Fairness without Demographics through Adversarially Reweighted Learning

Preethi Lahoti∗
plahoti@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarland Informatics Campus

Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, Ed H. Chi
Google Research

ABSTRACT

Much of the previous machine learning (ML) fairness literature assumes that protected features such as race and sex are present in the dataset, and relies upon them to mitigate fairness concerns. However, in practice factors like privacy and regulation often preclude the collection of protected features, or their use for training or inference, severely limiting the applicability of traditional fairness research. Therefore we ask: How can we train a ML model to improve fairness when we do not even know the protected group memberships?

In this work we address this problem by proposing Adversarially Reweighted Learning (ARL). In particular, we hypothesize that non-protected features and task labels are valuable for identifying fairness issues, and can be used to co-train an adversarial reweighting approach for improving fairness. Our results show that ARL improves Rawlsian Max-Min fairness, with significant AUC improvements for worst-case protected groups in multiple datasets, outperforming state-of-the-art alternatives.

1 INTRODUCTION

As machine learning (ML) systems are increasingly used for decision making in high-stakes scenarios, it is vital that they do not exhibit discrimination. However, recent research [5, 15, 26] has raised several fairness concerns, with researchers finding significant accuracy disparities across demographic groups in face detection [9], healthcare systems [18], and recommendation systems [11]. In response, there has been a flurry of research on fairness in ML, largely focused on proposing formal notions of fairness [15, 16, 16, 50], and offering “de-biasing” methods to achieve these goals. However, most of these works assume that the model has access to protected features (e.g., race and gender), at least at training [10, 51], if not at inference [16, 20].

In practice, however, many situations arise where it is not feasible to collect or use protected features for decision making due to privacy, legal, or regulatory restrictions. For instance, GDPR imposes heightened prerequisites to collect and use protected features. Yet, in spite of these restrictions on access to protected features, and their usage in ML models, it is often imperative for our systems to promote fairness. For instance, regulators like CFPB require that creditors comply by fairness, yet prohibit them from using demographic information for decision-making.1 Recent surveys of ML practitioners from both public-sector [45] and industry [20] highlight this conundrum, and identify “addressing fairness without demographics” as a crucial open-problem with high significance to ML practitioners. Therefore, in this paper, we ask the research question: How can we train a ML model to improve fairness when we do not have access to protected features neither at training nor inference time, i.e., we do not know protected group memberships?

Goal: We follow the Rawlsian principle of Max-Min welfare for distributive justice [43]. In Section 3.1, we formalize our Max-Min fairness goals: to train a model that maximizes the minimum expected utility across protected groups with the additional challenge that, we do not know protected group memberships. It is worth noting that, unlike parity based notions of fairness [16, 50], which aim to minimize gap across groups, Max-Min fairness notion permits inequalities. For many high-stakes ML applications, such as healthcare, improving the utility of worst-off groups is an important goal, and in some cases, parity notions that equally accept decreasing the accuracy of better performing groups are often not reasonable.

Exploiting Correlates: While the system does not have direct access to protected groups, we hypothesize that unobserved protected groups S are correlated with the observed features X (e.g., race is correlated with zip-code) and class labels Y (e.g., due to imbalanced class labels). As we will see in Table 6 ($S$), this is frequently true. While correlates of protected features are a common cause for concern in the fairness literature, we show this property can be valuable for improving fairness metrics. Next, we illustrate how this correlated information can be valuable with a toy example.

Illustrative Example: Consider a classification task wherein our dataset consists of individuals with membership to one of the two protected groups: “orange” data points and “green” data points. The trainer only observes their position on the x and y axis. Although the model does not have access to the group (color), y is correlated with the group membership.

Figure 1: Computational-identifiability example

1 This work was conducted while the author was an intern at Google Research, Mountain View.
2 Creditors may not request or collect information about an applicant’s race, color, religion, national origin, or sex. Exceptions to this rule generally involve situations in which the information is necessary to test for compliance with fair lending rules. [CFPB Consumer Law and Regulations, 12 CFR §1002.5]
Although each group alone is well-separable (Figure 1(a-b)), we see in Figure 1(c) that the empirical risk minimizing (ERM) classifier over the full data results in more errors for the green group. Even without color (groups), we can quickly identify a region of errors with low y value (bottom of the plot) and a positive label (+). In Section 3.2, we will define the notion of computationally-identifiable errors that correspond to this region. These errors are in contrast to outliers (e.g., from label noise) with larger errors randomly distributed across the x-y axes.

The closest prior work to ours is DRO [17]. Similar to us DRO algorithm has the goal of fairness without demographics, aiming to achieve Rawlsian Max-Min Fairness for unknown protected groups. However, to achieve this, DRO uses distributionally robust optimization to optimize for the worst-case groups by focusing on improving any worst-case distributions, but as the authors point out, this runs the risk of focusing optimization on noisy outliers. In contrast, we hypothesize that focusing on addressing computationally-identifiable errors will better improve fairness for the unobserved groups.

Adversarially Reweighted Learning: With this hypothesis, we propose Adversarially Reweighted Learning (ARL), an optimization approach that leverages the notion of computationally-identifiable errors through an adversary \(f_\phi(X, Y)\) to improve worst-case performance over unobserved protected groups \(S\). Our experimental results show that ARL achieves high AUC for worst-case protected groups, high overall AUC, and robustness against training data biases.

Taken together, we make the following contributions:

- **Fairness without Demographics**: In Section 3, we propose adversarially reweighted learning (ARL), a modeling approach that aims to improve the utility for worst-off protected groups, without access to protected features at training or inference time. Our key insight is that when improving model performance for worst-case groups, it is valuable to focus the objective on computationally-identifiable regions of errors.

- **Empirical Benefits**: In Section 4, we evaluate ARL on three real-world datasets. Our results show that ARL yields significant AUC improvements for worst-case protected groups, outperforming state-of-the-art alternatives on all the datasets, and even improves the overall AUC on two of three datasets.

- **Understanding ARL**: In Section 5 we do a thorough experimental analysis and present insights into the inner-workings of ARL by analyzing the learnt example weights. In addition, we perform a synthetic study to We observe that ARL is quite robust to representation bias, and differences in group base-rate. However, similar to prior approaches, ARL degrades with noisy ground-truth labels.

