https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 26, No. 4, July–August 2024, pp. 1567–1585
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Wasserstein Robust Classification with Fairness Constraints

Yijie Wang,[a,*] Viet Anh Nguyen,[b] Grani A. Hanasusanto[c]

[a] School of Economics and Management, Tongji University, Shanghai 200092, China; [b] Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong; [c] Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, Illinois 61801

*Corresponding author

Contact: yijiewang514@gmail.com, https://orcid.org/0000-0002-5705-892X (YW); nguyen@se.cuhk.edu.hk, https://orcid.org/0000-0002-9607-7891 (VAN); gah@illinois.edu, https://orcid.org/0000-0003-4900-2958 (GAH)

Received: May 15, 2022
Revised: January 19, 2024
Accepted: March 7, 2024
Published Online in Articles in Advance: April 30, 2024

https://doi.org/10.1287/msom.2022.0230

Abstract. ***Problem definition*:** Data analytics models and machine learning algorithms are increasingly deployed to support consequential decision-making processes, from deciding which applicants will receive job offers and loans to university enrollments and medical interventions. However, recent studies show these models may unintentionally amplify human bias and yield significant unfavorable decisions to specific groups. ***Methodology/results*:** We propose a distributionally robust classification model with a fairness constraint that encourages the classifier to be fair in the equality of opportunity criterion. We use a type-∞ Wasserstein ambiguity set centered at the empirical distribution to represent distributional uncertainty and derive a conservative reformulation for the worst-case equal opportunity unfairness measure. We show that the model is equivalent to a mixed binary conic optimization problem, which standard off-the-shelf solvers can solve. We propose a convex, hinge-loss-based model for large problem instances whose reformulation does not incur binary variables to improve scalability. Moreover, we also consider the distributionally robust learning problem with a generic ground transportation cost to hedge against the label and sensitive attribute uncertainties. We numerically examine the performance of our proposed models on five real-world data sets related to individual analysis. Compared with the state-of-the-art methods, our proposed approaches significantly improve fairness with negligible loss of predictive accuracy in the testing data set. ***Managerial implications*:** Our paper raises awareness that bias may arise when predictive models are used in service and operations. It generally comes from human bias, for example, imbalanced data collection or low sample sizes, and is further amplified by algorithms. Incorporating fairness constraints and the distributionally robust optimization (DRO) scheme is a powerful way to alleviate algorithmic biases.

Funding: This work was supported by the National Science Foundation [Grants 2342505 and 2343869] and the Chinese University of Hong Kong [Grant 4055191].
Supplemental Material: The online appendices are available at https://doi.org/10.1287/msom.2022.0230.

Keywords: math programming • stochastic methods

1. Introduction

High-quality individual analysis is recently attracting attention in operations management and data analytics because of the increasing availability of data (Mišić and Perakis 2020). To provide the target individuals with the most appropriate products, services, and offers, many companies, institutions, and governmental departments are deploying advanced data analytics models to analyze the characteristics of the users. For example, in loan audit (Bose and Mahapatra 2001), inductive learning systems and credit scoring models optimize the lending decisions based on the predicted default risk of the applicants (Shaw and Gentry 1988, Jacobson and Roszbach 2003). In retail, personalized strategies ranging from pricing (Chen et al. 2022), product offering (Baardman et al. 2023), to assortment planning (Golrezaei et al. 2014) are designed to meet the needs of different classes of customers. In hospital appointment scheduling, prediction models are deployed to identify patients with high nonshow probability to schedule them into or right after overbooked slots (Mak et al. 2014). In medical interventions, machine learning algorithms are trained to diagnose disease and provide treatment advice to doctors (Shipp et al. 2002, Obermeyer and Emanuel 2016). Furthermore, algorithmic recidivism scores in criminal justice support judges assessing defendants' future criminal risk (Monahan and Skeem 2016). Finally, in company recruitment (Lohr 2013, Dastin 2022) and university admissions (Chang 2006, Kabakchieva 2013), statistical learning models help reviewers screen out qualified candidates from a vast pool of applicants efficiently.

Data analytics models and algorithms can extract signals from large data sets to support consequential decision-making processes; however, they may not be entirely objective and can even amplify existing human biases. An exemplary case of algorithmic bias can be described by the well-known German data set containing the credit scores of 1,000 candidates with 20 demographical features such as age, deposit, and income (Dua and Graff 2017). Recent studies show that naively adopting plain vanilla prediction models to this data set yields remarkable biased outcomes for young people (Bellamy et al. 2018). The reason for these biased predictions is that the algorithms may identify age as a determining factor to the repayment and, thus, significantly prefer old candidates to the young. As the collected data sets may often not represent the true population across all groups, a plain vanilla prediction model trained on such data sets can unintentionally amplify the bias and yield highly unfavorable decisions for certain minority groups. Moreover, basing any algorithmic decision on sensitive attributes may be considered illegal if the learning algorithms are regulated by law. Thus, basing loan approval decisions on the age of the applicants may lead to possible lawsuits against discriminatory lending (Consumer Financial Protection Bureau 2013).

Similar algorithmic unfairness issues also arise in other service and operations management applications. For example, the hiring recommendation system of Amazon AI discriminated against female candidates for technical positions (Dastin 2022). Similarly, Google's personalized ad targeting algorithm recommended higher-paying executive jobs more often to male than female candidates (Datta et al. 2015). In healthcare, existing overbooking systems may unintentionally enlarge the correlation between races and no-show probabilities, resulting in significantly longer waiting times for patients of color (Samorani et al. 2022). In addition, the judicial unfairness brought by machine learning algorithms has also evoked widespread social concerns. An algorithm used by the U.S. justice system to predict future criminals is shown to be significantly biased against African Americans: It falsely flags black defendants as future criminals at almost twice the rate of white defendants (Angwin et al. 2022).

This paper focuses on the training phase of a linear classifier, arguably one of the most popular classification methods in the literature (Hastie et al. 2009). The classifier establishes a deterministic relationship between a feature vector $\boldsymbol{X} \in \mathcal{X} = \mathbb{R}^d$ and a binary response, or label, variable $Y \in \mathcal{Y} = \{-1, 1\}$. Without any loss of generality, we associate the positive response $Y = 1$ with the "desirable" outcome, such as "being hired" or "receiving a loan approval." In the linear setting, a classifier $\mathcal{C}: \mathcal{X} \to \mathcal{Y}$ is parameterized by a slope parameter $\boldsymbol{w} \in \mathbb{R}^d$ and an offset $b \in \mathbb{R}$, and the classification output is determined through an indicator function of the form

$$\mathcal{C}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{w}^\top \boldsymbol{x} + b \geq 0, \\ -1 & \text{if } \boldsymbol{w}^\top \boldsymbol{x} + b < 0. \end{cases}$$

In the context of classification, we need to find a classifier that maximizes the correct classification probability. To this end, we can consider the *correct classification probability* with respect to the distribution $\mathbb{Q}$ as

$$\mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) > 0).$$

Complementarily, the *misclassification probability* with respect to $\mathbb{Q}$ is defined as

$$\mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) \leq 0).$$

By definition, we consider that any $\boldsymbol{x}$ falling exactly on the hyperplane $\boldsymbol{w}^\top \boldsymbol{X} + b = 0$ is misclassified irrespective of the true label of $\boldsymbol{x}$. The optimal linear classifier can be defined as the solution to the *misclassification probability minimization problem:*

$$\min_{(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}} \mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) \leq 0). \tag{1}$$

Using the sample average approximation of the probability term in (1) and solving the resulting approximation problem, we can obtain an empirical classifier. Nevertheless, as previously discussed, this empirical classifier can be unfair because it may unjustifiably possess unequal predictive performances across different subgroups in the population.

To address the fairness concern, we assume that there is a single, binary sensitive attribute $A \in \mathcal{A} = \{0, 1\}$. In a real-world setting, this sensitive attribute can represent information such as the race, gender, or age of a person, and it distinguishes the privileged $A = 1$ from the unprivileged individuals $A = 0$. Hereby, we define the privileged group as the group for which the empirical classifier has higher predictive performance than for the *un*privileged group. Throughout this paper, we assume that we possess a training data set containing N samples of the form $\{(\hat{\boldsymbol{x}}_i, \hat{a}_i, \hat{y}_i)\}_{i=1}^N$, and these samples are generated independently from a single data-generating probability distribution. Moreover, we consider the privileged learning setting in which the sensitive information A is only available at the training stage but not at the testing stage (Vapnik and Vashist 2009, Quadrianto and Sharmanska 2017). It is therefore reasonable to consider only classifiers $\mathcal{C}$ that do *not* take the sensitive attribute A as input.

