

Fair Meta-Learning For Few-Shot Classification

Chen Zhao, Changbin Li, Jincheng Li, Feng Chen

Department of Computer Science

The University of Texas at Dallas

Richardson Texas, USA

{chen.zhao, changbin.li, jincheng.li, feng.chen}@utdallas.edu

Abstract—Artificial intelligence nowadays plays an increasingly prominent role in our life since decisions that were once made by humans are now delegated to automated systems. A machine learning algorithm trained based on biased data, however, tends to make unfair predictions. Developing classification algorithms that are fair with respect to protected attributes of the data thus becomes an important problem. Motivated by concerns surrounding the fairness effects of sharing and few-shot machine learning tools, such as the Model Agnostic Meta-Learning [1] framework, we propose a novel fair fast-adapted few-shot meta-learning approach that efficiently mitigates biases during meta-train by ensuring controlling the decision boundary covariance that between the protected variable and the signed distance from the feature vectors to the decision boundary. Through extensive experiments on two real-world image benchmarks over three state-of-the-art meta-learning algorithms, we empirically demonstrate that our proposed approach efficiently mitigates biases on model output and generalizes both accuracy and fairness to unseen tasks with a limited amount of training samples.

Index Terms—decision boundary covariance, statistical parity, few-shot, meta-learning

I. INTRODUCTION

In data mining and machine learning, the information system is becoming increasingly reliant on statistical inference and learning to give automated prediction and decision-making to solve regression and classification problems. Biased historical data or data containing biases, however, are often learned and thus lead to results with undesirability, inaccuracy, and even illegality. In recent years, there are increasing numbers of news reported that human bias is revealed in an artificial intelligence system applied by high-tech companies. [2] reported that a picture of two African Americans was automatically tagged as “Gorillas” by *Google Photos*. A 2016 study [3] found that the data-driven system developed by *Amazon* that used to determine the neighborhoods in which to offer free same-day delivery is highly biased and unfair to African American communities due to the stark disparities in the demographic makeup of neighborhoods: white residents were more than twice as likely as African American residents to live in one of the qualifying neighborhoods. Critics have voiced that human bias potentially has an influence on nowadays technology, which leads to outcomes with unfairness. Another example for biased image classification problem is shown in Figure 1: a dog classifier is trained with images of dogs lying on the grass. The training process goes through a feature

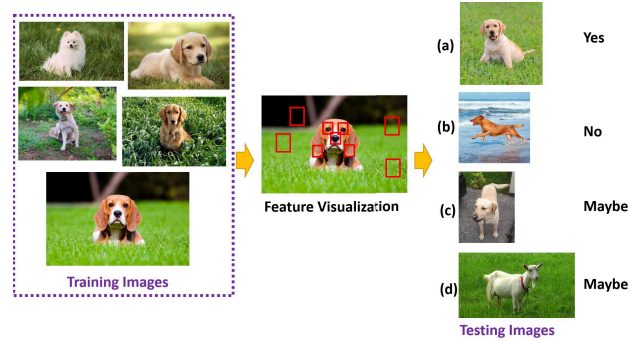


Fig. 1: An example of biased image classification problem. A dog classifier is trained using images of dogs lying on the grass. Features learned from the neural network, however, are not fully concentrated on the target objects (*i.e.* dogs). As a consequence, the biased learner leads to unfair or uncertain decision-makings: (a) a dog on the grass; (b) a dog running on the beach; (c) a dog on the stone step; (d) a goat on the grass.

extractor, but the captured features used for classifier training are not totally concentrated on target objects (*i.e.* dogs). As a consequence, the decision-making accuracy for testing images does not turn out well. To investigate the reason, we deduce that there is a non-negligible relationship between the predicted outcome and the protected feature (*i.e.* grass in this example), which leads to an unfair result.

To ameliorate this unfairness problem, one may attempt to make the automated decision-maker blind to the protected attributes [4]. This however, is difficult, as many attributes may be correlated with the protected one [5]. With biased input, the main goal of training an unbiased model is to make the output fair. In other words, the predicted outcomes are statistically independent on protected variables. Statistical parity, also known as group fairness, ensures that the overall proportion of members in a protected group receiving predictions (*i.e.* positive/negative classification) are identical to the proportion of the population as a whole.

To the best of our knowledge, unfortunately, 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