2 RELATED WORK

Fairness: There has been an increasing line of work to address fairness concerns in machine learning models. A number of fairness notions have been proposed. At a high level they can be grouped into three categories, including (i) individual fairness [10, 33, 34, 46, 52], (ii) group fairness [12, 16, 27, 28, 51] that expects parity of statistical performance across groups, and (iii) fairness notions that aim to improve per-group performance, such as Pareto-fairness [4] and Rawlsian Max-Min fairness [17, 39, 43, 54]. In this work we follow the third notion of improving per-group performance.

There is also a large body of work on incorporating these fairness notions into ML models, including learning better representations [52] and adding fairness constraints in the learning objective [51], through post-processing the decisions [16] and through adversarial learning [7, 38, 53]. These works generally assume the protected attribute information is known and thus the fairness metrics can be directly optimized. However, in many real world applications the protected attribute information might be missing or is very sparse.

**Fairness without demographics**: Some works address this issue approximately by using proxy features [14] or assuming that the attribute is slightly perturbed [3]. However, using proxies can in itself be prone to bias [25]. Further, proxy information might be hard to obtain for many applications.

An interesting line of recent work [21, 23, 30, 45] tackles this problem by relying on trusted third parties that selectively collect and store protected-data necessary for incorporating fairness constraints. They generally assume that the ML model has access to the protected-features, albeit in encrypted form via secure multiparty computation (MPC) [21, 30], or in a privacy-preserving form by employing differentially private learning [23]. Another closely related work is agnostic federated learning [39], wherein given training data over \(K\) clients with unknown sampling distributions, the model aims to learn worst-case mixture coefficient weights that optimize for a worst-case target distribution over these \(K\) clients.

As mentioned earlier, the work closest to ours is DRO [17], which uses techniques from distributionally robust optimization to achieve Rawlsian Max-Min fairness without access to demographics. However, a key difference between DRO and ARL is the type of groups identified by them: DRO considers any worst-case distribution exceeding a given size \(\alpha\) as a potential protected group. Concretely, given a lower bound on size of the smallest protected group, say \(\alpha\), DRO optimizes for improving the worst-case performance of any set of examples exceeding size \(\alpha\). In contrast, our work relies on a notion of computational-identifiability.

**Computational-Identifiability**: Related to our algorithm, a number of works [19, 29, 31, 32] address intersectional fairness concerns by optimizing for group fairness between all computationally-identifiable groups in the input space. While the perspective of learning over computationally identifiable groups is similar, they differ from us in that they assume the protected group features are available in their input space, and that they aim to minimize the gap in accuracy or calibration across groups via regularization.

**Modeling Technique Inspirations**: In terms of technical machinery, our proposed ARL approach draws inspiration from a wide variety of prior modeling techniques. Re-weighting [22, 24, 37] is a popular paradigm typically used to address problems such as class imbalance by up-weighting examples from minority class. Adversarial learning [2, 13, 40, 47] is typically used to train a model to be robust with respect to adversarial examples. Focal loss [35] encourages the learning algorithm to focus on more difficult examples by up-weighting examples proportionate to their losses. Domain adaptation work requires a model to be robust and generalizable across different domains, under either covariate shift [44, 49] or label shift [36].
We now dive into the precise problem formulation and our proposed modeling approach.

3.1 Problem Formulation

In this paper we consider a binary classification setup (though the approach can be generalized to other settings). We are given a training dataset consisting of \( n \) individuals \( D = \{(x_i, y_i)\}_{i=1}^n \) where \( x_i \sim X \) is an \( m \) dimensional input vector of non-protected features, and \( y_i \sim Y \) represents a binary class label. We assume there exist \( K \) protected groups where for each example \( x_i \) there exists an unobserved \( s_i \sim S \) where \( S \) is a random variable over \( \{1, \ldots, K\} \). The set of examples with membership in group \( s \) given by \( D_s := \{(x_i, y_i) : s_i = s\}_{i=1}^n \). Again, we do not observe distinct set \( D_s \) but include the notation for formulation of the problem. To be more precise, we assume that protected groups \( S \) are unobserved attributes not available at training or inference times. However, we will frame our definition and evaluation of fairness in terms of protected groups. A summary of the notation used in this paper is given in Tbl. 1.

**Problem Definition** Given dataset \( D \in X \times Y \), but no observed protected group memberships \( S \), e.g., race or gender, learn a model \( h_\theta : X \to Y \) that is fair to groups in \( S \).

A natural next question is: what is a “fair” model? As in DRO [17], we follow the Rawlsian Max Min fairness principle of distributive justice [43]: we aim to maximize the minimum utility \( U \) a model has across all groups \( s \in S \) as given by Definition 1. Here, we assume that when a model predicts an example correctly, it increases utility for that example. Such \( U \) can be considered any one of standard accuracy metrics in machine learning that models are designed to optimize for.

**Definition 1 (Rawlsian Max-Min Fairness).** Suppose \( H \) is a set of hypotheses, and \( U_{D_s}(h) \) is the expected utility of the hypothesis \( h \) for the individuals in group \( s \), then a hypothesis \( h^* \) is said to satisfy Rawlsian Max-Min fairness principle [43] if it maximizes the utility of the worst-off group, i.e., the group with the lowest utility.

\[
\hat{h}^* = \arg\max_{h \in H} \min_{s \in S} U_{D_s}(h) \tag{1}
\]

In our evaluation in Section 4, we use AUC as a utility metric, and report the minimum utility over protected groups \( S \) as AUC(min).

3.2 Adversarial Reweighted Learning

Given this fairness definition and goal, how do we achieve it? As with traditional machine learning, accuracy/metrics are not differentiable, and instead convex loss functions are used. The traditional ML task is to learn a model \( h \) that minimizes the loss over the training data \( D \):

\[
h_{\text{avg}}^* = \arg\min_{h \in H} L_D(h) \tag{2}
\]

where \( L_D(h) = \mathbb{E}_{(x_i, y_i) \sim D}[\ell(h(x_i), y_i)] \) for some loss function \( \ell(\cdot) \) (e.g., cross entropy).

Therefore, we take the same perspective in turning Rawlsian Max-Min Fairness as given in Eq. (1) into a learning objective. Replacing the expected utility with an appropriate loss function \( L_{D_s}(h) \) over the set of individuals in group \( s \), we can formulate our fairness objective as:

\[
h^*_s = \arg\min_{h \in H} \max_{s \in S} L_{D_s}(h) \tag{3}
\]

where \( L_{D_s}(h) = \mathbb{E}_{(x_i, y_i) \sim D_s}[\ell(h(x_i), y_i)] \) is the expected loss for the individuals in group \( s \).