To make the linear classifier fair, we can incorporate a measure of fairness into Problem (1), either as a constraint or a regularization term added to the objective function. There are a plethora of fairness measures that we can use to promote fairness in this case, including

the demographic parity (Calders et al. 2009), equalized odds, and equal opportunity (Hardt et al. 2016, Zafar et al. 2017) among many others. The demographic parity criterion requires that the predictor be statistically independent of the sensitive attribute A. Intuitively, demographic parity enforces the probability of getting good outcomes to be the same across the privileged and unprivileged groups. However, demographic parity does not consider the actual label Y. It has been argued that demographic parity is not the most relevant notion of fairness in cases where we have ground truth on the quality of the candidates (Hardt et al. 2016, Zafar et al. 2017). In contrast, equalized odds is a much stronger definition by using the true label Y: It requires that the positive outcome is conditionally independent of the sensitive attributes given the true label. However, as we associate $Y = 1$ with positive outcomes such as being hired, decision makers are generally more interested in the true-positive rate than the false-positive rate. Also, the equalized odds criterion can be too strict to hurt accuracy (Hardt et al. 2016). A reasonable relaxation of equalized odds only imposes fairness within the desirable outcome ($Y = 1$) group, also known as equal opportunity (EO). The EO criterion requires the true-positive rate of the classifier to be invariant across the sensitive groups, and it will be the focus of this paper. We refer the reader to the references (Corbett-Davies et al. 2017, Chouldechova and Roth 2020, Berk et al. 2021, Mehrabi et al. 2021) for comprehensive treatments of fairness in machine learning in general and in the classification problem in particular. Unfortunately, the EO unfairness measure is challenging to formulate due to its nonconvexity (Donini et al. 2018). Moreover, one can verify that the EO unfairness constraint leads to an open feasible set, which prohibits exact mixed binary programming reformulations (Jeroslow 1987). To alleviate intractability, simple functions such as linear functions (Agarwal et al. 2018, Donini et al. 2018) and log functions (Taskesen et al. 2020) have been used to approximate the unfairness measure. Recently, the paper (Ye and Xie 2020) proposes a mixed binary model incorporating nonconvex approximations of the fairness measures as a regularization term to enhance fairness.

The existing notions of fairness proposed in the literature necessitate precise knowledge about the joint probability distribution that governs $(\boldsymbol{X}, A, Y)$. In practice, this distribution is rarely available to the decision makers and is typically estimated using the empirical distribution generated from the imbalanced—and possibly biased—historical observations. Although the empirical-based methods may work well on the observed data set, they often fail to yield fairness in practice because they do not generalize to out-of-sample data that have not been observed. For example, since there are fewer females in the technical positions at Amazon, relying on the empirical distribution can give rise to severe overfitting that yields an unfair hiring decision. Conversely, even if the true underlying distribution is available, computing the fairness of the decision is generically intractable (#P-hard; Dyer and Frieze 1988) because it involves evaluating a multidimensional integration (e.g., computing the probability of getting hired conditionally on being an unprivileged person).

In this paper, we endeavor to address this problem using the ideas of *distributionally robust optimization (DRO)*. The DRO approach does not impose a single distribution of the features, the sensitive attributes, and the response label of the entities in the population. Instead, it constructs a set of plausible probability distributions that are locally consistent with the available data set. The DRO approach then optimizes for a safe classifier that performs best under the most adverse distribution from within the prescribed distribution set. This approach thus may yield a fair classifier that has provable guarantees on the out-of-sample data.

Our paper belongs to an emerging class of fairness aware distributionally robust algorithms. Recently, a repeated loss minimization model with a χ^2-divergence ambiguity set is considered in Hashimoto et al. (2018). Alternatively, Rezaei et al. (2020) embeds the fairness constraint in the ambiguity set and proposes a robust classification model. When only the labels are noisy, robust fairness constraints based on a total variation ambiguity set are described in Wang et al. (2020). In this paper, we consider adversarial perturbations based on the Wasserstein distance (Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019, Kuhn et al. 2019, Gao and Kleywegt 2023, Ho-Nguyen and Wright 2023), in particular, the type-∞ Wasserstein distance (Givens and Shortt 1984; Bertsimas et al. 2018, 2022; Nguyen et al. 2020; Xie 2020). The Wasserstein distance has attracted significant attention in machine learning and robust optimization due to its statistical properties and metric interpretation. We remark that Wasserstein distributionally robust classification has been proposed to promote individual fairness (Yurochkin et al. 2020). Unfortunately, incorporating Wasserstein distance with the aforementioned fairness measures to encourage group fairness is more challenging because the decision maker has to solve a Wasserstein min-max statistic learning problem (Blanchet et al. 2019, Shafieezadeh-Abadeh et al. 2019, Nguyen et al. 2022) with nonconvex conditional probability terms. The recent study (Taskesen et al. 2020) considers uncertainty only in the feature space and convexifies the probability terms in the EO unfairness measure using the log function. The convexified log-EO unfairness measure is introduced to a distributionally robust logistic regression model as a fairness-driven regularization term to promote group fairness. Although the trained log-probabilistic fair logistic classifier demonstrates its effectiveness in the

empirical experiments, it cannot offer any guarantee on the misclassification probability or fairness score, even within the training data set. Compared with Taskesen et al. (2020), our model minimizes the misclassification probability while controlling the EO unfairness measure, which is more interpretable than the log-prob loss and log-EO unfairness measure used in Taskesen et al. (2020). Unfortunately, the nominal model leads to an open feasible region, which is challenging to write an exact reformulation that off-the-shelf solvers can readily solve (MOSEK ApS 2024). We thus propose a tight conservative approximation to the open safety set and derive a mixed-binary conic reformulation. In addition, we invoke the Wasserstein robust learning framework to handle uncertainty from the features, the sensitive attributes, and the response label of the entities in the population. Considering the mixed-binary model may encounter computational difficulties with large data sets, we also develop a conservative convex model that can be solved efficiently with large instances. Both models provide performance guarantees on the misclassification probability and unfairness score and achieve attractive performance in the numerical experiments.

1.1. Contributions

The contributions of this paper can be summarized as follows.

- **A new distributionally robust fairness aware classifier model:** We propose a one-sided unfairness measure motivated by the EO criterion and impose this unfairness measure as a constraint of a distributionally robust misclassification probability minimization problem. Compared with the generally adopted two-sided unfairness measures (Agarwal et al. 2018, Taskesen et al. 2020, Ye and Xie 2020), this one-sided unfairness measure reduces the number of constraints by explicitly tracking the difference of true-positive rate between the privileged and unprivileged groups. We then consider the worst-case unfairness measure and the worst-case misclassification probability under the most unfavorable distributions within the type-∞ Wasserstein ambiguity set constructed around the empirical distribution. The developed distributionally robust fairness aware classifier can manage multiple sources of uncertainty, such as those from features, labels, and marginal probabilities. If the radius of the ambiguity set diminishes to zero, our formulation reverts to the unfairness measure evaluated at the empirical distribution. As such, our proposed robust learning scheme can be leveraged as a regularization of the empirical-based method.
- **Tight conservative approximation and its reformulation:** Unfortunately, the nominal distributionally robust fairness aware classification problem suffers from the openness issue: its feasible region may be open, and its objective function may not be lower-semicontinuous. This issue prohibits an exact reformulation in a form that can be readily understood and solved by off-the-shelf optimization solvers. We present an arbitrarily precise approximation by substituting the open safety set with a tight inner closed approximation. The approximation admits a mixed-binary conic reformulation, which off-the-shelf solvers can solve.
- **Hinge-loss–based fairness aware model:** To enhance scalability, we propose a *convex* distributionally robust fairness aware classification model: This model uses the convex hinge loss function to approximate the unfairness measure and the objective function. Experimental results demonstrate that this classifier generates a marked improvement in fairness, with a negligible loss of predictive accuracy. Interestingly, minimizing the expected hinge loss is precisely the conditional value at risk (CVaR) approximation of the misclassification probability minimization problem.

The paper is organized as follows. Section 2 describes the distributionally robust fairness aware classification problem. Section 3 proposes a conservative approximation to the original problem and provides a binary optimization reformulation for training the model. Section 4 further proposes a convex fairness aware model for large instances, and a convex optimization reformulation is derived for training. Section 5 discusses the situation of uncertainties in the sensitive attribute and label. Finally, Section 6 reports on the numerical experiments.

1.2. Notations

For any set $\mathcal{S}$, we use $\mathcal{M}(\mathcal{S})$ to denote the set of probability measures supported on $\mathcal{S}$ and $|\mathcal{S}|$ to denote its cardinality. For any logical expression $\mathcal{E}$, the indicator function $\mathbb{I}(\mathcal{E})$ admits value 1 if $\mathcal{E}$ is true and value 0 if $\mathcal{E}$ is false. For any norm $\|\cdot\|$ on $\mathbb{R}^d$, we use $\|\cdot\|_*$ to denote its dual norm. We use $\mathbb{R}_+$ to denote the set of nonnegative real numbers and $\mathbb{R}_{++}$ to denote the set of strictly positive real numbers. For any vector $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$, we define $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{(n+m)}$ to be the combination of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

2. Distributionally Robust Fairness Aware Linear Classifiers

Throughout this section, we focus on promoting the fairness of a linear classifier with respect to the criterion of equal opportunity, also known as equality of opportunity (Hardt et al. 2016). This criterion is formally defined as follows.