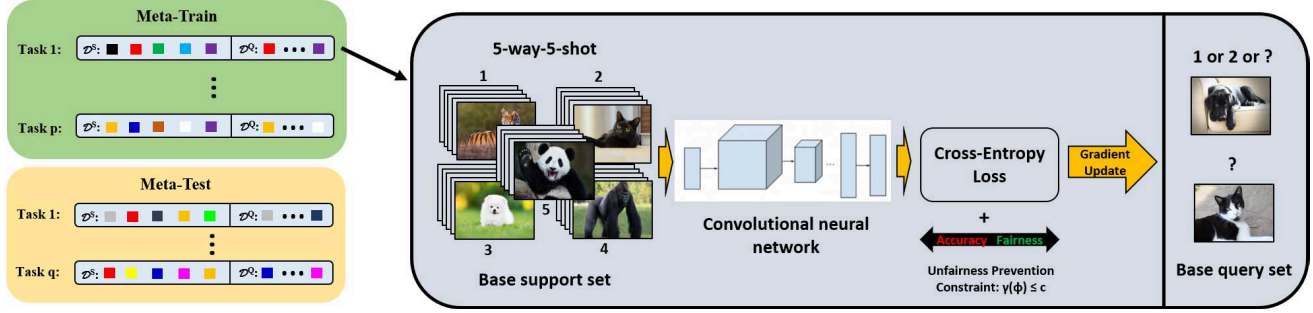


Fig. 2: An overview of our proposed bias controlling approach in few-shot meta-learning. (Left) The meta-learning framework include two processes: meta-train and meta-test, where each includes multiple tasks. (Right) Taking 5-way-5-shot classification problem as an example, $\gamma(\phi)$ refers to a discrimination measure function, and c is a predefined small threshold to account for a degree of randomness in the decision making process and sampling. The constraint here is considered to ensure no discrimination on the prediction model for each task.

human intelligence. In meta-learning, also known as learning to learn, the goal of trained model is to quickly learn a new task from a small amount of new data (*i.e.* few-shot), and the model is trained by the meta-learner to be able to learn on a large number of different tasks [1]. In contrast to traditional machine learning algorithms, such as multi-task learning [6]–[9] and transfer learning [10]–[13], meta-learning framework has advantages: (1) it learns across tasks where each task takes one or few samples as input; (2) it therefore efficiently speeds up model adaptation (3) and generalizes accuracy to unseen tasks.

The overall idea of existing methods of meta-learning, however, is to train a model which is capability of generalizing accuracy, rather than fairness, to unseen data tasks. But techniques for unfairness prevention and bias control in the few-shot meta-learning study are challenging and rarely touched. To ensure prediction without biases, the main contribution to this paper is that we feed each support set of a task with unified group fairness constraints and minimize meta-loss overall episodes. Specifically, we mitigate biases in each episode during meta-training by controlling the decision boundary covariance [14] which is defined as the covariance between the protected attribute and the signed distance from the feature vectors to the decision boundary. A value of zero signifying no dependency or attribute effect. Our experimental results demonstrate our approach is capability of controlling bias and decreasing loss as well as generalizing both to unseen tasks. In the context of classification, for example, as shown in Figure 2, each support set of a task used for training contains images sampled from 5 different classes ($N = 5$ ways) and each class includes 5 images ($K = 5$ shots). By giving an unified meta-initialization for each task, a task specific local model parameter is learned through one or few steps gradient update of the loss function that is constrained by fairness condition. To update the meta-parameter, the generalization error, *i.e.* the summation of the query loss across all tasks, is minimized. In summary, the main contributions of this paper are listed:

- For the first time the issue of bias control in meta-

learning multi-class classification problem is applied to image data sets. We mitigate biases by controlling the decision boundary covariance.

- We develop a novel algorithm to solve this constrained classification problem under the Model Agnostic Meta-Learning (MAML) few-shot framework.
- We validate the performance of our proposed approach of controlling biases on three state-of-the-art meta-learning techniques through extensive experiments based on real-world data sets. Our results demonstrate the proposed approach is capability of mitigating biases and generalizing both accuracy and fairness to unseen tasks, with the input training data is minimal.

In Section 2, some related works are referred. In Section 3, we see how unfairness is important in a machine learning model by introducing a simple causal based knowledge graph and how a statistical parity constraint, *i.e.* decision boundary covariance, is able to be used for bias-control in a single task. In Section 4, the fair few-shot meta-learning problem is formulated and how to solve it by applying the Model-Agnostics Meta-Learning framework is presented in detail. In Section 5, to validate the proposed approach, we conduct experiments by using two real-world benchmarks and three cutting-edge techniques, and we conclude this paper in Section 6.

II. RELATED WORK

In recent years, researches involving processing biased data became increasingly significant. Fairness-aware in data mining is classified into unfairness discovery and prevention. Based on the taxonomy by tasks, it can be further categorized to classification [14]–[22], regression [9], [23]–[26], clustering [27], [28], recommendation [29]–[31] and dimension reduction [32].

A. Unfairness Prevention in Classification

Majority of works in unfairness prevention is concentrated on data classification. According to approaches studied in

fairness, bias-prevention in classification is subcategorized into pre-processing [15], in-processing [14], [16] and post-processing [17]–[19]. Recent works [20] and [21] developed new approaches resulting in increasing the binary classification accuracy through reduction of fair classification to a sequence of cost-sensitive problems and applications of multi-task techniques with convex fairness constraints, respectively.

Non-discrimination (unfairness-free) 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 [33]. The first condition is related to direct discrimination. The second condition ensures that there is no indirect discrimination, also referred to as redlining. These types of discrimination (direct and indirect) are supported by two legal frameworks applied in large bodies of cases, disparate treatment and disparate impact [34]. The disparate treatment framework enforces procedural fairness, namely, the equality of treatments that prohibits the use of the protected attribute in the decision process. The disparate impact framework guarantees outcome fairness, namely, the equality of outcomes among protected groups [35]. Many of the prior studies, however, suffer from one or more of the following limitations: (i) they are restricted to a narrow range of classifiers, (ii) they only accommodate a single, binary sensitive attribute, and (iii) they cannot eliminate disparate treatment and disparate impact simultaneously. To overcome such limitations, in this paper, we consider the measure of decision boundary fairness [14], which enables us to ensure fairness with respect to one or more sensitive attributes, in terms of both disparate treatment and disparate impact.

