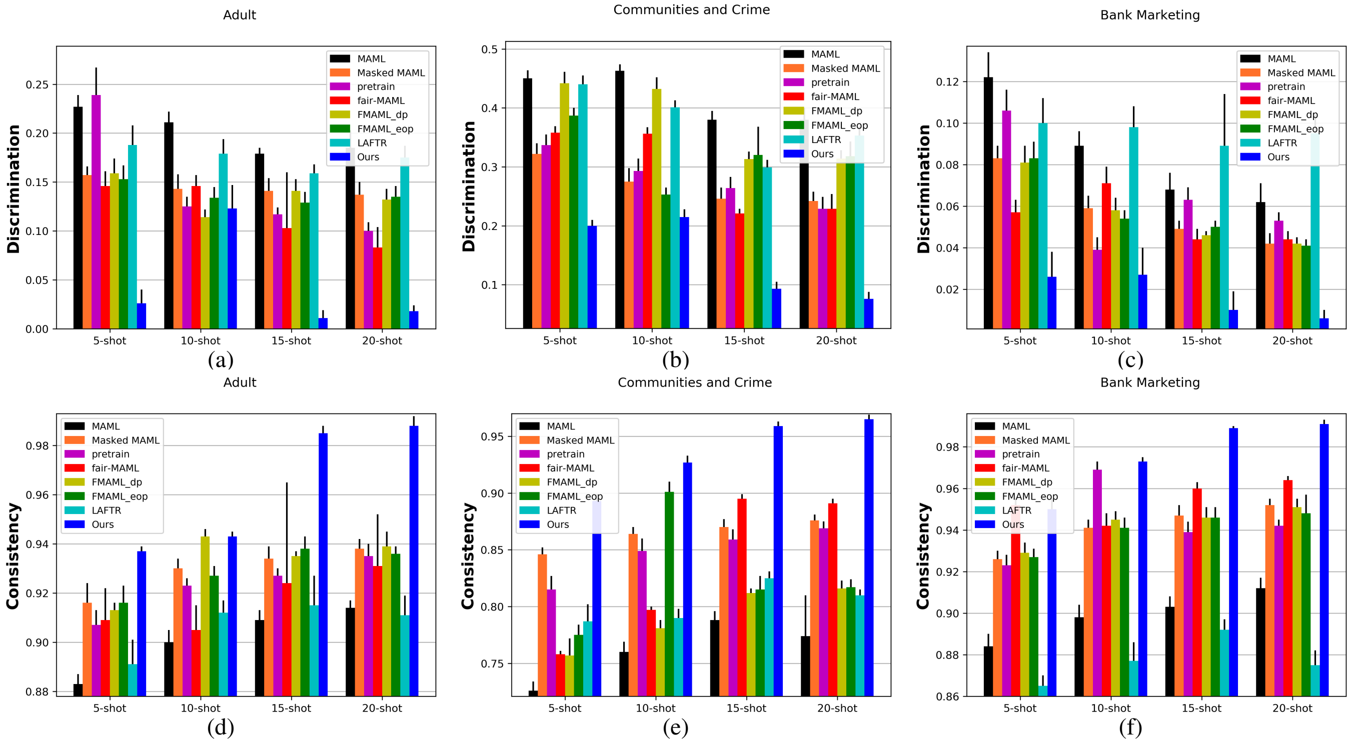
All baseline models used to compare with our proposed approach share the same neural network architecture and parameter settings. Hyperparameters are selected by a held-out validation procedure. All experiments are repeated 10 times with the same settings. Results shown with these methods in this paper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

VII. Experiment Results

This section evaluates the effectiveness of the proposed approach and its competitors on a classification task. We focus on generalization of statistical parity on unseen tasks and trade-off between validation loss and fairness that the proposed dual subgradient method alleviates when used to train classifiers. 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 *DBC* and all other evaluation metrics.

Consolidated and detailed performance of the different techniques over real-world data are listed in Table II. We evaluate performance by fine-tuning the model learned by all methods on K-shot of $\{5, 10, 15, 20\}$ datapoints of each class for each dataset. Best performance 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 in Table II demonstrates our proposed approach out-performs than other baseline methods in terms of controlling biases. It efficiently reduces *DBC* from the original dataset and values of *DBC* are limited to close zero that signify a fair prediction. In addition, fairness results based on two fair evaluation metrics, *i.e. Disc* (Figure 2 (a-c)) and *Cons* (Figure 2 (d-f)), are plotted 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.