Definition 2.1 (EO). A classifier $\mathcal{C} : \mathcal{X} \to \mathcal{Y}$ satisfies the equal opportunity criterion relative to a probability measure $\mathbb{Q}$ if

$$\mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1 \,|\, A = 1, Y = 1) = \mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1 \,|\, A = 0, Y = 1).$$

The definition requires that the true-positive rate is the same across the privileged and unprivileged groups. In practice, we often observe that the classifier may have *higher* true-positive rate for the privileged group $A = 1$. We track this performance discrepancy using the one-sided *EO unfairness measure* defined by

$$\mathbb{U}(\boldsymbol{w}, b, \mathbb{Q}) \triangleq \mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1 \,|\, A = 1, Y = 1) - \mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1 \,|\, A = 0, Y = 1), \tag{2}$$

which measures the difference between the true-positive rate of the privileged group ($A = 1$) and that of the unprivileged group ($A = 0$).

We say that a classifier is *trivial* if it is parametrized by $(\boldsymbol{w}, b) = (\boldsymbol{0}, 0) \in \mathbb{R}^{d+1}$; in this case, $\mathcal{C}(\boldsymbol{x}) = 1$ for any input $\boldsymbol{x} \in \mathcal{X}$. It is easy to verify that the trivial classifier is also fair with respect to any possible distribution $\mathbb{Q}$, but it may not attain a desirable level of predictive performance. This paper aims to search for a *nontrivial* classifier that balances fairness and predictive power. To this end, suppose that $\mathbb{P}^\star \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$ is the data-generating distribution of the joint random vector $(\boldsymbol{X}, A, Y)$. The fair linear classifier solves the constrained misclassification probability minimization problem:

$$\begin{array}{ll} \min & \mathbb{P}^\star(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) \le 0) \\ \text{s.t.} & \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R} \\ & \mathbb{U}(\boldsymbol{w}, b, \mathbb{P}^\star) \le \eta. \end{array} \tag{3}$$

The objective function of (3) minimizes the misclassification probability, whereas the constraint of (3) imposes an upper bound η on the unfairness measure with respect to $\mathbb{P}^\star$. A major challenge of Problem (3) is that the data-generating distribution $\mathbb{P}^\star$ is elusive to the decision maker. Even if $\mathbb{P}^\star$ is known, the probabilistic program (3) is, in general, computationally intractable because computing the probability of an event involving multiple random variables belongs to the complexity class #P-hard (Dyer and Frieze 1988)—which is perceived to be harder than the class NP-hard. In a data-driven setting, we assume that we have access to N training samples generated from $\mathbb{P}^\star$. Let $\hat{\mathbb{P}}$ be the empirical distribution supported on $\{(\hat{\boldsymbol{x}}_i, \hat{a}_i, \hat{y}_i)\}_{i=1}^N$, we will construct an ambiguity set around $\hat{\mathbb{P}}$ using the Wasserstein distance. Let $\Xi \triangleq \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ be the joint outcome space of the covariate, the sensitive attribute, and the label; we endow Ξ with a metric c. A formal definition of the Wasserstein distance is as follows.

Definition 2.2 (Wasserstein Distance). The type-l ($1 \le l < +\infty$) Wasserstein distance between two distributions $\mathbb{Q}$ and $\mathbb{Q}'$ supported on Ξ is defined as

$$\mathbb{W}_l(\mathbb{Q}, \mathbb{Q}') \triangleq \inf\left\{ (\mathbb{E}_\pi[c(\boldsymbol{\xi}, \boldsymbol{\xi}')^l])^{\frac{1}{l}} : \pi \in \Pi(\mathbb{Q}, \mathbb{Q}') \right\},$$

where $\Pi(\mathbb{Q}, \mathbb{Q}')$ is the set of all probability measures on $\Xi \times \Xi$ with marginals $\mathbb{Q}$ and $\mathbb{Q}'$, respectively. The type-∞ Wasserstein distance is defined as the limit of $\mathbb{W}_l$ as l tends to ∞ and amounts to

$$\mathbb{W}_\infty(\mathbb{Q}, \mathbb{Q}') \triangleq \inf\left\{ \operatorname*{ess\,sup}_\pi \{c(\boldsymbol{\xi}, \boldsymbol{\xi}') : (\boldsymbol{\xi}, \boldsymbol{\xi}') \in \Xi \times \Xi\} : \pi \in \Pi(\mathbb{Q}, \mathbb{Q}') \right\}.$$

The Wasserstein distance is an intuitive way of comparing two distributions when one is derived from the other by small, nonuniform perturbations. The decision variable π can be interpreted as a *transportation plan* for moving a mass distribution denoted by $\mathbb{Q}$ to another one denoted by $\mathbb{Q}'$, where the transportation cost between two points $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ is measured using $c(\boldsymbol{\xi}, \boldsymbol{\xi}')$. Thus, the type-$l$ Wasserstein distance can be viewed as the lth root of the minimum transportation cost between $\mathbb{Q}$ and $\mathbb{Q}'$. When l tends to ∞, the type-l Wasserstein distance $\mathbb{W}_l(\mathbb{Q}, \mathbb{Q}')$ converges to the type-∞ Wasserstein distance $\mathbb{W}_\infty(\mathbb{Q}, \mathbb{Q}')$ (Givens and Shortt 1984), where esssup is the essential supremum (Rudin 1964). This paper constructs the ambiguity set based on the type-∞ Wasserstein distance because it offers tractable reformulations with attractive convergent properties (Xie 2020).

The ground metric on Ξ is supposed to be separable, meaning that c can be written as a sum of three components as

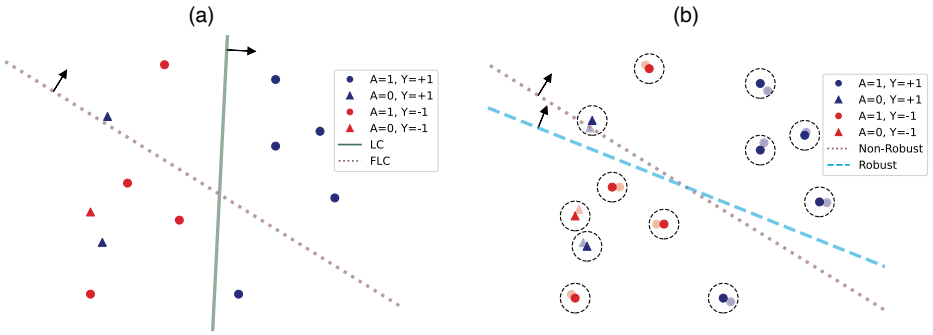
$$c((\boldsymbol{x}', a', y'), (\boldsymbol{x}, a, y)) = \|\boldsymbol{x} - \boldsymbol{x}'\| + \kappa_{\mathcal{A}} |a - a'| + \kappa_{\mathcal{Y}} |y - y'|$$

for some parameters $\kappa_{\mathcal{A}} \in [0, +\infty]$ and $\kappa_{\mathcal{Y}} \in [0, +\infty]$. Moreover, let $\hat{p}_{ay} = \hat{\mathbb{P}}(A = a, Y = y)$ denote the empirical marginals constructed from the training samples. We will consider the following marginally-constrained ambiguity set

$$\mathbb{B}(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) : \begin{array}{l} \mathbb{W}_\infty(\mathbb{Q}, \hat{\mathbb{P}}) \le \rho, \\ \mathbb{Q}(A = a, Y = y) = \hat{p}_{ay} \quad \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\}, \tag{4}$$

which is a neighborhood around the empirical distribution $\hat{\mathbb{P}}$. Intuitively, $\mathbb{B}(\hat{\mathbb{P}})$ contains all the distributions of $(\boldsymbol{X}, A, Y)$, which is of a type-∞ Wasserstein distance less than or equal to ρ from $\hat{\mathbb{P}}$, and at the same time has the same marginal distribution on (A, Y) as $\hat{\mathbb{P}}$. The ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ is thus parametrized by ρ, and the marginals $\hat{p}$; however, the dependence on these parameters is implicit. Adding a marginal constraint to the ambiguity set is an expedient practice to achieve tractable reformulation, especially when dealing with conditional expectation constraints that are prevalent in

Figure 1. (Color online) Classification Hyperplanes (Dashed) Obtained by Different Approaches



Notes. Color encodes the labels and shape encodes the sensitive attributes. The arrows point to the positive halfspace determined by the classification hyperplanes. Light-colored samples depict an exemplary distribution $\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})$. One can verify that this exemplary distribution also serves as an extremal distribution for the purple classification hyperplane. (a) Fair and unfair. (b) Robust and nonrobust.

fairness (Taskesen et al. 2020). Indeed, the conditional expectation is a *non*-linear function of the probability measure. However, when confining inside the set $\mathbb{B}(\hat{\mathbb{P}})$, we have

$$\mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1 \,|\, A = a, Y = y)$$
$$= \hat{p}_{ay}^{-1} \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{\{x:\mathcal{C}(x)=1\}}(\boldsymbol{X}) \mathbb{1}_{(a,y)}(A, Y)] \qquad \forall (a, y) \in \mathcal{A} \times \mathcal{Y},$$

which are linear functions of $\mathbb{Q}$ and conveniently simplifies the problem. We now provide a concrete example to illustrate the aforementioned fair classification problems and distributionally robust classification framework.