B. Few-shot Meta-learning

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. Much efforts have been devoted to overcome the data efficiency issue. One popular category of few-shot learning techniques is distance metric learning based method, which addresses the few-shot classification problem by “learning to compare”. The intuition is that if a model can determine the similarity of two images, it can classify an unseen input image with the labeled instances. [36] introduced Matching Networks which employed ideas from k-nearest neighbors algorithm and metric learning based on a bidirectional Long-Short Term Memory (LSTM) to encode in the context of the support set. Prototypical networks [37] 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, such as [1], [38]–[42], aim to learn good model initialization so that the meta-loss is minimum.

The overall idea of these state-of-the-art is to train a meta-learning model which is capability of generalizing accuracy,

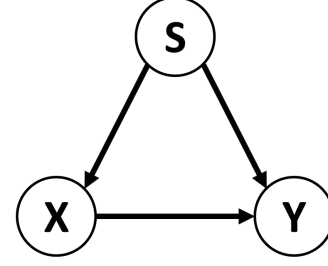


Fig. 3: A simple diagram demonstrates the causal relationship in fairness learning. X , S , and Y represent input feature, the protected attribute and target outcome, respectively.

but less attention on fairness generalization to unseen data tasks. In this paper, our proposed approach makes up for this regret of unfairness prevention using few-shot meta-learning techniques in multi-class classification problems.

III. MODELING OF FAIRNESS BASED ON CAUSAL KNOWLEDGE GRAPH

In this section, we first present how unfairness/bias affects decision-making by introducing a simple causal knowledge graph and then explain the mechanism of mitigating bias in a single task using the decision boundary covariance.

A. Causation in Fairness Learning

To understand how past decisions may bias a prediction model, we must first understand how the protected attribute may have affected the outcome by answering such questions: What would this outcome have been under different protected values? How would the outcome change if the protected attribute were changed, all else being equal? These questions are core to the mission of learning fair systems which aim to inform decision-making.

Unfairness can be broadly partitioned into two types: direct and indirect. The directed bias is concerned with settings where individuals received less favorable treatments on the basis of the protected attribute. The indirect one is concerned with individuals who receive treatments on the basis of inadequately justified factors that are somewhat related with the protected attribute [35].

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. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ be the data space, where $\mathcal{X} \subset \mathbb{R}^n$ is an input space, $\mathcal{S} = \{0, 1\}$ is a protected space, and $\mathcal{Y} = \{1, 2, \dots, N\}$ is an output space for multi-class classification where N is the number of classes. We consider a single task data $\mathcal{D} = \{(\mathbf{x}_i, y_i, s_i)\}_{i=1}^h$, $i = 1, \dots, h$, where $\mathbf{x}_i \in \mathbb{R}^n$ denotes the i -th observation, y_i denotes the corresponding output, $s_i \in \{s_+, s_-\}$ represents the binary protected attribute, and h is the number of observations in each task. A practical definition of fair causality is:

Definition 1 (Fair Causality and Causal Effect). X causes Y if and only if changing X leads to a change in Y , while keeping everything else (*i.e.* S) constant. Causal effect is defined as the magnitude by which Y is changed by a unit change in X .

Therefore, a fair prediction, shown in Figure 3, indicates there is no either direct ($S \rightarrow Y$) or indirect ($S \rightarrow X \rightarrow Y$) dependency effect of outcome on the protected attribute. These types of discrimination (direct and indirect) are supported by two legal frameworks applied in large bodies of cases throughout the disparate treatment and disparate impact [34]. The disparate treatment framework enforces procedural fairness, namely, the equality of treatments that prohibits the use of the protected attribute in the decision process. The disparate impact framework guarantees outcome fairness, namely, the equality of outcomes among protected groups [35].

To comply with disparate treatment criterion we specify that sensitive attributes are not used in decision making, *i.e.* $\{x_i\}_{i=1}^h$ and $\{s_i\}_{i=1}^h$ consist of disjoint feature sets. As discussed in [14], our definition of disparate impact leverages the 80%-rule [43]. A decision boundary satisfies the 80%-rule if the ratio between the percentage of users with a particular protected attribute value having $d_\alpha(\mathbf{x}) \geq 0$, where α is the decision boundary parameter, and the percentage of users without that value having $d_\alpha(\mathbf{x}) \geq 0$ is no less than 0.8 [14].

$$\min \left(\frac{P(d_\alpha(\mathbf{x}) \geq 0 | s = 1)}{P(d_\alpha(\mathbf{x}) \geq 0 | s = 0)}, \frac{P(d_\alpha(\mathbf{x}) \geq 0 | s = 0)}{P(d_\alpha(\mathbf{x}) \geq 0 | s = 1)} \right) \geq 0.8 \quad (1)$$

B. Decision Boundary Covariance in Statistical Parity

In this section, we introduce a measure of decision boundary fairness, which enables us to ensure fairness with respect to one or more protected attributes, in terms of both disparate treatment and disparate impact. The decision boundary covariance (DBC) which measures the decision boundary (un)fairness is defined as