**Minimax Problem:** Similar to Agnostic Federal Learning (AFL) [39], we can formulate the Rawlsian Max-Min Fairness objective function in Eq. (3) as a zero-sum game between two players \( \theta \) and \( \lambda \). The optimization comprises of \( T \) game rounds. In round \( t \), player \( \theta \) learns the best parameters \( \theta \) that minimizes the expected loss. In round \( t + 1 \), player \( \lambda \) learns an assignment of weights \( \lambda \) that maximizes the weighted loss.

\[
J(\theta, \lambda) := \min_{\theta} \max_{\lambda} \mathbb{E}_{(x_i, y_i)} \ell(h(x_i), y_i) = \min_{\theta} \sum_{s \in S} \lambda_s L_{D_s}(h) \tag{4}
\]

To derive a concrete algorithm we need to specify how the players pick \( \theta \) and \( \lambda \). For the \( \theta \) player, one can use any iterative learning algorithm for classification tasks. For \( \lambda \), if the group memberships were known, the optimization problem in Eq. 4 can be solved by projecting \( \theta \) on a probability simplex over \( S \) groups given by \( \lambda = \{0, 1\}^K : ||\lambda|| = 1 \) as in AFL [39]. Unfortunately, for us, because we do not observe \( S \) we cannot directly optimize this objective as in AFL [39].

DRO [17] deals with this by effectively setting weights \( \lambda_i \) based on \( \ell(h(x_i), y_i) \) to focus on the largest errors. Instead, we will leverage the concept of computationally-identifiable subgroups [19]. Given a family of binary functions \( F \), we say that a group \( S \) is computationally-identifiable if there is a function \( f : X \times Y \to \{0, 1\} \in F \) such that \( f(x, y) = 1 \) if and only if \( (x, y) \in S \).

Building on this definition, we define \( f_\phi : X \times Y \to \{0, 1\} \) to be an adversarial neural network parameterized by \( \phi \) whose task, implicitly, is to identify regions where the learner makes significant errors \( Z := \{(x, y) : \ell(h(x), y) \geq \epsilon\} \). The adversarial examples weights \( \lambda_\phi : f_\phi \to \mathbb{R} \) can then be defined by appropriately rescaling \( f_\phi \) to put a high weight on regions with a high likelihood of errors, forcing the hypothesis \( h_\theta \) to improve in these regions. Rather than explicitly enforce a binary set of weights, as would be implied by the original definition of computational identifiability, our adversary uses a sigmoid activation to map \( f_\phi(x, y) \) to \( [0, 1] \). While this does not explicitly enforce a binary set of weights, we empirically observe that the rescaled weights \( \lambda_\phi(x, y) \) results in the weights clustering in two distinct regions as we see in Fig. 4 (with low weights near 1 and high weights near 4).

**ARL Objective:** We formalize this intuition, and propose an Adversarially Reweighted Learning approach, called ARL, which considers a minimax game between a learner and adversary. Both learner and adversary are learnt models, trained alternatively. The learner optimizes for the main classification task, and aims to learn the best parameters \( \theta \) that minimizes expected loss. The adversary learns a function mapping \( f_\phi : X \times Y \to \{0, 1\} \) to computationally-identifiable regions with high loss, and makes an adversarial assignment of weight vector \( \lambda_\phi : f_\phi \to \mathbb{R} \) so as to maximize the expected loss. The learner then adjusts itself to minimize the adversarial loss.
### Table 1: A Summary of Notation

| Notation | Definition |
|----------|------------|
| \(x_i \sim X\) | \(m\) dimensional input vector of non-protected features |
| \(y_i \sim Y\) | Binary class label for the prediction task |
| \(D = \{(x_i, y_i)\}_{i=1}^{n}\) | Training dataset consisting of \(n\) individuals |
| \(K\) | Number of protected groups |
| \(s_i \sim S\) | Random variable over \(K\) protected group representing protected group membership |
| \(D_s := \{(x_i, y_i) : s_i = s\}_{i=1}^{n}\) | Subset of training examples with membership in protected group \(s\) |
| \(h_\theta : X \rightarrow Y\) | Learner parameterized by \(\theta\) |
| \(f_\phi : X \times Y \rightarrow [0, 1]\) | Adversary parameterized by \(\phi\) |
| \(\lambda_\phi : f_\phi \rightarrow \mathbb{R}\) | Adversarial example weights defined by rescaling \(f_\phi\) |

\[
J(\theta, \phi) = \min_{\theta} \max_{\phi} \sum_{i=1}^{n} \lambda_\phi(x_i, y_i) \cdot \ell_{ce}(h_\theta(x_i), y_i) \tag{5}
\]

If the adversary was perfect it would adversarially assign all the weight \((\lambda)\) on the computationally-identifiable regions where learner makes significant errors, and thus improve learner performance in such regions. It is worth highlighting that, the design and complexity of the adversary model \(f_\phi\) plays an important role in controlling the granularity of computationally-identifiable regions of error. More expressive \(f_\phi\) leads to finer-grained upweighting but runs the risk of overfitting to outliers. While any differentiable model can be used for \(f_\phi\), we observed that for the small academic datasets used in our experiments, a linear adversary performed the best (further implementation details follow).

Observe that without any constraints on \(\lambda\) the objective in Eq. 5 is ill-defined. There is no finite \(\lambda\) that maximizes the loss, as an even higher loss could be achieved by scaling up \(\lambda\). Thus, it is crucial that we constrain the values \(\lambda\). In addition, it is necessary that \(\lambda_i \geq 0\) for all \(i\), since minimizing the negative loss can result in unstable behaviour. Further, we do not want \(\lambda_i\) to fall to 0 for any examples, so that all examples can contribute to the training loss. Finally, to prevent exploding gradients, it is important that the weights are normalized across the dataset (or current batch). In principle, our optimization problem is general enough to accommodate a wide variety of constraints. In this work we perform a normalization step that rescales the adversary \(f_\phi(x, y)\) to produce the weights \(\lambda_\phi\). We center the output of \(f_\phi\) and add 1 to ensure that all training examples contribute to the loss.

\[
\lambda_\phi(x_i, y_i) = 1 + n \cdot \frac{f_\phi(x_i, y_i)}{\sum_{i=1}^{n} f_\phi(x_i, y_i)}
\]

**Implementation:** In the experiments presented in Section 4, we use a standard feed-forward network to implement both learner and adversary. Our model for the learner is a fully connected two layer feed-forward network with 64 and 32 hidden units in the hidden layers, with ReLu activation function. While our adversary is general enough to be a deep network, we observed that for the small academic datasets used in our experiments, a linear adversary performed the best. Fig. 2 summarizes the computational graph of our proposed ARL approach.