Example 2.3 (Robust Fair Classification). We consider a simple two-dimensional fair classification problem with 12 points partitioned into two classes. The sensitive attribute A is denoted by the shape of samples (triangle or circle), which in real life could represent gender, race, or any other sensitive features. A brief summary of samples is presented in Table 1. In Figure 1(a), the solid hyperplane represents a linear classifier that is optimal to the misclassification probability minimization problem (1), denoted by LC. From the classification outcomes of LC, all circles with a positive label are classified correctly, whereas a triangle with a positive label is classified to the negative halfspace. Therefore, this classifier is unfair in terms of the EO criterion because the difference in the true-positive rate between the privileged (circle) and the unprivileged (triangle) groups is

$$\mathbb{U}_{\mathrm{LC}} = 5/5 - 0/2 = 1.$$

Table 1. Sample Classes and Sensitive Attributes

	Circle ($A = 1$)	Triangle ($A = 0$)
Negative class (red)	4	1
Positive class (blue)	5	2

Note. In both classes, most samples come from the privileged group (circle).

On the contrary, the dotted line is one example of a fair linear classifier that is optimal to Problem (3) under the empirical distribution, denoted by FLC. In the outcomes of FLC, the triangle with a positive label is classified correctly, and the unfairness score decreases to

$$\mathbb{U}_{\mathrm{FLC}} = 4/5 - 1/2 = 0.3.$$

Figure 1(b) compares a nonrobust and a robust FLC. The robust FLC is optimal to the distributionally robust model (5). Intuitively, the type-∞ Wasserstein ambiguity set can be visualized using a combination of balls with radius ρ centered at each sample. Any perturbation from the original sample within the corresponding ball constitutes a distribution $\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})$ (e.g., light-colored points visualize a discrete distribution $\mathbb{Q}$ in the Wasserstein ambiguity set). Considering all distributions within the ambiguity set, the DRO model will yield a robust classifier against noises and disturbances. Conversely, although the nonrobust fair classifier achieves the same scores under the empirical distribution, it may suddenly fail under small perturbations from the empirical distribution.

Equipped with the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$, we can consider the fairness aware distributionally robust linear classification problem:

$$\begin{aligned} \min \quad & \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) \le 0) \\ \text{s.t.} \quad & \boldsymbol{w} \in \mathbb{R}^d, \ b \in \mathbb{R}, \\ & \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{U}(\boldsymbol{w}, b, \mathbb{Q}) \le \eta. \end{aligned} \tag{5}$$

The constraint of Problem (5) depends on a tolerance $\eta \in \mathbb{R}_+$: It requires that the true-positive rate for the privileged group $A = 1$ cannot be larger than the true-positive rate for the unprivileged group $A = 0$ plus a tolerance η, uniformly over all distributions in the

ambiguity set. It is easy to verify that the trivial classifier with $(\boldsymbol{w}, b) = (\mathbf{0}, 0)$ is feasible for (5) with an objective value of one.

Unfortunately, it is challenging to write an exact reformulation of Problem (5) that off-the-shelf solvers can solve. Given any probability measure $\mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$, we can leverage the finite cardinality of $\mathcal{A}$ and $\mathcal{Y}$ to decompose $\mathbb{Q}$ using its conditional measures $\mathbb{Q}_{ay}(\boldsymbol{X} \in \cdot) = \mathbb{Q}(\boldsymbol{X} \in \cdot | A = a, Y = y)$. The EO unfairness measure $\mathbb{U}$ defined in (2) can be written as

$$\mathbb{U}(\boldsymbol{w}, b, \mathbb{Q}) \triangleq \mathbb{Q}_{11}(\boldsymbol{w}^\top \boldsymbol{X} + b \geq 0) - \mathbb{Q}_{01}(\boldsymbol{w}^\top \boldsymbol{X} + b \geq 0). \tag{6}$$

The set $(\boldsymbol{w}, b)$ satisfying the constraint $\sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{U}(\boldsymbol{w}, b, \mathbb{Q}) \leq \eta$ is, in general, an open set, and the root cause stems from the inequality inside $\mathbb{Q}_{11}(\boldsymbol{w}^\top \boldsymbol{X} + b \geq 0)$ in the previous equation. To see this, consider the special case where $\rho = 0$, which implies that $\mathbb{B}(\hat{\mathbb{P}}) = \{\hat{\mathbb{P}}\}$, and the constraint of (5) becomes $\mathbb{U}(\boldsymbol{w}, b, \hat{\mathbb{P}}) \leq \eta$. Let us define the index set $\mathcal{I}_{a1} = \{i \in [N] : \hat{a}_i = a, \hat{y}_i = 1\}$ containing indices of the samples with sensitive attribute a and label 1. We have

$$\begin{aligned}
\mathbb{U}(\boldsymbol{w}, b, \hat{\mathbb{P}}) &= \mathbb{E}_{\hat{\mathbb{P}}}[\hat{p}_{11}^{-1} \mathbb{I}(\boldsymbol{w}^\top \boldsymbol{X} + b \geq 0) \mathbb{1}_{(1,1)}(A, Y) \\
&\quad - \hat{p}_{01}^{-1} \mathbb{I}(\boldsymbol{w}^\top \boldsymbol{X} + b \geq 0) \mathbb{1}_{(0,1)}(A, Y)] \\
&= \frac{1}{N} \left(\hat{p}_{11}^{-1} \sum_{i \in \mathcal{I}_{11}} \mathbb{I}(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b \geq 0) \right. \\
&\quad \left. + \hat{p}_{01}^{-1} \sum_{i \in \mathcal{I}_{01}} \mathbb{I}(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b < 0) - \hat{p}_{01}^{-1} |\mathcal{I}_{01}| \right).
\end{aligned}$$

Thus, the fairness constraint $\mathbb{U}(\boldsymbol{w}, b, \hat{\mathbb{P}}) \leq \eta$ can be written as

$$\frac{1}{N} \left(\hat{p}_{11}^{-1} \sum_{i \in \mathcal{I}_{11}} \mathbb{I}(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b \geq 0) \right.$$
$$\left. + \hat{p}_{01}^{-1} \sum_{i \in \mathcal{I}_{01}} \mathbb{I}(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b < 0) - \hat{p}_{01}^{-1} |\mathcal{I}_{01}| \right) \leq \eta.$$

Consider now the simplest case where $\mathcal{I}_{01}$ is empty. The previous constraint is simplified into

$$\frac{1}{N \hat{p}_{11}} \sum_{i \in \mathcal{I}_{11}} \mathbb{I}(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b \geq 0) \leq \eta.$$

It can be verified that for $\eta \in (0, 1)$, the feasible region of $(\boldsymbol{w}, b)$ with the previous constraint is an open set, and unfortunately, this open set cannot be reformulated to a bounded mixed integer program (MIP) problem (Jeroslow 1987, theorem 2.1). For example, if $\eta < 1/(N \hat{p}_{11})$, then $(\boldsymbol{w}, b)$ must satisfy $\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b < 0$ for all $i \in \mathcal{I}_{a1}$. Because the intersection of open sets is open, the previous constraint is not bounded-MIP representable. Similarly, the objective function of (5) encounters a similar issue. By setting $\rho = 0$, Problem (5) is equivalent to

$$\begin{aligned}
\min \quad & \tau \\
\text{s.t.} \quad & \boldsymbol{w} \in \mathbb{R}^d, \ b \in \mathbb{R}, \ \tau \in \mathbb{R}, \ \|(\boldsymbol{w}, b)\| \leq 1, \\
& \mathbb{U}(\boldsymbol{w}, b, \hat{\mathbb{P}}) \leq \eta \\
& \hat{\mathbb{P}}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) \leq 0) \leq \tau.
\end{aligned}$$

When $\tau = 0$, the last constraint requires $(\boldsymbol{w}, b)$ satisfying $\hat{y}_i(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b) > 0 \ \ \forall i \in [N]$, which cannot be reformulated to a bounded MIP constraint.

In the following sections, we will develop approximations to Problem (5) that can resolve these openness issues of the feasible set and the objective function. To conclude this section, we describe one possible approach to choosing η in practice. Given a training data set, the decision maker first finds the empirical classifier by solving (1) with $\mathbb{Q}$ being replaced by the empirical probability measure. From the empirical classifier, the decision maker can identify the group with a higher true-positive rate as the privileged group ($A = 1$). Next, the empirical unfairness score $\hat{\eta}$ is calculated by taking the difference in the true-positive rate between the privileged and the unprivileged group. If this empirical unfairness score is less than the tolerance level of the decision maker, there is no need to impose fairness constraints and resolve the fair classification problem. If the empirical unfairness score is too large, the decision maker could gradually decrease η starting from the empirical unfairness score $\hat{\eta}$. During this process, the decision maker should actively monitor the classifier's performance until a fair classifier that satisfies the requirements is found.