Definition 2 (Decision Boundary Covariance [14]). 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, $d_\alpha(\mathbf{x}) = \{d_\alpha(\mathbf{x}_i)\}_{i=1}^h$, where α is the decision boundary parameter.

$$\begin{aligned} DBC(\mathbf{s}, d_\alpha(\mathbf{x})) &= \mathbb{E}[(\mathbf{s} - \bar{\mathbf{s}})d_\alpha(\mathbf{x})] - \mathbb{E}[\mathbf{s} - \bar{\mathbf{s}}]\mathbb{E}[d_\alpha(\mathbf{x})] \\ &\approx \frac{1}{h} \sum_{i=1}^h (s_i - \bar{s})d_\alpha(\mathbf{x}_i) \end{aligned} \quad (2)$$

where $\mathbb{E}[\mathbf{s} - \bar{\mathbf{s}}]\mathbb{E}[d_\alpha(\mathbf{x})]$ is cancels out since $\mathbb{E}[\mathbf{s} - \bar{\mathbf{s}}] = 0$. Taking linear model as an example, the decision boundary is simply the hyperplane defined by $\alpha^T \mathbf{x} = 0$. Then the DBC reduces to $\frac{1}{h} \sum_{i=1}^h (s_i - \bar{s})\alpha^T \mathbf{x}_i$.

An example of fair binary classification with a linear decision boundary is given in Figure 4. Red markers represent the protected group (*i.e.* $s = 1$) and blue ones are the unprotected group (*i.e.* $s = 0$). In the left of Figure 4, we calculate $P(d_\alpha(\mathbf{x}) \geq 0 | s = 1) = 1/4 = 0.25$, where $d_\alpha(\mathbf{x}) \geq 0$

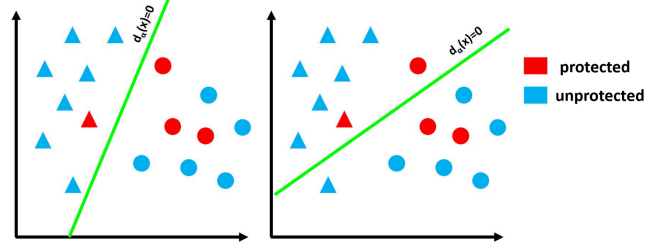


Fig. 4: An example of fair binary classification (unfair (left) and fair (right) classifier) by controlling the decision boundary covariance between the protected variable and the signed distance from user's feature to the decision boundary.

indicates the triangle class and $P(d_\alpha(\mathbf{x}) \geq 0 | s = 0) = 7/(16 - 4) = 0.583$. By applying the 80%-rule indicated in Eq.(1), the disparate impact value of the left classifier is $0.25/0.583 = 0.43$, which is lower than the threshold of 0.8 and returns an unfair classification prediction. Similarly, in the right case of Figure 4, however, $P(d_\alpha(\mathbf{x}) \geq 0 | s = 1) = P(d_\alpha(\mathbf{x}) \geq 0 | s = 0) = 0.5$ and thus the disparate impact is 1.0. Note that, if a decision boundary satisfies the 100%-rule, *i.e.*

$$P(d_\alpha(\mathbf{x}) \geq 0 | s = 1) = P(d_\alpha(\mathbf{x}) \geq 0 | s = 0)$$

then the empirical covariance will be approximately zero for a sufficiently large training set.

IV. FAIR META-LEARNING

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. MAML (Model-Agnostic Meta-Learning) proposed by [1] is one of the popular gradient based meta-learning frameworks, which leads to state-of-the-art performance and fast adaptation to unseen tasks. To generalize fairness in a classification problem with minimal samples, we propose a novel approach by modifying MAML in which we uniformly control DBC for each task. The goal of the proposed approach is to estimate a good meta-parameter such that the summation of empirical risks for each task is minimized and meanwhile each task is fair.

A. Settings

In this work, we consider a collection of supervised learning tasks $\mathcal{T} = \{(\mathcal{D}_j^S, \mathcal{D}_j^Q)\}_{j=1}^T$ 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 $(\mathcal{D}_j^S, \mathcal{D}_j^Q)$ explicitly contains a pair of a support \mathcal{D}_j^S and a query \mathcal{D}_j^Q data sets. For each task $j \in \{1, 2, \dots, T\}$, we let $\{\mathbf{x}_{j,i}, y_{j,i}, s_{j,i}\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y} \times \mathcal{S})$ 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 j -th task, it contains N categories and each consists of K datapoints. We emphasize that we need to sample without replacement, *i.e.*, $\mathcal{D}_j^S \cap \mathcal{D}_j^Q = \emptyset$.