4 EXPERIMENTAL RESULTS

We now demonstrate the effectiveness of our proposed adversarially re-weighted learning ARL approach through experiments over three real datasets well used in the fairness literature:

- **Adult:** The UCI Adult dataset [42] contains US census income survey records. We use the binarized “income” feature as the target variable for our classification task to predict if an individual’s income is above 50k.
- **LSAC:** The Law School dataset [48] from the law school admissions council’s national longitudinal bar passage study to predict whether a candidate would pass the bar exam. It consists of law school admission records. We use the binary feature “isPassBar” as the raget variable for classification.
- **COMPAS:** The COMPAS dataset [1] for recidivism prediction consists of criminal records comprising offender’s criminal history, demographic features (sex, race). We use the ground truth on whether the offender was re-arrested (binary) as the target variable for classification.

Key characteristics of the datasets, including a list of all the protected groups are in Tbl. 2. We transform all categorical attributes using one-hot encoding, and standardize all features vectors to have zero mean and unit variance. Python scripts for preprocessing the public datasets are open accessible along with the rest of the code of this paper.

**Evaluation Metrics:** We choose AUC (area under the ROC curve) as our utility metric as it is robust to class imbalance, i.e., unlike Accuracy it is not easy to receive high performance for trivial predictions. Further, it encompasses both FPR and FNR, and is threshold agnostic.
To evaluate fairness we stratify the test data by groups, compute AUC per protected group $s \in S$, and report:

- AUC(min): minimum AUC over all protected groups,
- AUC(macro-avg): macro-average over all protected group AUCs
- AUC(minority): AUC reported for the smallest protected group in the dataset.

For all metrics higher values are better. Note that the protected features are removed from the dataset, and are not used for training, validation or testing. The protected features are only used to compute subgroup AUC in order to evaluate fairness.

### Baselines and other approaches:

Our main comparison is with fairness without demographics approaches, which aim to improve worst-case subgroup performance. To this end, we compare ARL with a standard group-agnostic Baseline which performs standard ERM with uniform weights, and DRO [17], which is the current state-of-the-art, and summarize the results in subsection 4.1.

In subsection 4.2, we illustrate the strengths of ARL over standard re-weighting approaches like inverse probability weighting (IPW) [22] by comparing group-agnostic ARL with group-aware IPW.

Finally, while our fairness formulation is not the same as traditional group-fairness approaches, in order to better understand relationship between improving subgroup performance vs minimizing gap in error rates, we compare ARL with a group-fairness approach that aims to equalize false positive rates across groups. Results and key take-aways from this experiment are reported in Subsection 4.3.

Following is a summary of all the approaches.

- **Baseline Model:** a group-agnostic baseline, which performs standard empirical risk minimization (ERM) with uniform weights optimizing for the best overall performance.
- **DRO [17]:** a group-agnostic distributionally robust optimization approach that optimizes for worst-case subgroup.
- **IPW [22]:** A group-aware common re-weighting approach, which assigns weights to examples inverse proportionate to the probability of their observation in training data.
- **Min-Diff [41]:** A group-aware group-fairness approach that aims to minimize the difference between false positive rates across groups via constrained optimization.

### Setup and Parameter Tuning:

We use the same experimental setup, architecture, and hyper-parameter tuning for all the approaches. As our proposed ARL model has additional model capacity in the form of example weights $\lambda$, we increase the model capacity of the baselines by adding more hidden units in the intermediate layers of their DNN in order to ensure a fair comparison. Refer to Supplementary §7 for further details.

Best hyper-parameter values for all approaches are chosen via grid-search by performing 5-fold cross validation optimizing for best overall AUC. We do not use subgroup information for training or tuning. DRO has a separate fairness parameter $\eta$. For the sake of fair comparison, we report results for two variants of DRO: (i) DRO, with $\eta$ tuned as detailed in their paper and (ii) DRO (auc) with $\eta$ tuned to achieve best overall AUC performance. All results reported are averages across 10 independent runs (with different model parameter initialization).

### 4.1 Fairness without Demographics

Our main comparison is with DRO [17], a group-agnostic distributionally robust optimization approach that optimizes for the worst-case subgroup. Additionally, we report results for the vanilla group-agnostic Baseline, which performs standard ERM with uniform weights. Tbl. 3 summarizes the main results. Additional results with AUC (mean ± std) for all protected groups are reported in the Tbl. 8. Best values in each table are highlighted in bold. We make the following key observations:

#### Table 2: Description of datasets

| Dataset | Size | No. of features | Base-rate Protected features | Protected groups $(S)$ |
|---------|------|-----------------|-------------------------------|-------------------------|
| Adult   | 40791| 15              | 0.23 Race, Sex {White, Black} × {Male, Female} |                          |
| LSAC    | 27479| 12              | 0.80 Race, Sex {White, Black} × {Male, Female} |                          |
| COMPAS  | 7215 | 11              | 0.47 Race, Sex {White, Black} × {Male, Female} |                          |

#### Table 3: Main results: ARL vs DRO

| dataset | method | AUC (mean) | AUC (median) | AUC (min) | AUC (macro-avg) | AUC (minority) |
|---------|--------|------------|--------------|-----------|----------------|----------------|
| Adult   | Baseline | 0.898      | 0.891        | 0.867     | 0.875          |                |
| Adult   | DRO    | 0.874      | 0.862        | 0.843     | 0.891          |                |
| Adult   | DRO (auc) | 0.899      | 0.908        | 0.869     | 0.933          |                |
| Adult   | ARL    | 0.907      | 0.915        | 0.881     | 0.942          |                |
| LSAC    | Baseline | 0.813      | 0.813        | 0.790     | 0.824          |                |
| LSAC    | DRO    | 0.662      | 0.656        | 0.638     | 0.677          |                |
| LSAC    | DRO (auc) | 0.709      | 0.710        | 0.683     | 0.729          |                |
| LSAC    | ARL    | 0.823      | 0.820        | 0.798     | 0.832          |                |
| COMPAS  | Baseline | 0.748      | 0.730        | 0.674     | 0.774          |                |
| COMPAS  | DRO    | 0.619      | 0.601        | 0.572     | 0.593          |                |
| COMPAS  | DRO (auc) | 0.699      | 0.678        | 0.616     | 0.704          |                |
| COMPAS  | ARL    | 0.743      | 0.727        | 0.658     | 0.785          |                |