3. The ε-Distributionally Robust Fairness Aware Classifier

In this section, we propose an approximation of the original problem (5) and derive its reformulation. We first introduce a norm constraint $\|(\boldsymbol{w}, b)\| \leq 1$ to restrict the feasible region to a compact set. This constraint is a general approach to reformulate indicator functions (Liittschwager and Wang 1978, equation (2.9)), and it does not alter the classification result because of the scaling-invariant property (Shalev-Shwartz and Ben-David 2014). Recall that the openness of the feasible set as previously described is because the function $\mathbb{I}(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b \geq 0)$ is an upper-semicontinuous function in the variable $(\boldsymbol{w}, b)$. Hence, to generate a closed approximation of the feasible set, it is necessary to switch the inequality sign highlighted in red in (6) to a *strict* inequality. For notation simplicity, we use $\mathbb{Q}_{ay}$ to denote the conditional distribution, that is, $\mathbb{Q}_{ay}(\boldsymbol{X} \in \cdot) = \mathbb{Q}(\boldsymbol{X} \in \cdot | A = a, Y = y)$. We consider the modified one-

sided EO *un*fairness measure $\mathbb{U}_\varepsilon$ as

$$\mathbb{U}_\varepsilon(\boldsymbol{w}, b, \mathbb{Q}) \triangleq \mathbb{Q}_{11}(\boldsymbol{w}^\top \boldsymbol{X} + b > -\varepsilon) - \mathbb{Q}_{01}(\boldsymbol{w}^\top \boldsymbol{X} + b \geq 0), \tag{7}$$

which is parametrized by a strictly positive value $\varepsilon \in \mathbb{R}_{++}$. Similarly, as the objective function of (5) does not admit an exact reformulation, we replace it with $\mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) < \varepsilon)$, which is a conservative approximation of the misclassification probability for any $\varepsilon \in \mathbb{R}_{++}$. The next proposition demonstrates that these approximations are tight in the limit as ε tends to zero.

Proposition 3.1 (Convergence). *Fix a measure $\mathbb{Q}$, the ε-unfairness measure $\mathbb{U}_\varepsilon$ converges to the EO unfairenss measure $\mathbb{U}$ as $\varepsilon \to 0$, that is, $\lim_{\varepsilon \to 0} \mathbb{U}_\varepsilon(\boldsymbol{w}, b, \mathbb{Q}) = \mathbb{U}(\boldsymbol{w}, b, \mathbb{Q})$. Similarly, we have $\lim_{\varepsilon \to 0} \mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) < \varepsilon) = \mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) \leq 0)$.*

Proposition 3.1 indicates that for any fixed distribution $\mathbb{Q}$ and classifier $(\boldsymbol{w}, b)$, the modified objective function and unfairness measure converges to the original terms when ε goes to zero. However, we cannot set ε to zero because minimizing $\mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) < 0)$ returns a trivial optimal solution $(\boldsymbol{w}, b) = (\boldsymbol{0}, 0)$ with optimal value $\mathbb{Q}(Y(\boldsymbol{0}^\top \boldsymbol{X} + 0) < 0) = 0$. Hence, although a small ε value offers a good approximation, a too-small ε may induce computational errors in optimization solvers. For example, when ε is smaller than the constraint error tolerance level of the solvers, the solver may still take the trivial value $(\boldsymbol{w}, b) = (\boldsymbol{0}, 0)$ as optimal. In practice, the user can decrease ε from a positive value and stop until the solver returns the trivial solution to obtain a tight approximation. Combining the modified objective function and unfairness measure, we consider the following ε-distributionally robust fairness aware classification (ε-DRFC) problem:

$$\begin{aligned} \min \quad & \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) < \varepsilon) \\ \text{s.t.} \quad & \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R},\ \|(\boldsymbol{w}, b)\| \leq 1, \\ & \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{U}_\varepsilon(\boldsymbol{w}, b, \mathbb{Q}) \leq \eta. \end{aligned} \tag{8}$$

We now show that the ε-DRFC problem (8) is a conservative approximation of (5), that is, the objective value of (8) and η provide upper bounds on the misclassification rate and EO unfairness measure, respectively.

Proposition 3.2 (Conservative Approximation). *Let $(\boldsymbol{w}^\star, b^\star)$ be the optimal solution to Problem (8). Then $(\boldsymbol{w}^\star, b^\star)$ is feasible for Problem (5). Moreover, let $v^\star$ be the corresponding optimal value of Problem (8), then*

$$\mathbb{Q}(Y((\boldsymbol{w}^\star)^\top \boldsymbol{X} + b^\star) \leq 0) \leq v^\star \qquad \forall \mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}).$$

The ε-DRFC model (8) enables decision makers to bound the unfairness measure in the training set explicitly using η. Moreover, as shown in Proposition 3.2, the optimal value of Problem (8) constitutes an upper bound on the misclassification probability.

For any $\varepsilon \in \mathbb{R}_{++}$, we show that Problem (8) admits a mixed binary conic reformulation. We first consider the case where we have absolute trust in sensitive attributes and labels; that is, we use the ground metric

$$c((\boldsymbol{x}', a', y'), (\boldsymbol{x}, a, y)) = \|\boldsymbol{x} - \boldsymbol{x}'\| + \infty |a - a'| + \infty |y - y'|, \tag{9}$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^d$. In this setting, we set $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$, which indicates that we have absolute trust in the value of the sensitive attribute A and the label Y. When c is chosen as in (9), a simple modification of the proof of (Taskesen et al. 2021, theorem 3.2) shows that any distribution $\mathbb{Q}$ with $\mathbb{W}_\infty(\mathbb{Q}, \hat{\mathbb{P}}) < \infty$ should satisfy $\mathbb{Q}(A = a, Y = y) = \hat{p}_{ay}$ for all $(a, y) \in \mathcal{A} \times \mathcal{Y}$. Consequently, the marginal constraint defining the set $\mathbb{B}(\hat{\mathbb{P}})$ becomes redundant and can be omitted. This simplification with absolute trust in the sensitive attribute and label has been previously exploited to derive hypothesis tests for fair classifiers (Taskesen et al. 2021) and to train fair logistic classifier (Taskesen et al. 2020). In Section 5, we will further discuss the general ground metric with finite positive values of κ_A and κ_Y. The next theorem asserts the reformulation of the min-sup problem (8) as a mixed binary conic optimization problem.

Theorem 3.3 (ε-DRFC Reformulation). *Suppose that the ground metric is prescribed using (9), then the ε-DRFC model (8) is equivalent to the conic mixed binary optimization problem*

$$\begin{aligned} \min \quad & \frac{1}{N} \sum_{i=1}^N t_i \\ \text{s.t.} \quad & \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R},\ \boldsymbol{t} \in \{0,1\}^N,\ \boldsymbol{\lambda} \in \{0,1\}^N, \|(\boldsymbol{w}, b)\| \leq 1, \\ & -\hat{y}_i(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b) + \rho \|\boldsymbol{w}\|_* \leq M t_i - \varepsilon \qquad \forall i \in [N], \\ & \frac{1}{|\mathcal{I}_{11}|} \sum_{i \in \mathcal{I}_{11}} \lambda_i + \frac{1}{|\mathcal{I}_{01}|} \sum_{i \in \mathcal{I}_{01}} \lambda_i - 1 \leq \eta, \\ & \boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + \rho \|\boldsymbol{w}\|_* + b + \varepsilon \leq M \lambda_i \quad \forall i \in \mathcal{I}_{11}, \\ & -\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + \rho \|\boldsymbol{w}\|_* - b \leq M \lambda_i \quad \forall i \in \mathcal{I}_{01}, \end{aligned} \tag{10}$$

where M is the big-M parameter.

For notational simplicity, we present the reformulation (10) with $2N$ binary variables. A closer investigation into Problem (10) reveals that it suffices to use $N + |\mathcal{I}_1|$ binary variables, where $\mathcal{I}_1 = \{i \in [N] : \hat{y}_i = 1\}$ is the index set of training samples with positive labels. If $\|\cdot\|$ is either the 1-norm or the ∞-norm on $\mathbb{R}^d$, Problem (10) is a mixed binary *linear* optimization problem.

If $\|\cdot\|$ is the Euclidean norm, Problem (10) becomes a mixed binary second-order cone optimization problem. Both problems can be solved using off-the-shelf solvers such as MOSEK (MOSEK ApS 2024).

For the remainder of this section, we will provide the proof for Theorem 3.3. This proof relies on the following auxiliary result.

Lemma 3.4 (Indicator Function Reformulation). *Fix any index set $\mathcal{K} \subseteq \{1,\ldots,N\}$, a radius $\rho \in \mathbb{R}_+$, a classifier $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$ and a collection of samples $\{\hat{\boldsymbol{x}}_k\}_{k\in\mathcal{K}}$. For any $\varepsilon \in \mathbb{R}$, we have*

$$\sum_{k\in\mathcal{K}} \sup_{\boldsymbol{x}_k:\|\boldsymbol{x}_k-\hat{\boldsymbol{x}}_k\|\le\rho} \mathbb{I}(\boldsymbol{w}^\top\boldsymbol{x}_k + b > \varepsilon) = \begin{cases} \min & \sum_{k\in\mathcal{K}} \lambda_k \\ \text{s.t.} & \boldsymbol{\lambda} \in \{0,1\}^N, \\ & \boldsymbol{w}^\top\hat{\boldsymbol{x}}_k + \rho\|\boldsymbol{w}\|_* + b - \varepsilon \le M\lambda_k \quad \forall k \in \mathcal{K}, \end{cases}$$

where M is the big-M parameter.