In a general meta-learning setting, it consists of meta-train and meta-test partitions where each contains a number of mini-batches of episodes (see Figure 2). We consider a distribution over tasks $p(\mathcal{T})$ that we want our model to be able to adapt to. In a N -way- K -shot learning setting, a task \mathcal{T}_j is sampled from $p(\mathcal{T})$, where the subscript j represents the j -th task of a mini-batch. In the supervised learning setting, supposing the meta-model is a parameterized function f_ϕ with parameters ϕ . In a general meta-learning model, the goal is to learn an optimized ϕ so that the summation of query losses $l_{\mathcal{T}_j}(f_\phi)$ over all meta-training tasks is minimum. During meta-training, ϕ is updated iteratively.

$$\phi^* = \arg \min_{\phi} \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} l_{\mathcal{T}}(f_\phi) \quad (3)$$

B. Model-Agnostic Meta-Learning with convex constraints

Meta-learning approaches for few-shot learning are often assumed 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 $\mathcal{L}^{inner}(\mathcal{D}_j^S, \phi)$ such that it is constrained by a function $g_j(\phi)$.

$$\begin{aligned} \min_{\phi_j} \quad & \mathcal{L}^{inner}(\mathcal{D}_j^S, \phi) \\ \text{subject to} \quad & g_j(\phi) \leq 0 \end{aligned} \quad (4)$$

where $\mathcal{L}^{inner} : \mathbb{R}^n \rightarrow \mathbb{R}$ is a loss function, such as cross-entropy for classification, and $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is an appropriate complexity function ensuring the existence and the uniqueness of the above minimizer. A point ϕ in the domain of the problem is feasible if it satisfies the constraint $g_j(\phi) \leq 0$. Specifically, $g_j(\phi)$ is defined by the definition of decision boundary covariance in Eq.(2), *i.e.*

$$g_j(\phi) = \left| \frac{1}{N \times K} \sum_{\mathbf{s}_i, \mathbf{x}_i \sim \mathcal{T}_j} (\mathbf{s}_i - \bar{\mathbf{s}}) d_\alpha(\mathbf{x}_i) \right| - c \quad (5)$$

where c is a small positive fairness relaxation, $d_\alpha(\mathbf{x}_i) \approx \max\{\mathbf{p}_n \in [0, 1]^N\}$ and \mathbf{p} denotes the class probabilities of \mathbf{x}_i . Here, for a N -way- K -shot classification task, we include $N \times K$ data points in the support set \mathcal{D}_j^S of each task \mathcal{T}_j .

To solve the optimization problem, we thus introduce an unified Lagrange multiplier $\lambda \geq 0$ for all tasks and the Lagrange function $\mathcal{L}_{\mathcal{T}_j}(\phi, \lambda)$ for each task is defined by

$$\mathcal{L}_{\mathcal{T}_j}(\phi, \lambda) = \mathcal{L}^{inner}(\phi) + \lambda(g_j(\phi)) \quad (6)$$

Therefore the original problem can be finally seen by minimizing $\mathcal{L}_{\mathcal{T}_j}(\phi, \lambda)$ for each task and thus mitigates dependency of prediction on the protected attribute. The goal of training a single task is to output a local parameter ϕ_j given the meta-parameter ϕ such that it minimizes the task loss \mathcal{L}^{inner} subject to the task constraint $g_j(\phi) \leq 0$. Next, to update the meta-parameter, we minimize the generalization error \mathcal{L}^{meta} using query sets across every task in the batch such that the query

constraints are satisfied. Formally, the learning objective across all tasks is

$$\min_{\phi} \mathcal{L}^{meta} \left(\sum_{j=1}^T \mathcal{D}_j^Q, \phi \right) = \sum_{j=1}^T \mathcal{L}^{inner}(\mathcal{D}_j^Q, \phi_j) \quad (7)$$

where $\phi_j = \arg \min_{\phi_j, g_j(\phi) \leq 0} \mathcal{L}^{inner}$ is the local optimum for each task. A step-by-step learning algorithm for unfairness prevention in few-shot regression is proposed in Algorithm 1.

Algorithm 1 Unfairness Prevention in Few-Shot Classification.

Require: $p(\mathcal{T})$: distribution over tasks.

Require: α, β : step size hyperparameters.

Require: q : inner gradient update steps.

```

1: Randomly initialize  $\phi$ 
2: while not done do
3:   Sample batch of tasks  $\mathcal{T}_j$ 
4:   for all  $\mathcal{T}_j = \{\mathcal{D}_j^S, \mathcal{D}_j^Q\}$  do
5:     Sample  $N$ -way- $K$ -shot datapoints from  $\mathcal{D}_j^S = \{\mathbf{x}^j, \mathbf{y}^j, \mathbf{s}^j\}$ 
6:      $\phi_j \leftarrow \phi$ 
7:     for  $q = 1, 2, \dots$  do
8:       Evaluate  $\nabla_{\phi_j} \mathcal{L}_{\mathcal{T}_j}(\phi_j, \lambda)$  using  $\mathcal{D}_j^S$ 
9:       Compute adapted local parameter  $\phi_j \leftarrow \phi_j - \alpha \nabla_{\phi_j} \mathcal{L}_{\mathcal{T}_j}(\phi_j, \lambda)$ 
10:    end for
11:    Sample datapoints from  $\mathcal{D}_j^Q = \{\mathbf{x}^j, \mathbf{y}^j, \mathbf{s}^j\}$ 
12:    Evaluate query loss  $l_{\mathcal{T}_j}(\phi_j)$  and query fairness  $g_{\mathcal{T}_j}(\phi_j)$  using  $\mathcal{D}_j^Q$ 
13:  end for
14:  Update  $\phi \leftarrow \phi - \beta \nabla_{\phi} \sum_{\mathcal{T}_j \sim p(\mathcal{T})} l_{\mathcal{T}_j}(\phi_j)$ 
15:  Evaluate training fairness  $mean(g_{\mathcal{T}_j}(\phi_j))$ 
16: end while
```