**ARL improves worst-case performance:** ARL significantly outperforms DRO, and achieves best results for AUC (minority) for all datasets. We observe a 6.5 percentage point (pp) improvement over the baseline for Adult, 0.8 pp for LSAC, and 1.1 pp for COMPAS. Similarly, ARL shows 2 pp and 1 pp improvement in AUC (min) over baseline for Adult and LSAC datasets respectively. For COMPAS dataset there is no significant difference in performance over baseline, yet significantly better than DRO, which suffers a lot.

These results are inline with our observations on computational-identifiability of protected groups in Tbl. 6 (§5). As we will later see, unlike Adult and LSAC datasets, protected-groups in COMPAS dataset are not computationally-identifiable. Hence, ARL shows no gain or loss. In contrast, we believe DRO is picking on noisy outlier in the dataset as high loss example, hence the significant drop in its performance. This result clearly highlights the merit of optimizing for distributional robustness over computationally-identifiable groups as in ARL, as opposed to any worst-case distribution as in DRO.

**ARL improves overall AUC:** Further, in contrast to the general expectation in fairness approaches, wherein utility-fairness trade-off is implicitly assumed, we observe that for Adult and LSAC datasets ARL in fact shows ~ 1 pp improvement in AUC (avg) and AUC (macro-avg). This is because ARL’s optimization objective of minimizing maximal loss is better aligned with improving overall AUC.
4.2 ARL vs Inverse Probability Weighting

Next, to better understand and illustrate the advantages of ARL over standard re-weighting approaches, we compare ARL with inverse probability weighting (IPW)[22], which is the most common re-weighting choice used to address representational disparity problems. Specifically, IPW performs a weighted ERM with example weights set as $1/p(s)$ where $p(s)$ is the probability of observing an individual from group $s$ in the empirical training distribution. In addition to vanilla IPW, we also report results for a IPW variant with inverse probabilities computed jointly over protected-features $s$ and class-label $Y$ reported as IPW(S+Y). Tbl. 4 summarizes the results. We make following observations and key takeaways:

**Table 4: ARL vs Inverse Probability Weight**

| dataset | method  | AUC  | AUC  | AUC  | AUC  |
|---------|---------|------|------|------|------|
|         |         | macro-avg | min | minority |
| Adult   | IPW(S)  | 0.897 | 0.872 | 0.876 | 0.883 |
| Adult   | IPW(S+Y)| 0.897 | 0.909 | 0.877 | 0.932 |
| Adult   | ARL     | 0.907 | 0.915 | 0.881 | 0.942 |
| LSAC    | IPW(S)  | 0.794 | 0.789 | 0.772 | 0.775 |
| LSAC    | IPW(S+Y)| 0.799 | 0.798 | 0.784 | 0.785 |
| LSAC    | ARL     | 0.823 | 0.820 | 0.798 | 0.832 |
| COMPAS  | IPW(S)  | 0.744 | 0.727 | 0.679 | 0.759 |
| COMPAS  | IPW(S+Y)| 0.727 | 0.724 | 0.678 | 0.764 |
| COMPAS  | ARL     | 0.743 | 0.727 | 0.658 | 0.785 |

Firstly, observe that in spite of not having access to demographic features, ARL has comparable if not better results than both variants of the IPW on all datasets. This results shows that even in the absence of group labels, ARL is able to appropriately assign adversarial weights to improve errors for protected-groups.

Further, not only does ARL improve subgroup fairness, in most settings it even outperforms IPW, which has perfect knowledge of group membership. This result further highlights the strength of ARL. We observed that this is because unlike IPW, ARL does not equally upweight all examples from protected groups, but does so only if the model needs much more capacity to be classified correctly. We present evidence of this observation in Section 5.3

4.3 ARL vs Group-Fairness Approaches

While our fairness formulation is not the same as traditional group-fairness approaches, in order to better understand relationship between improving subgroup performance vs minimizing gap, we compare our group-agnostic ARL with a group-aware group-fairness approach that aims to achieve equal opportunity (EqOpp) [16], i.e., equalize false positive rates (FPR) across groups. Amongst many EqOpp approaches [6, 16, 50], we choose Min-Diff [6] as a comparison as it is the closest to ARL in terms of implementation and optimization. To ensure fair comparison we instantiate Min-Diff with similar neural architecture and model capacity as ARL. Further, as we are interested in performance for multiple protected groups, we add one Min-Diff loss term for each protected feature (sex and race). Details of the implementation are described in Supplementary §7. Tbl. 5 summarizes these results. We make the following observations:

**Table 5: ARL vs Group-Fairness**

| dataset | method  | AUC  | AUC  | AUC  | AUC  |
|---------|---------|------|------|------|------|
|         |         | macro-avg | min | minority |
| Adult   | MinDiff | 0.847 | 0.856 | 0.835 | 0.863 |
| Adult   | ARL     | 0.907 | 0.915 | 0.881 | 0.942 |
| LSAC    | MinDiff | 0.826 | 0.825 | 0.805 | 0.840 |
| LSAC    | ARL     | 0.823 | 0.820 | 0.798 | 0.832 |
| COMPAS  | MinDiff | 0.730 | 0.712 | 0.645 | 0.748 |
| COMPAS  | ARL     | 0.743 | 0.727 | 0.658 | 0.785 |

Min-Diff improves gap but not worst-off group: True to its goal, Min-Diff decreases the FPR gap between groups: FPR gap on sex is between 0.02 and 0.05, and FPR gap on race is between 0.1 and 0.19 for all datasets. However, lower-gap between groups doesn’t always lead to improved AUC for worst-off groups (observe AUC min and AUC minority). ARL significantly outperforms Min-Diff for Adult and COMPAS datasets, and achieves comparable performance on LSAC dataset. This is especially remarkable given that Min-diff approach has explicit access to protected group information.