Equipped with Lemma 3.4, we are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. By exploiting the choice of c with an infinite unit cost on $\mathcal{A}$ and $\mathcal{Y}$, the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ can be re-expressed as

$$\mathbb{B}(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X}\times\mathcal{A}\times\mathcal{Y}) : \begin{array}{l} \exists \pi_i \in \mathcal{M}(\mathcal{X}) \quad \forall i \in [N], \\ \mathbb{Q}(\mathrm{d}\boldsymbol{x}\times\mathrm{d}a\times\mathrm{d}y) = N^{-1} \\ \sum_{i=1}^N \pi_i(\mathrm{d}\boldsymbol{x})\delta_{(\hat{a}_i,\hat{y}_i)}(\mathrm{d}a\times\mathrm{d}y), \\ \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\| \le \rho \\ \quad \forall \boldsymbol{x}_i \in \mathrm{supp}(\pi_i) \end{array} \right\}, \tag{11}$$

where $\mathrm{supp}(\pi_i)$ denotes the support of the probability measure π_i (Aliprantis and Border 2006, p. 441). We first provide the reformulation for the objective function of (8). For any $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$, we have

$$\begin{aligned} &\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}(Y(\boldsymbol{w}^\top\boldsymbol{X} + b) < \varepsilon) \\ &= \frac{1}{N}\sum_{i=1}^N \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|\le\rho} \mathbb{I}(\hat{y}_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) < \varepsilon) \\ &= \begin{cases} \min & \frac{1}{N}\sum_{i=1}^N t_i \\ \text{s.t.} & \boldsymbol{t} \in \{0,1\}^N, \\ & -\hat{y}_i(\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i + b) + \rho\|\boldsymbol{w}\|_* \le Mt_i - \varepsilon \quad \forall i \in [N], \end{cases} \end{aligned}$$

where the last equality follows from an epigraphical reformulation and the result of Lemma 3.4. Next, we provide the reformulation for the constraints of (8). Define the following index sets $\mathcal{I}_{a1} = \{i \in [N] : \hat{a}_i = a, \hat{y}_i = 1\}\ \forall a \in \mathcal{A}$, for any $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$, we have

$$\begin{aligned} &\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}(\boldsymbol{w}^\top\boldsymbol{X} + b > -\varepsilon \,|\, A = 1, Y = 1) \\ &\quad - \mathbb{Q}(\boldsymbol{w}^\top\boldsymbol{X} + b \ge 0 \,|\, A = 0, Y = 1) \\ &= \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\hat{p}_{11}^{-1}\mathbb{I}(\boldsymbol{w}^\top\boldsymbol{X} + b > -\varepsilon)\mathbb{1}_{(1,1)}(A, Y) \\ &\quad - \hat{p}_{01}^{-1}\mathbb{I}(\boldsymbol{w}^\top\boldsymbol{X} + b \ge 0)\mathbb{1}_{(0,1)}(A, Y)] \\ &= \frac{1}{N}\left(\hat{p}_{11}^{-1}\sum_{i\in\mathcal{I}_{11}} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|\le\rho} \mathbb{I}(\boldsymbol{w}^\top\boldsymbol{x}_i + b > -\varepsilon) \right. \\ &\quad \left. - \hat{p}_{01}^{-1}\sum_{i\in\mathcal{I}_{01}} \inf_{\boldsymbol{x}_i:\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|\le\rho} \mathbb{I}(\boldsymbol{w}^\top\boldsymbol{x}_i + b \ge 0) \right) \\ &= \frac{1}{N}\left(\hat{p}_{11}^{-1}\sum_{i\in\mathcal{I}_{11}} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|\le\rho} \mathbb{I}(\boldsymbol{w}^\top\boldsymbol{x}_i + b > -\varepsilon) \right. \\ &\quad \left. - \hat{p}_{01}^{-1}(|\mathcal{I}_{01}| - \sum_{i\in\mathcal{I}_{01}} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|\le\rho} \mathbb{I}(\boldsymbol{w}^\top\boldsymbol{x}_i + b < 0)) \right) \\ &= \frac{1}{N}\left(\frac{N}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|\le\rho} \mathbb{I}(\boldsymbol{w}^\top\boldsymbol{x}_i + b > -\varepsilon) \right. \\ &\quad \left. + \frac{N}{|\mathcal{I}_{01}|}\sum_{i\in\mathcal{I}_{01}} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|\le\rho} \mathbb{I}(\boldsymbol{w}^\top\boldsymbol{x}_i + b < 0) - \frac{N}{|\mathcal{I}_{01}|}|\mathcal{I}_{01}| \right) \\ &= \begin{cases} \min & \frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}} \lambda_i + \frac{1}{|\mathcal{I}_{01}|}\sum_{i\in\mathcal{I}_{01}} \lambda_i - 1 \\ \text{s.t.} & \boldsymbol{\lambda} \in \{0,1\}^N \\ & \boldsymbol{w}^\top\hat{\boldsymbol{x}}_i + \rho\|\boldsymbol{w}\|_* + b + \varepsilon \le M\lambda_i \quad \forall i \in \mathcal{I}_{11} \\ & -\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i + \rho\|\boldsymbol{w}\|_* - b \le M\lambda_i \quad \forall i \in \mathcal{I}_{01}, \end{cases} \end{aligned}$$

where the last equality follows by applying Lemma 3.4 twice and by noticing that $\mathcal{I}_{11} \cap \mathcal{I}_{01} = \emptyset$. Setting the optimal value of the above minimization problem to be less than η completes the proof. □

Remark 3.5 (Big-M Value). For practical implementation, it is sufficient to set the big-M parameter to $M = C + \rho d + \varepsilon$, where $C = \max_{i\in[N]}\|(\hat{\boldsymbol{x}}_i, 1)\|_*$ is the largest dual norm value for all combined vectors $(\hat{\boldsymbol{x}}_i, 1)$. A short proof is provided in Online Appendix A.

Remark 3.6 (Out-of-Sample Guarantee). We also investigate the out-of-sample performance of Model (8). The ambiguity set (4) contains marginal constraints that require probability measures in the ambiguity sets to have the same marginal distribution as the empirical distribution. This constraint invalidates the

finite sample guarantees unless the true distribution shares the same marginal distribution with the empirical distribution. To see this, consider a simple example: When the true marginal probability is given by $\mathbb{P}^\star(A=1,Y=1)=1/\sqrt{2}$, for any finite sample size N, the underlying distribution will not be contained in the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ even if $\rho\to\infty$. In Online Appendix B, we illustrate that relaxing the marginal constraints does admit a solvable model with attractive theoretical results; however, the model is more computationally intensive as we add an extra layer of robustness.

Remark 3.7 (Balanced Accuracy). In the aforementioned model, we minimize the misclassification rate because the accuracy (one minus misclassification rate) is one of the most popular model evaluation metrics. However, accuracy can be misleading when the data set is imbalanced. In such cases, the decision maker should adopt other metrics that are more suitable for the imbalanced data set, such as balanced accuracy. We remark that our modeling framework and reformulation tricks can be easily extended to maximize balanced accuracy, the average accuracy obtained from both the positive and negative classes. A detailed discussion can be found in Online Appendix A.

The deterministic reformulation (10) may encounter computational difficulties as the sample size N grows because it involves $\mathcal{O}(N)$ binary variables. Thus, there is merit in studying tractable approximations that scale better with the sample size. The following section proposes one such approximation.

4. The Hinge Distributionally Robust Fairness Aware Classifier

We propose a convex approximation of Problem (5), which requires no binary variables. Observe that Problem (5) involves probability values in both the objective function and the unfairness constraint, and we will use conservative approximations of these probabilities in the sequel. First, we have for any distribution $\mathbb{Q}$ and for any classifier $(\boldsymbol{w},b)$:

$$\begin{aligned}\mathbb{Q}(Y(\boldsymbol{w}^\top\boldsymbol{X}+b)\le 0) &= \mathbb{E}_{\mathbb{Q}}[\mathbb{I}(Y(\boldsymbol{w}^\top\boldsymbol{X}+b)\le 0)]\\ &\le \mathbb{E}_{\mathbb{Q}}[\max\{0,1-Y(\boldsymbol{w}^\top\boldsymbol{X}+b)\}],\end{aligned}$$

where the previous inequality follows from the fact that $\mathbb{I}(z\le 0)\le\max\{0,1-z\}$. As a consequence, the objective function of Problem (5) can be upper-bounded as

$$\begin{aligned}&\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})}\ \mathbb{Q}(Y(\boldsymbol{w}^\top\boldsymbol{X}+b)\le 0)\\ &\le \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})}\ \mathbb{E}_{\mathbb{Q}}[\max\{0,1-Y(\boldsymbol{w}^\top\boldsymbol{X}+b)\}].\end{aligned}$$

This upper bound is also known as the hinge loss upper bound of the misclassification error that is well known in the machine learning literature (Chapelle et al. 2008, Bach 2021). Next, we consider an approximation of the EO unfairness measure. We rewrite the EO unfairness measure (2) as

$$\begin{aligned}\mathbb{U}(\boldsymbol{w},b,\mathbb{Q}) &= \mathbb{Q}_{11}(\boldsymbol{w}^\top\boldsymbol{X}+b\ge 0)+\mathbb{Q}_{01}(\boldsymbol{w}^\top\boldsymbol{X}+b<0)-1\\ &= \mathbb{E}_{\mathbb{Q}_{11}}[\mathbb{I}(\boldsymbol{w}^\top\boldsymbol{X}+b\ge 0)]\\ &\quad+\mathbb{E}_{\mathbb{Q}_{01}}[\mathbb{I}(\boldsymbol{w}^\top\boldsymbol{X}+b<0)]-1.\end{aligned}$$