C. Algorithm Analysis

Since the proposed Algorithm 1 is modified following [1], the convergence is guaranteed and detailed analysis is stated in [44]. Accessing to sufficient samples, the running time of the algorithm is $O(n \cdot b \cdot q)$, where n is the number of outer iterations, b is batch-size, and q is gradient steps of inner loop. For a N -way- K -shot learning, the best accuracy is achieved when $\|\nabla \phi\| \leq O(\tilde{\sigma}/\sqrt{NK})$, where $\phi = \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} l_{\mathcal{T}}(f_\phi)$, σ is a bound on the standard deviation of $\nabla \mathcal{L}_{\mathcal{T}_j}(\phi_j, \lambda)$ from its mean $\nabla \mathcal{L}_{\mathcal{T}}(\phi, \lambda)$, and $\tilde{\sigma}$ is a bound on the standard deviation of estimating $\nabla \mathcal{L}_{\mathcal{T}_j}(\phi_j, \lambda)$ using a single data point.

V. EXPERIMENTS

To validate our approach of unfairness prevention in few-shot meta-learning models, we conduct experiments with two real-world image data sets.

A. Data

Omniglot [45] is a data set of 1623 handwritten characters collected from 50 alphabets. To avoid overfitting, data augmentation is used on images in the form of rotations of 90 degrees increments, *i.e.* 90, 180, and 270 degrees. Rotated class samples are considered new classes and thus we have 1623×4 classes in total. We shuffle all character classes and randomly split the data set into three sets, 1150×4 for the training set, 50×4 for validation, and 423×4 for testing. We follow the procedure of [36] by resizing the grayscale images to 28×28 . In order to study fairness, an arbitrary probability $p(\mathbf{x}|s = 1)$ is assigned to each class and thus each image sample is given a protected attribute $s \in \{0, 1\}$.

TABLE I: Key Characteristics of Experimental Data

	Omniglot		mini-ImageNet
images	grey		color
augmentation	yes		no
scale	28×28		84×84
classes for training	1150×4		64
classes for validation	50×4		12
classes for test	423×4		24
N -way	5	20	5
local gradient step(s) for training	1	5	5
local gradient step(s) for evaluation	3	5	10
local learning rate for training	0.4	0.1	0.01
local learning rate for evaluation	0.4	0.1	0.01
meta batch-size	32	16	4 (1-shot), 2 (5-shot)

mini-ImageNet is originally proposed by [36]. It consists of 60,000 color images scaled down to 84×84 divided into 100 classes with 600 examples each. We use the split proposed in [38], which consists of 64 classes for training, 12 classes for validation and 24 classes for testing. The protected attribute is randomly added following the same procedure stated in our Omniglot settings.

B. Parameter Tuning

In order to provide a fair comparison for all methods, our embedding architecture mirrors that used by [36]. It consists of four modules and each comprises a 64-filter 3×3 convolution, batch normalization layer [46], a ReLU nonlinearity and a 2×2 max-pooling layer. For Omniglot, since each image is resized to 28×28 , it results in a 64-dimensional output space. Due to the increased size of images in mini-ImageNet, the resulting feature map is 1600-dimensional. All models are trained with Adam optimizer.

For N -way, K -shot classification, each gradient is computed using a batch size of $N \times K$ examples. For Omniglot, the 5-way convolutional model is trained with 1 gradient step with step size of $\alpha = 0.4$, and a meta batch-size of 32 tasks. The

network is evaluated using 3 gradient steps with the same step size $\alpha = 0.4$. The 20-way convolutional model is trained and evaluated with 5 gradient steps with step size of $\alpha = 0.1$. During training, the meta batch-size is set to 16 tasks. For MiniImagenet, the model is trained using 5 gradient steps of size $\alpha = 0.01$, and evaluated using 10 gradient steps at test time. Following [38], 15 examples per class are used for evaluating the post-update meta-gradient. We used a meta batch-size of 4 and 2 tasks for 1-shot and 5-shot training respectively. All model are trained for 60000 iterations.

Besides, to comply with disparate treatment criterion, we specify that protected attributes are not used in the decision making, *i.e.* protected and explanatory attributes consist of disjoint feature sets. Key characteristics for all data set are listed in Table I.

C. Baseline Methods

We compared our work with three well known meta-learning state-of-the-arts, MAML [1], Matching Networks [36], and Prototypical Networks [37]. These methods apply the same meta-learning framework but differ in their strategies to make predictions conditioned on the support set. MAML is a gradient based meta-learning algorithm, where each support set is used to adapt a uniform initialized model parameters using one or few gradient steps. After several updates, the meta-loss reaches the minimum along with each episode loss reaching its local minimum.

For both Matching and Prototypical Networks (*i.e.* MatchingNet and ProtoNet), the prediction of the samples in a query set is based on comparing the distance between embedded query feature and support feature within each class. MatchingNet was the first to both train and test on N -way- K -shot tasks. The appeal of this is training and evaluating on the same tasks lets us optimise for the target task in an end-to-end fashion. Different from earlier approaches such as siamese networks [47], MatchingNet combines both embedding and classification to form an end-to-end differentiable nearest neighbours classifier. Specifically, it applied a bidirectional Long-Short Term Memory (LSTM) to encode in the context of the support set and compares cosine distance between the query feature and each support feature.