This result highlights the intrinsic mismatch between fairness goals of group-fairness approaches vs the desire to improve performance for protected groups. We believe making models more inclusive by improving the performance for groups, not just decreasing the gap, is an important complimentary direction for fairness research.

Utility-Fairness Trade-off: Further, observe that Min-diff incurs a 5 pp drop in overall AUC for Adult dataset, and 2 pp drop for COMPAS dataset. In contrast, as noted earlier ARL in-fact shows an improvement in overall AUC for Adult and LSAC datasets. This result shows that unlike Min-Diff (or group fairness approaches in general) where there is an explicit utility-fairness trade-off, ARL achieves a better pareto allocation of overall and subgroup AUC performance. This is because the goal of ARL, which explicitly strives to improve the performance for protected groups is aligned with achieving better overall utility.

5 ANALYSIS

Next, we conduct analysis to gain insights into ARL.

5.1 Are groups computationally-identifiable?

We first test our hypothesis that unobserved protected groups $S$ are correlated with observed features $X$ and class label $Y$. Thus, even when they are unobserved, they can be computationally-identifiable. We test this hypothesis by training a predictive model to infer $S$ given $X$ and $Y$. Tbl. 6 reports the predictive accuracy of a linear model.

**Table 6: Identifying groups**

|          | Adult | LSAC | COMPAS |
|----------|-------|------|--------|
| Race     | 0.90  | 0.94 | 0.61   |
| Sex      | 0.84  | 0.58 | 0.78   |

We observe that Adult and LSAC datasets have significant correlations with unobserved protected groups, which can be adversarially exploited to computationally-identify protected-groups.
In contrast, for COMPAS dataset the protected-groups are not computationally-identifiable. As we saw earlier in Tbl. 3 (§4) these results align with ARL showing no gain or loss for COMPAS dataset, but improvements for Adult and LSAC.

5.2 Robustness to training distributions

In this experiment, we investigate robustness of ARL and DRO approaches to training data biases [8], such as bias in group sizes (representation-bias) and bias due to noisy or incorrect ground-truth labels (label-bias). We use the Adult dataset and generate several semi-synthetic training sets with worst-case distributions (e.g., few training examples of “female” group) by sampling points from original training set. We then train our approaches on these worst-case training sets, and evaluate their performance on a fixed untainted original test set.

Concretely, to replicate representation-bias, we vary the fraction of female examples in training set by under/over-sampling female examples from training set. Similarly, to replicate label-bias, we vary fraction of incorrect labels by flipping ground-truth class labels uniformly at random for a fraction of training examples. In all experiments, training set size remains fixed. To mitigate the randomness in data sampling and optimization processes, we repeat the process 10 times and report results on a fixed untainted original test set (e.g., without adding label noise). Fig. 3 reports the results. In the interest of space, we limit ourselves to the protected-group “Female”. For each training setting shown on X-axis, we report the corresponding AUC for Female subgroup on Y-axis. The vertical bars in the plot are confidence intervals over 10 runs. We make the following observations:

**Representation Bias**: Both DRO and ARL are robust to the representation bias. ARL clearly outperforms DRO and baseline at all points. Surprisingly, we see a drop in AUC for baseline as the group-size increases. This is an artifact of having fixed training data size. As the fraction of female examples increases, we are forced to oversample female examples and downsample male examples; this leads to a decrease in the information present in training data and in turn leads to a worse performing model. In contrast, ARL and DRO cope better with this loss of information.

**Label Bias**: Both ARL and DRO are quite sensitive to label bias, much more than baseline. This is however expected, as both approaches aim to up-weight examples with prediction error. However, they cannot distinguish between true and noisy labels, which leads to performance degradation. This result highlights that distributionally robust optimization techniques like ARL and DRO should be used with caution for datasets wherein ground-truth labels may not be trustworthy.

5.3 Are learnt example weights meaningful?

Next, we investigate if the example weights learnt by ARL are meaningful through the lens of training examples in the Adult dataset. Fig. 4 visualizes the example weights assigned by ARL stratified into four quadrants of a confusion matrix. Each subplot visualizes the learnt weights λ on x-axis and their corresponding density on y-axis. We make following observations:

- **Misclassified examples are upweighted**: As expected, misclassified examples are upweighted (in Fig. 4b and 4c), whereas correctly classified examples are not upweighted (in Fig. 4a). Further, we observe that even though this was not our original goal, as an interesting side-effect ARL has also learnt to address class imbalance problem in the dataset. Recall that our Adult dataset has class imbalance, and only 23% of examples belong to class 1. Observe that, in spite of making no errors ARL assigns high weights to all class 1 examples as shown in Fig. 4d (unlike in Fig. 4a where all class 0 example have weight 1).
ARL adjusts weights to base-rate. To investigate this further, we smoothly vary the base-rate of female group in training data (i.e., we synthetically control fraction of female examples with class label 1 in training data). Fig. 5 visualizes training data base-rate on x-axis and mean example weight learnt for the subgroup on y-axis. Observe that at female base-rate 0.1, i.e., when only 10% of female training examples belong to class 1, the mean weight assigned for examples in class 1 is significantly higher than class 0. As base-rate increases, i.e., as the number of class 1 examples increases, ARL correctly learns to decrease the weights for class 1 examples, and increases the weights for class 0 examples. These insights further explain the reason why ARL manages to improve overall AUC.

5.4 Significance of inputs to the Adversary

Our proposed adversarially reweighted learning ARL approach is flexible and generalizes to many related works by varying the inputs to the adversary. For instance, if the domain of our adversary \( f_S(.) \) was, i.e., it took only protected features \( S \) as input, the sets \( Z \) computationally-identifiable by the adversary boil down to an exhaustive cross over all the protected features \( S \), i.e., \( Z \subseteq \mathbb{Z}^2 \). Thus, our ARL objective in Eq. 5 would reduce to being very similar to the objective of fair agnostic federated learning by Mohri et al. [39]: to minimize the loss for the worst-off group amongst all known intersectional subgroups.

In this experiment, we further gain insights into our proposed adversarially re-weighting approach by comparing a number of variants ARL:

- ARL (adv: X+Y): vanilla ARL where the adversary takes non-protected features \( X \) and class label \( Y \) as input.
- ARL (adv: S): variant of ARL where the adversary takes only protected features \( S \) as input.
- ARL (adv: S+Y): variant of ARL with access to protected features \( S \) and class label \( Y \) as input.
- ARL(adv: X+Y+S): variant of ARL where the adversary takes all features \( X + S \) and class label \( Y \) as input.