Inspired by the previous hinge loss approximation, we propose the *hinge EO unfairness measure*:

$$\begin{aligned}\mathbb{H}(\boldsymbol{w},b,\mathbb{Q}) \triangleq\ &\mathbb{E}_{\mathbb{Q}_{11}}[\max\{0,1+\boldsymbol{w}^\top\boldsymbol{X}+b\}]\\ &+\mathbb{E}_{\mathbb{Q}_{01}}[\max\{0,1-\boldsymbol{w}^\top\boldsymbol{X}-b\}]-1.\end{aligned}$$

The hinge unfairness measure $\mathbb{H}$ is convex in $(\boldsymbol{w},b)$. To see this, each term $\max\{0,1+\boldsymbol{w}^\top\boldsymbol{x}+b\}$ and $\max\{0,1-\boldsymbol{w}^\top\boldsymbol{x}-b\}$ are convex in $(\boldsymbol{w},b)$ for any realization $\boldsymbol{X}=\boldsymbol{x}$. Because taking expectation preserves convexity (Boyd and Vandenberghe 2004, section 3.2.1), all the expectation terms in the previous equation are hence convex in $(\boldsymbol{w},b)$. The function $\mathbb{H}$ is convex for any fixed distribution $\mathbb{Q}$ because it is a pointwise maximum of two convex functions (Boyd and Vandenberghe 2004, section 3.2.3). Contrary to the unfairness measure $\mathbb{U}_\varepsilon$ defined in Section 3, the hinge unfairness measure does not constitute a tight upper bound for the EO unfairness measure.

Combining the hinge loss objective and the hinge unfairness measure, we propose the hinge distributionally robust fairness aware classification (HDRFC) problem:

$$\begin{aligned}\min\quad &\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})}\ \mathbb{E}_{\mathbb{Q}}[\max\{0,1-Y(\boldsymbol{w}^\top\boldsymbol{X}+b)\}]\\ \text{s.t.}\quad &\boldsymbol{w}\in\mathbb{R}^d,\ b\in\mathbb{R},\\ &\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})}\ \mathbb{H}(\boldsymbol{w},b,\mathbb{Q})\le\zeta.\end{aligned}\tag{12}$$

The constraint of Problem (8) depends on a tolerance $\zeta\in\mathbb{R}_+$: It requires that the hinge unfairness measure be smaller than ζ, uniformly over all distributions in the ambiguity set. Because $\mathbb{H}$ shares a different domain with $\mathbb{U}$, we deliberately use a separate parameter $\zeta\in\mathbb{R}_+$ in (12), which can differ from the parameter $\eta\in\mathbb{R}_+$ in (5). The way to choose ζ is similar to the approach of choosing η. Given a training data set, the decision maker first finds the empirical classifier by solving Problem (12) without the fairness constraint under the empirical distribution. From the empirical classifier, the decision maker can identify the group with a higher true-positive rate as the privileged group ($A=1$). Next,

the empirical unfairness score $\hat{\eta}$ and $\hat{\zeta}$ are calculated based on the EO and hinge EO unfairness measure, respectively. If the empirical EO unfairness score $\hat{\eta}$ is less than the tolerance level of the decision maker, there is no need to impose fairness constraints and resolve the fair classification problem. If the empirical unfairness score is too large, the decision maker could gradually decrease ζ starting from the empirical unfairness score $\hat{\zeta}$. During this process, the decision maker should actively monitor the performance of Model (12) until a fair classifier that satisfies the requirements is found. The following proposition illustrates that the HDRFC model is a conservative approximation to the original problem (5).

Proposition 4.1 (Conservative Approximation). *Suppose that Problem (12) with parameter ζ admits an optimal solution $(\boldsymbol{w}^\star, \boldsymbol{b}^\star)$. Let $v^\star$ be the corresponding optimal value of Problem (12) associated with $(\boldsymbol{w}^\star, \boldsymbol{b}^\star)$. Then*

$$\mathbb{Q}(Y((\boldsymbol{w}^\star)^\top \boldsymbol{X} + \boldsymbol{b}^\star) \le 0) \le v^\star \qquad \forall \mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}).$$

Furthermore, if $\zeta = \eta$, then $(\boldsymbol{w}^\star, \boldsymbol{b}^\star)$ is also feasible for Problem (5).

We next present the main result of this section, which asserts that the HDRFC problem (12) can be reformulated as a conic optimization problem.

Theorem 4.2 (HDRFC Reformulation). *Suppose that the ground metric is prescribed using (9), the HDRFC model (12) is a convex optimization, and it is equivalent to the conic optimization problem:*

$$\begin{array}{ll}
\min & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} t_i \\
\text{s.t.} & \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R},\ \boldsymbol{t} \in \mathbb{R}_+^N,\ \boldsymbol{\lambda} \in \mathbb{R}_+^N, \\
& -\hat{y}_i(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b) + \rho\|\boldsymbol{w}\|_* \le t_i - 1 \qquad \forall i \in [N], \\
& \dfrac{1}{|\mathcal{I}_{11}|}\displaystyle\sum_{i\in\mathcal{I}_{11}} \lambda_i + \dfrac{1}{|\mathcal{I}_{01}|}\sum_{i\in\mathcal{I}_{01}} \lambda_i - 1 \le \zeta, \\
& 1 + \boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + \rho\|\boldsymbol{w}\|_* + b \le \lambda_i \qquad \forall i \in \mathcal{I}_{11}, \\
& 1 - \boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + \rho\|\boldsymbol{w}\|_* - b \le \lambda_i \qquad \forall i \in \mathcal{I}_{01}.
\end{array} \tag{13}$$

Proof of Theorem 4.2. Because taking pointwise supremum over an infinite set of convex functions preserves convexity (Boyd and Vandenberghe 2004, section 3.2.3), we can observe that the objective function and the constraint function of (12) are both convex. Hence, (12) is a convex optimization problem. Exploiting Reformulation (11) of the set $\mathbb{B}(\hat{\mathbb{P}})$, we first reformulate the objective function of (12). For any $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$, we have

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(\boldsymbol{w}^\top \boldsymbol{X} + b)\}] \\
&= \frac{1}{N}\sum_{i=1}^{N} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\| \le \rho} \max\{0, 1 - \hat{y}_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\} \\
&= \frac{1}{N}\sum_{i=1}^{N} \max\left\{0, 1 - \inf_{\boldsymbol{x}_i:\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\| \le \rho} \hat{y}_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right\} \\
&= \begin{cases}
\min & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} t_i \\
\text{s.t.} & \boldsymbol{t} \in \mathbb{R}_+^N, \\
& -\hat{y}_i(\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + b) + \rho\|\boldsymbol{w}\|_* \le t_i - 1 \quad \forall i \in [N],
\end{cases}
\end{aligned}$$

where the last equality follows from an epigraphical reformulation and the properties of the dual norm. Next, we provide the reformulation for the constraints of (12). For any $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$, we have

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}_{11}}[\max\{0, 1 + \boldsymbol{w}^\top \boldsymbol{X} + b\}] \\
&\quad + \mathbb{E}_{\mathbb{Q}_{01}}[\max\{0, 1 - \boldsymbol{w}^\top \boldsymbol{X} - b\}] - 1 \\
&= \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\hat{p}_{11}^{-1}\max\{0, 1 + \boldsymbol{w}^\top \boldsymbol{X} + b\}\mathbb{1}_{(1,1)}(A, Y) \\
&\quad + \hat{p}_{01}^{-1}\max\{0, 1 - \boldsymbol{w}^\top \boldsymbol{X} - b\}\mathbb{1}_{(0,1)}(A, Y)] - 1 \\
&= \frac{1}{N}\left(\hat{p}_{11}^{-1}\sum_{i\in\mathcal{I}_{11}} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\| \le \rho} \max\{0, 1 + \boldsymbol{w}^\top \boldsymbol{x}_i + b\}\right. \\
&\quad \left. + \hat{p}_{01}^{-1}\sum_{i\in\mathcal{I}_{01}} \sup_{\boldsymbol{x}_i:\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\| \le \rho} \max\{0, 1 - \boldsymbol{w}^\top \boldsymbol{x}_i - b\} - \hat{p}_{01}^{-1}|\mathcal{I}_{01}|\right) \\
&= \begin{cases}
\min & \dfrac{1}{|\mathcal{I}_{11}|}\displaystyle\sum_{i\in\mathcal{I}_{11}} \lambda_i + \dfrac{1}{|\mathcal{I}_{01}|}\sum_{i\in\mathcal{I}_{01}} \lambda_i - 1 \\
\text{s.t.} & \boldsymbol{\lambda} \in \mathbb{R}_+^N, \\
& 1 + \boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + \rho\|\boldsymbol{w}\|_* + b \le \lambda_i \qquad \forall i \in \mathcal{I}_{11}, \\
& 1 - \boldsymbol{w}^\top \hat{\boldsymbol{x}}_i + \rho\|\boldsymbol{w}\|_* - b \le \lambda_i \qquad \forall i \in \mathcal{I}_{01},
\end{cases}
\end{aligned}$$

where the last equality follows by applying Lemma 3.4 twice and by noticing that $\mathcal{I}_{a1} \cap \mathcal{I}_{a'1} = \emptyset$. Setting the optimal value of the previous minimization problem to be less than η completes the proof. □

If $\|\cdot\|$ is either a 1-norm or an ∞-norm on $\mathbb{R}^d$, Problem (13) is a linear optimization problem. If $\|\cdot\|$ is

an Euclidean norm, Problem (13) becomes a second-order cone optimization problem. Both problems can be solved using off-the-shelf solvers such as MOSEK (MOSEK ApS 2024).