In Prototypical Networks, the authors apply a compelling inductive bias in the form of class prototypes to achieve impressive few-shot performance that exceeds Matching Networks without the complication of full context embeddings or FCE for short. The key assumption is that there exists an embedding in which samples from each class cluster around a single prototypical representation which is simply the mean of the individual samples. This idea streamlines K -shot classification in the case of $K > 1$ as classification is simply performed by taking the label of the closest class prototype.

In order to output fair predictions, fairness constraints are applied. Experimental results shown with prefix “Fair-” in the front indicate models adjusted using our proposed approach.

TABLE II: Consolidated overall result for few-shot classification accuracies and fairness. Methods with superscript * and ‡ respectively indicates results reported by the original paper and our local replication with addition of protected attributes.

Omniglot	5-way				20-way			
	1-shot		5-shot		1-shot		5-shot	
Approach	Accuracy	DBC	Accuracy	DBC	Accuracy	DBC	Accuracy	DBC
MAML*	98.7±0.4%	-	99.9±0.1%	-	95.8±0.3%	-	98.9±0.2%	-
MAML‡	98.1±0.5%	0.39±0.01	98.1±0.1%	0.59±0.09	96.2±0.6%	0.50 ±0.10	98.7±0.4%	0.59±0.03
Fair-MAML	89.4±0.5%	0.18±0.01	92.2±0.4%	0.19±0.01	85.4±0.3%	0.21±0.03	90.4±0.4%	0.27 ±0.01
MatchingNet*	98.1%	-	98.9%	-	93.8%	-	98.7%	-
MatchingNet‡	98.0±0.4%	0.45±0.02	98.6±0.3%	0.45±0.01	95.2±0.5%	0.47±0.02	97.9±0.3%	0.48±0.03
Fair-MatchingNet	96.5±0.5%	0.14±0.01	98.0±0.5%	0.15±0.01	94.1±0.5%	0.13±0.03	98.5±0.5%	0.13±0.01
ProtoNet*	98.8%	-	99.7 %	-	96.0%	-	98.9%	-
ProtoNet‡	98.5±1.8%	0.44±0.01	99.1±0.3%	0.44±0.03	96.6±1.7%	0.46±0.01	98.6±0.1%	0.48±0.01
Fair-ProtoNet	86.4±2.0%	0.10±0.07	94.3±0.5%	0.10±0.02	88.7±1.5%	0.03±0.02	89.4±0.5%	0.02±0.01

mini-ImageNet	5-way	1-shot		5-shot	
		Accuracy	DBC	Accuracy	DBC
MAML*		48.7±1.8%	-	63.1±0.9%	-
MAML‡		44.2±1.1%	0.91±0.01	61.1±0.8%	0.89±0.01
Fair-MAML		35.5±0.9%	0.61±0.05	54.5±0.8%	0.58±0.07
MatchingNet*		43.6±0.8%	-	55.3±0.7%	-
MatchingNet‡		44.8±0.1%	0.10±0.01	60.9±1.0%	0.13±0.03
Fair-MatchingNet		37.1±2.0%	0.06±0.05	56.5±2.0%	0.07±0.05
ProtoNet*		49.4±0.8%	-	68.2±0.7%	-
ProtoNet‡		43.3±2.9%	0.09±0.03	64.2±1.0%	0.14±0.01
Fair-ProtoNet		39.9±1.0%	0.05±0.02	59.9±2.0%	0.06±0.05

D. Experimental Results

Table II showcases results experimented through three cutting-edge meta-learning methods and those associated with our proposed unfairness prevention approach (noted with “Fair-”), which examined with two real-world image data sets, *i.e.* Omniglot [45] and mini-ImageNet [36]. The problem of N -way classification is set up as follows: select N unseen classes, provide the model with K different instances of each of the N classes, and evaluate the model’s ability to classify new instances within the N classes.

In order to check the generalized fairness of these state-of-the-arts on unseen tasks, we produce local replications labeled with superscript ‡ that are used to compare with the results reported in the original paper labeled with *. Note that for local replication methods, input images are slightly different, where in this paper we additionally consider the protected attribute as one of the input features. Besides, in the proposed unfairness prevention approach (labeled with the prefix “Fair-” in Table II), cross-entropy losses are calculated with using images without the protected attribute, as the fairness constraint is applied to control the covariance between the protected variable and the signed distance from the feature vectors to the decision boundary.

Besides, our approaches (Fair-MAML, Fair-MatchingNet, and Fair-ProtoNet) are outperformed than the original methods and the gap between all methods is narrowing as the number of training data increases (*i.e.* from 1-shot to 5-shot). However,

in Omniglot data, as more image classes are considered in the experiment, accuracy decreases. Even in the 20-way classification problem, our approach is able to control the decision boundary covariance efficiently.

Our proposed approach is empirically shown to reduce DBC and thus improve outcome fairness in multi-class decision-making for all selected meta-learning methods. Another notable observation is that decreasing unfairness is brought at the sacrifice of classification accuracy in a bit. This comes down to the trade-off between accuracy and fairness. In summary, our approach significantly controls biases for few-shot meta-learning models and generalizes to unseen tasks.