A summary of results is reported in Tbl. 7. We make the following observations:

**Group-agnostic ARL is competitive:** Firstly, observe that contrary to general expectation our vanilla ARL without access to protected groups \( S \), i.e., ARL (adv: X+Y) is competitive, and its results are comparable with ARL variants with access to protected-groups \( S \) (except in the case of COMPAS dataset as observed earlier). These results highlight the strength of ARL as an approach achieve fairness without access to demographics.

**Access to class label \( Y \) is crucial:** Further, we observe that variants with class label \( Y \) generally outperform variants without class label. For instance, for ARL(S+Y) has higher AUC than ARL(S) for all groups across all datasets. Especially for Adult and LSAC datasets, which are known to have class imbalance problem (observe base-rate in Tbl.2). A similar trend was observed for IPW(S) vs IPW(S+Y) in Tbl.4 (§4). This is expected and can be explained as follows: variants without access to class label \( Y \) such as ARL(S) are forced to give the same weight to both positive and negative examples of a group. As a consequence, they do not cope well with differences in base-rates, especially across groups, as they cannot treat majority and minority class differently.

**Blind Fairness:** Finally, in this work, we operated under the assumption that protected features are not available in the dataset. However, in practice there are scenarios where protected features \( S \) are available in the dataset, however, we are blind to them. More concretely, we do not know a priori which subset of features amongst all features \( X + S \) might be candidates for protected groups \( S \). Examples of this setting include scenarios wherein a number of demographics features (e.g., age, race, sex) are present in the dataset. However, we do not known which subgroup(s) amongst all intersectional groups (given by the cross-product over demographic features) might need potential fairness treatment.

Our proposed ARL approach naturally generalizes to this setting as well. We observe that the performance of our ARL variant ARL(adv: X+Y+S) is comparable to the performance of ARL(adv: Y+S). In certain cases (e.g., Adult dataset), access to remaining features \( X \) even improves fairness. We believe this is because access to \( X \) helps the adversary to make fine-grained distinctions amongst a subset of disadvantaged candidates in a given group \( s \in S \) that need fairness treatment.

Table 7: A comparison of variants of ARL

| dataset  | method          | AUC   | AUC                  | AUC                  | AUC              |
|----------|-----------------|-------|----------------------|----------------------|------------------|
|          |                 |       | macro-avg            | min                  | minority         |
| Adult    | Baseline        | 0.698 | 0.891                | 0.867                | 0.875            |
| Adult    | ARL (adv: S)    | 0.990 | 0.894                | 0.875                | 0.879            |
| Adult    | ARL (adv: S+Y)  | 0.907 | 0.907                | 0.882                | 0.907            |
| Adult    | ARL (adv: X+Y+S)| 0.907 | 0.911                | 0.881                | 0.932            |
| Adult    | ARL (adv: X+Y)  | 0.907 | 0.915                | 0.881                | 0.942            |
| LSAC     | Baseline        | 0.813 | 0.813                | 0.790                | 0.824            |
| LSAC     | ARL (adv: S)    | 0.820 | 0.823                | 0.799                | 0.846            |
| LSAC     | ARL (adv: S+Y)  | 0.824 | 0.826                | 0.801                | 0.845            |
| LSAC     | ARL (adv: X+Y+S)| 0.826 | 0.825                | 0.808                | 0.838            |
| LSAC     | ARL (adv: X+Y)  | 0.823 | 0.820                | 0.798                | 0.832            |
| COMPAS   | Baseline        | 0.748 | 0.730                | 0.674                | 0.774            |
| COMPAS   | ARL (adv: S)    | 0.747 | 0.729                | 0.675                | 0.768            |
| COMPAS   | ARL (adv: S+Y)  | 0.747 | 0.731                | 0.681                | 0.771            |
| COMPAS   | ARL (adv: X+Y+S)| 0.747 | 0.731                | 0.673                | 0.778            |
| COMPAS   | ARL (adv: X+Y)  | 0.743 | 0.727                | 0.658                | 0.785            |

6 CONCLUSION

Improving model fairness without directly observing protected features is a difficult and under-studied challenge for putting machine learning fairness goals into practice. The limited prior work has focused on improving model performance for any worst-case distribution, but as we show this is particularly vulnerable to noisy outliers. Our key insight is that when improving model performance for worst-case groups, it is valuable to focus the objective on computationally-identifiable regions of errors i.e., regions of the input and label space with significant errors.

In practice, we find that our proposed group-agnostic Adversarially Reweighted Learning (ARL) approach yields significant improvement in AUC for worst-case protected groups, outperforming state-of-the-art alternatives across multiple dataset, and is robust to multiple types of training data biases. As a result, we believe this insight and the ARL method provides a foundation for how to pursue fairness without access to demographics.

Acknowledgements: The authors would like to thank Jay Yagmik for encouragement along this research direction and insightful
connections that contributed to this work. We thank Alexander D’Amour, Gerhard Weikum, Krishna P. Gummadi, and Yoni Halpern for helpful feedback that improved the manuscript.

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There’s software used across the country to predict future criminals and it’s biased against blacks. ProPublica (2016).

[2] K. Asif, W. Xing, S. Behpour, and B. D. Ziebart. 2015. Adversarial Cost-Sensitive Classification. In IJAI 92–101.

[3] P. Awasthi, M. Kleindessner, and J. Morgenstern. 2019. Equalized odds post-processing under imperfect group information. stat 1050.

[4] A. Balashankar, A. Leis, C. Welby, and L. Subramanian. 2019. Pareto-Efficient Fairness for Skewed Subgroup Data. In IJCAI AI for Social Good Workshop.

[5] S. Barocas and A. D. Selbst. 2016. Big data’s disparate impact. Cal. L. Rev. 2016 (2016).

[6] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Liu, P. Kreitmann, J. Bischof, and E. H. Chi. 2019. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In AIES ACM, 453–459.

[7] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In 2017 KDD Workshop on Fairness, Accountability, and Transparency in Machine Learning.

[8] A. Blum and K. Chang. 2019. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? arXiv preprint arXiv:1902.01994 (2019).

[9] J. Buolamwini and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. 77–91.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. Fairness through Awareness (ITCS ’12) 214–226.