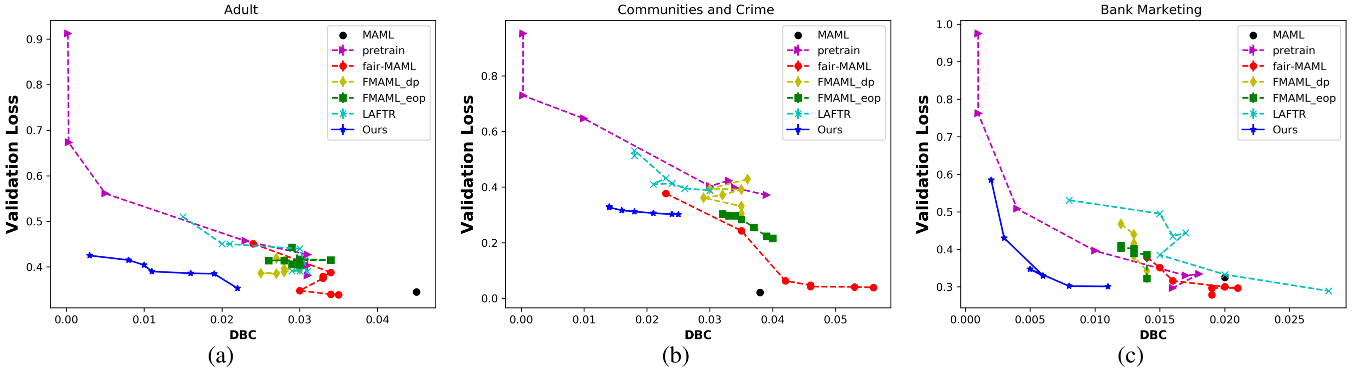
MAML became a famous meta-learning algorithm because of its fast adaptation and good generalization performance on losses [1]. However, our results shows it fails to control biases nor performs success in fairness generalization in a few-shot meta-learning, although *MAML* is stably able to produce high generalization accuracy. *Masked MAML* shows an improvement in fairness; however, there is still substantial unfairness hidden in the data in the form of correlated attributes. *F-MAML*$_{dp}$



(a) (b) (c)

(d) (e) (f)

Fig. 2: Experiment results of real-world datasets in controlling biases.



(a) (b) (c)

Fig. 3: The validation loss/fairness trade off sweeping over a range of dual variables.

and *F-MAML$_{eop}$* proposed by *Slack et al.,* in [33] intuitively control unfairness by taking advantage of demographic parity and equal opportunity, respectively. Our results in Figure 2 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*. Furthermore, though *LAFTR* offers a way to transfer machine learning models between tasks, consistent with [33], we observe it is unsuccessful in very data light situations. Besides, it is worth noting that we outperform baseline methods in bias controlling with better results as the number of training data increases.

Although our proposed approach, *PDFM*, returns a bit smaller predictive accuracies (see Table II), 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 variables: $[0.001, 0.01, 0.1, 1, 10, 100, 1000]$. Taking 10-shot as an example, results presented in Figure 3 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 fine-tuning epochs and as the dual variable increases, its validation losses decrease and thus *DBC* increases. Moreover, *LAFTR* is not successful at learning with minimal data and a small number of fine-tuning epochs for the new task. At low values, *fair-MAML, F-MAML$_{dp}$*, and *F-MAML$_{eop}$* are able to achieve lower validation losses than the *pretrain* and *LAFTR* baselines. Crucially, the results stated in Figure 3 confirm and further illustrate the findings that our proposed *PDFM* is able to learn more accurate representations that are also fairer for the swept range than all baseline techniques.

VIII. 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 Primal-Dual Fair Meta-learning (PDFM) 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," *ICML*, 2017.

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

[3] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *AISTATS*, 2017.

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

[15] 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.

[16] R. L. Raffard, 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. 2453–2459, 2004.

[17] 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.

[18] H. Yao, Y. Wei, J. Huang, and Z. Li, "Hierarchically structured meta-learning," *ICML*, 2019.

[19] 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.

[20] 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.

[21] 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.

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

[23] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," *ICLR*, 2019.

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

[25] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," *ICDM*, 2013.

[26] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact." *KDD*, 2015.

[27] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning." *NeurIPS*, 2016.

[28] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression." *FAT ML*, 2018.

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

[30] D. Gondek and T. Hofman, "Non-redundant clustering with conditional ensembles." *KDD*, 2005.

[31] T. Kamishima and S. Akaho, "Considerations on recommendation independence for a find-good-items task." *In Workshop on Responsible Recommendation*, 2017.

[32] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *KDD(2018)*, pp. 2219–2228.

[33] 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.

[34] C. Zhao and F. Chen, "Unfairness discovery and prevention for few-shot regression." *ICKG*, 2020.

[35] C. Zhao, C. Li, J. Li, and F. Chen, "Fair meta-learning for few-shot classification." *ICKG*, 2020.

[36] L. Zhang, Y. Wu, and X. Wu, "Fairness-aware classification: Criterion, convexity, and bounds." *AAAI*, 2019.

[37] N. Goel, M. Yaghini, , and B. Faltings, "Non-discriminatory machine learning through convex fairness criteria." *AAAI*, 2018.

[38] A. Rush and M. Collins, "A tutorial on dual decomposition and lagrangian relaxation for inference in natural language processing." *Journal of Artificial Intelligence Research*, 2012.

[39] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[40] A. Nedic and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM Journal on Optimization*, vol. 19, pp. 1757–1780, 01 2009.

[41] A. Fallah, A. Mokhtari, and A. Ozdaglar, "On the convergence theory of gradient-based model-agnostic meta-learning algorithms." *AISTATS*, 2020.

[42] R. Kohavi and B. Becker, "Uci machine learning repository," 1994.

[43] M. Lichman, "Uci machine learning repository," 2013.

[44] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing." *Decision Support Systems*, 2014.

[45] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations." *ICML*, 2018.

[46] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition." *ICML*, 2014.