VI. CONCLUSION AND FUTURE WORK

In this paper, for the first time we develop deep into the few-shot supervised meta-learning classification model and propose a bias-control approach by adding statistical parity constraint, namely decision boundary covariance, which significantly mitigates dependence of prediction on the protected variable in each task and generalize both accuracy and fairness to unseen tasks. Experimental results on two real-world image data sets indicate the proposed approach works for three cutting-edge few-shot meta-learning models with multi-class classification problems. Further research in this area can make multitask parameters a standard ingredient in deep fairness learning.

ACKNOWLEDGEMENT

This work is supported by the National Science Foundation (NSF) under Grant No #1815696 and #1750911.

REFERENCES

- [1] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- [2] A. Barr, "Google mistakenly tags black people as 'gorillas,' showing limits of algorithms," *The Wall Street Journal*, 2015.
- [3] D. Ingold and S. Soper, "Amazon doesn't consider the race of its customers. should it?" *Bloomberg*, 2016.
- [4] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
- [5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," *International Conference on Machine Learning (ICML)*, 2013.
- [6] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l21-norm minimization," *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.
- [7] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.
- [8] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, 2018.
- [9] C. Zhao and F. Chen, "Rank-based multi-task learning for fair regression," *In Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 916–925, 2019.
- [10] N. Dong and E. P. Xing, "Domain adaption in one-shot learning," *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2018.
- [11] Y.-C. Hsu, Z. Lv, , and Z. Kira, "Learning to cluster in order to transfer across domains and tasks," *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.
- [12] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *In Proceedings of the International Conference on Machine Learning (ICML)*, 2015.
- [13] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2010.
- [14] M. B. Zafar, I. Valera, M. G. Rodríguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- [15] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.
- [16] B. Fish, J. Kun, and Ádám D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," *arXivpreprint:1601.05764*, 2015.
- [17] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *30th Conference on Neural Information Processing Systems (NIPS)*, 2016.
- [18] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Model-based and actual independence for fairness-aware classification," *Data Mining and Knowledge Discovery*, 2018.
- [19] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, "Discrimination-and privacy-aware patterns," *Data Mining and Knowledge Discovery*, 2015.
- [20] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," *International Conference on Machine Learning (ICML)*, 03 2018.
- [21] L. Oneto, M. Donini, A. Elders, and M. A. Pontil, "Taking advantage of multitask learning for fair classification," *the 2019 AAAI/ACM Conference (AIES)*, 2019.
- [22] Z. Wang, B. Dong, Y. Lin, Y. Wang, M. S. Islam, and L. Khan, "Co-representation learning framework for the open-set data classification," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 239–244.
- [23] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," *IEEE International Conference on Data Mining (ICDM)*, 2013.
- [24] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, 2015.
- [25] J. Komiyama, A. Takeda, J. Honda, and H. Shimao, "Nonconvex optimization for regression with fairness constraints," *In Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.
- [26] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair kernel learning," *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2017.
- [27] D. Gondek and T. Hofmann, "Non-redundant data clustering," *In Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2004.
- [28] D. Gondek and T. Hofman, "Non-redundant clustering with conditional ensembles," *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.
- [29] T. Kamishima, S. Akaho, H. Asoh, and I. Sato, "Model-based approaches for independence-enhanced recommendation," *The IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2017.
- [30] T. Kamishima and S. Akaho, "Considerations on recommendation independence for a find-good-items task," *In Workshop on Responsible Recommendation*, 2017.
- [31] A. Singh and T. Joachims, "Fairness of exposure in rankings," *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.
- [32] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *arXiv:1607.06520 [cs.CL]*, 2016.
- [33] I. Zliobaite, "A survey on measuring indirect discrimination in machine learning," *arXiv preprint arXiv:1511.00148*, 2015.
- [34] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review* 104, 2016.
- [35] J. Zhang and E. Bareinboim, "Fairness in decision-making – the causal explanation formula," *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
- [36] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *In 30th Conference on Neural Information Processing Systems (NIPS)*, 2016.
- [37] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *In 31th Conference on Neural Information Processing Systems (NIPS)*, 2017.
- [38] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," *International Conference for Learning Representations (ICLR)*, 2017.
- [39] C. Finn, K. Xu, , and S. Levine, "Probabilistic model-agnostic meta-learning," *In Advances in Neural Information Processing Systems (NIPS)*, 2018.
- [40] A. Nichol and J. Schulman, "Reptile: a scalable metalearning algorithm," *arXiv preprint arXiv:1803.02999*, 2018.
- [41] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
- [42] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," *International Conference for Learning Representations (ICLR)*, 2019.
- [43] D. Biddle, "Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing," *Gower*, 2005.
- [44] A. Fallah, A. Mokhtari, and A. Ozdaglar, "On the convergence theory of gradient-based model-agnostic meta-learning algorithms," *In the proceedings of the 23rd International Conference of Artificial Intelligence and Statistics (AISTATS)*, 2020.
- [45] B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum, "One shot learning of simple visual concepts," *The 41st Annual Meeting of the Cognitive Science Society (CogSci)*, 2011.
- [46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [47] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.