[11] M. D. Ekstrand, M. Tian, Ion Madrazo Aspiazu, J. D. Ekstrand, O. Anuyah, D. Mckeen, and M. S. Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Conference on Accountability and Transparency in Machine Learning. 172–186.

[12] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatsubramanian. 2015. Certifying and Removing Disparate Impact. In ICML.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In Advances in neural information processing systems. 2672–2680.

[14] M. R. Gupta, A. Cotter, M. M. Fard, and S. Wang. 2018. Proxy Fairness. ArXiv abs/1806.11212 (2018).

[15] S. Hajan, F. Bosch, and C. Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In SIGKDD.

[16] M. Hardt, E. Price, and N. Srebro. 2016. Equality of Opportunity in Supervised Learning. In NIPS.

[17] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. 2018. Fairness Without Demographics in Repeated Lot-Loss Minimization. In IJCAI.

[18] R. Haasnoot-Wyna, D. W. Baker, D. Nerenz, J. Feinglass, Anne C Beal, M. B. Landrum, R. Behal, and J. S. Weissman. 2007. Disparities in health care are driven by where minority patients seek care: examination of the hospital quality alliance.

[19] A. P. Lahoti, K. P. Gummadi, and G. Weikum. 2019. Operationalizing Individual Fairness with Fair pairwise Representations. Proceedings of the VLDB Endowment 13, 4 (2019).

[20] M. Kim, O. Reingold, and G. Rothblum. 2018. Fairness through computationally-bounded awareness. In Advances in Neural Information Processing Systems. 4842–4852.

[21] M. P. Kim, A. Ghorbani, and J. Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 247–254.

[22] P. Lahoti, K. P. Gummadi, and G. Weikum. 2019. Liar: Learning individually fair data representations for algorithmic decision making. In ICDE.

[23] P. Lahoti, K. P. Gummadi, and G. Weikum. 2019. Operationalizing Individual Fairness with Fair pairwise Representations. Proceedings of the VLDB Endowment 13, 4 (2019).

[24] Z. Lipton, Y. Wang, and A. Smola. 2018. Detecting and Correcting for Label Shift with Black Box Predictors. (2018).

[25] Roderick J A Little and Donald B Rubin. 1986. Statistical Analysis with Missing Data. John Wiley & Sons, Inc., USA.

[26] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In ICML. 3381–3390.

[27] M. Mohri, G. Sivek, and A. T. Suresh. 2019. Agnostic Federated Learning. (2019).

[28] E. Montaehaei, M. Ghorbani, M. Soleymani Baghshah, and Hamid R Rabbani. 2018. Adversarial classifier for imbalanced problems. arXiv preprint arXiv:1811.08812 (2018).

[29] F. Provost, H. Qian, C. Chen, E. E. Chi, J. Chen, and A. Beutel. 2019. Towards a better trade-off between performance and fairness with kernel-based distribution matching. arXiv preprint arXiv:1910.11779 (2019).

[30] B. Becker R Kohavi. 1996. UCI ML Repository. http://archive.ics.uci.edu/ml

[31] J. Rawls. 2001. Justice as fairness: A restatement. Harvard University Press.

[32] H. Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference (2000).

[33] M. Veale and R. Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society 4, 2 (2017). 2053551717743530.

[34] H. Wang, N. Grjige-Hlaca, P. Lahoti, K. P. Gummadi, and A. Weller. 2019. An Empirical Study on Learning Fairness Metrics for COMPAS Data with Human Supervision. arXiv preprint arXiv:1910.10255 (2019).

[35] J. Wang, T. Zhang, S. Liu, P. Chen, J. Xu, M. Faradon, and Bo Li [n.d.]. Towards A Unified Min-Max Framework for Adversarial Exploration and Robustness. (n.d.).

[36] L. F. Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).

[37] Bianca Zadrozny. 2004. Learning and Evaluating Classifiers under Sample Selection Bias (ICML ’04). 114.

[38] M.B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. 2017. From parity to preference-based notions of fairness in classification. In NIPS.

[39] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In WWW.

[40] R. S. Zemel, L. Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. Learning Fair Representations. In ICML.

[41] B. H. Zhang, B. Lemoine, and M. Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In AIES ACM, 335–340.

[42] C. Zhang and J. A. Shah. 2014. Fairness in multi-agent sequential decision-making. In Advances in Neural Information Processing Systems. 2636–2644.

7 SUPPLEMENTARY MATERIAL

Reproducibility: All the datasets used in this paper are publicly available. Python and Tensorflow implementations of all code required to reproduce the results reported in this paper is available at https://github.com/google-research/google-research/tree/master/group_agnostic_fairness.

Baselines and Implementation: We compare our proposed approach ARL with the two naive baselines and one state-of-the-art approach. All the implementations are open accessible along with the rest of the code of this paper. All approaches have the same DNN
As we are interested in improving performance for multiple subgroups at a time, we add one Min-Diff loss terms for each protected attribute (sex and race).

Setup and Parameter Tuning: Each dataset is randomly split into 70% training and 30% test sets. On the training set, we perform a 5-fold cross validation to find the best hyper-parameters for each model (details follow). Once the hyperparameters are tuned, we use the second part as an independent test set to get an unbiased estimate of their performance. We use the same experimental setup, data split, and parameter tuning techniques for all the methods.

For each approach, we choose the best learning-rate, and batch size by performing a grid search over an exhaustive hyperparameter space given by batch size (32, 64, 128, 256, 512) and learning rate (0.001, 0.01, 0.1, 1, 2, 5). All the parameters are chosen via 5-fold cross validation by optimizing for best overall AUC.

In addition to batch size, and learning rate, DRO approach [17] has an additional fairness hyper-parameter $\eta$, which controls the performance for the worst-case subgroup. In their paper, the authors present a specific hyperparameter tuning approach to choose the best value for $\eta$. Hence for the sake of fair comparison, we report results for two variants of DRO: (i) DRO, original approach with $\eta$ tuned as detailed in their paper and (ii) DRO(auc) with $\eta$ tuned to achieve best overall AUC performance.

Omitted Tables: In Section 4, we performed our main comparison with fairness without demographics approaches DRO [17], and group-agnostic Baseline, and present main results. Additional omitted results summarizing AUC (mean ± std) for all protected groups in each dataset are reported in Tbl. 8. Best values in each table are highlighted in bold.