Remark 4.3 (Balanced Hinge Loss). For imbalanced data sets, one can also convexify the balanced misclassification rate using the hinge loss function. A brief discussion is provided in Online Appendix A.

We now benchmark the ε-DRFC model with the HDRFC model. The reformulation of the ε-DRFC problem (10) involves $2N$ binary variables and $2N$ big-M constraints. In contrast, the HDRFC reformulation (13) only contains $2N$ continuous variables and $2N$ convex constraints. HDRFC is more suitable for large instances because it is a continuous problem, a significant advantage compared with the ε-DRFC problem. The numerical results in Section 6 demonstrate that the hinge unfairness measure performs competitively.

5. Training with General Ground Metric

Previous sections have considered the absolute trust case of the ground cost (9) in which $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$. Here, we consider a general ground metric: For some finite values of $\kappa_{\mathcal{A}}$ and $\kappa_{\mathcal{Y}}$, we set

$$c((\boldsymbol{x}', a', y'), (\boldsymbol{x}, a, y)) = \|\boldsymbol{x} - \boldsymbol{x}'\| + \kappa_{\mathcal{A}}|a - a'| + \kappa_{\mathcal{Y}}|y - y'|. \tag{14}$$

The case for finite $\kappa_{\mathcal{A}}$ and $\kappa_{\mathcal{Y}}$ is particularly relevant when we have noisy observations of the sensitive attributes and class labels (Shafieezadeh-Abadeh et al. 2019). Without any loss of generality, we will illustrate how to incorporate this general ground metric using the ε-DRFC model (8). For the HDRFC model (12), we will provide the corresponding results in Online Appendix A. At the same time, we will consider a more general definition of the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ in this section. To this end, we first observe that the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ can be re-expressed as follows (a formal proof can be found in Online Appendix A):

$$\mathbb{B}(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) : \begin{array}{r} \exists \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \\ \forall i \in [N]: \\ \mathbb{Q} = N^{-1} \sum_{i \in [N]} \pi_i, \\ \mathbb{W}_\infty(\pi_i, \delta_{(\hat{\boldsymbol{x}}_i, \hat{a}_i, \hat{y}_i)}) \le \rho \\ \forall i \in [N], \\ \mathbb{Q}(A = a, Y = y) = \hat{p}_{ay} \\ \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\}.$$

Let $\gamma \in [0, 1]$ and consider the ambiguity set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ parametrized by γ as

$$\mathcal{B}_\gamma(\hat{\mathbb{P}}) \triangleq \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) : \begin{array}{r} \exists \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \\ \forall i \in [N]: \\ \mathbb{Q} = N^{-1} \sum_{i \in [N]} \pi_i, \\ \mathbb{W}_\infty(\pi_i, \delta_{(\hat{\boldsymbol{x}}_i, \hat{a}_i, \hat{y}_i)}) \le \rho \\ \forall i \in [N], \\ \mathbb{Q}(A = a, Y = y) = \hat{p}_{ay} \\ \forall (a, y) \in \mathcal{A} \times \mathcal{Y}, \\ \sum_{i \in [N]} \pi_i(A = \hat{a}_i, Y = \hat{y}_i) \\ \ge (1 - \gamma) N \end{array} \right\}. \tag{15}$$

Notice that $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ differs from $\mathbb{B}(\hat{\mathbb{P}})$ solely based on the last constraint defining $\mathcal{B}_\gamma(\hat{\mathbb{P}})$. Intuitively, the parameter γ indicates the maximum proportion of the training sample points that can be flipped in the (A, Y) dimension. When $\gamma = 1$, then the last constraint defining $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ collapses into

$$\sum_{i \in [N]} \pi_i(A = \hat{a}_i, Y = \hat{y}_i) \ge 0,$$

which holds trivially. Thus, we can deduce that $\mathcal{B}_1(\hat{\mathbb{P}}) = \mathbb{B}(\hat{\mathbb{P}})$. At the other extreme, when $\gamma = 0$, we arrive at the constraint

$$\sum_{i \in [N]} \pi_i(A = \hat{a}_i, Y = \hat{y}_i) \ge N \Rightarrow \pi_i(A = \hat{a}_i, Y = \hat{y}_i) = 1 \quad \forall i \in [N].$$

The latter constraint resembles the case in Sections 2 and 4 with absolute trust in the sensitive attribute and the label. Any value $\gamma \in (0, 1)$ thus can be considered an interpolation of the robustness condition between these two previously mentioned extreme cases.

We consider in this section the modified problem of (8) that uses the ambiguity set (15):

$$\begin{array}{lll} \min & \sup\limits_{\mathbb{Q} \in \mathcal{B}_\gamma(\hat{\mathbb{P}})} & \mathbb{Q}(Y(\boldsymbol{w}^\top \boldsymbol{X} + b) < \varepsilon) \\ \text{s.t.} & \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R},\ \|(\boldsymbol{w}, b)\| \le 1 & \\ & \sup\limits_{\mathbb{Q} \in \mathcal{B}_\gamma(\hat{\mathbb{P}})} & \mathbb{U}_\varepsilon(\boldsymbol{w}, b, \mathbb{Q}) \le \eta. \end{array} \tag{16}$$

We now present the main result of this section, which provides the reformulation for (16).

Theorem 5.1 (ε-DRFC Reformulation). *Suppose that the ground metric is prescribed using* (14). *For any* $\gamma \in (0,1)$, *Problem* (16) *is equivalent to the mixed binary conic program:*

$$
\begin{aligned}
\inf\ & \frac{1}{N}\sum_{i\in[N]} \nu_i + \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{ay}\mu_{ay} - \theta(1-\gamma) \\
\text{s.t.}\ & \boldsymbol{w}\in\mathbb{R}^d,\ b\in\mathbb{R},\ \boldsymbol{\nu}\in\mathbb{R}^N,\ \theta\in\mathbb{R}_+,\ \boldsymbol{\mu}\in\mathbb{R}^{2\times 2},\boldsymbol{\tau}\in\{0,1\}^N \\
& \|(\boldsymbol{w},b)\| \le 1, \\
& \boldsymbol{\nu}'\in\mathbb{R}^N,\ \theta'\in\mathbb{R}_+,\ \boldsymbol{\mu}'\in\mathbb{R}^{2\times 2},\boldsymbol{\lambda}^{\mathbf{1}}\in\{0,1\}^N,\ \boldsymbol{\lambda}^{\mathbf{0}}\in\{0,1\}^N \\
& \left.\begin{array}{l} \text{If } \kappa_{\mathcal{A}}|a-\hat{a}_i| + \kappa_{\mathcal{Y}}|y-\hat{y}_i| \le \rho: \\ \quad \tau_i \le \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y) + \nu_i, \\ \quad -\hat{y}_i(\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i + b) + (\rho - \kappa_{\mathcal{A}}|a-\hat{a}_i| \\ \qquad - \kappa_{\mathcal{Y}}|y-\hat{y}_i|)\|\boldsymbol{w}\|_* \le M\tau_i - \varepsilon \end{array}\right\} \\
& \qquad\qquad \forall i\in[N] \quad \forall (a,y)\in\mathcal{A}\times\mathcal{Y}, \\
& \left.\begin{array}{l} \text{If } \kappa_{\mathcal{A}}|1-\hat{a}_i| + \kappa_{\mathcal{Y}}|1-\hat{y}_i| \le \rho: \\ \quad \hat{p}_{11}^{-1}\lambda_i^1 \le \mu'_{1,1} - \theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,1) + \nu'_i, \\ \quad \boldsymbol{w}^\top\hat{\boldsymbol{x}}_i + (\rho - \kappa_{\mathcal{A}}|1-\hat{a}_i| - \kappa_{\mathcal{Y}}|1-\hat{y}_i|)\|\boldsymbol{w}\|_* \\ \qquad + b \le M\lambda_i^1 - \varepsilon \\ \text{If } \kappa_{\mathcal{A}}|0-\hat{a}_i| + \kappa_{\mathcal{Y}}|1-\hat{y}_i| \le \rho: \\ \quad \hat{p}_{01}^{-1}(\lambda_i^0 - 1) \le \mu'_{01} - \theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,1) + \nu'_i, \\ \quad -\boldsymbol{w}^\top\hat{\boldsymbol{x}}_i + (\rho - \kappa_{\mathcal{A}}|0-\hat{a}_i| - \kappa_{\mathcal{Y}}|1-\hat{y}_i|)\|\boldsymbol{w}\|_* \\ \qquad - b \le M\lambda_i^0 \\ \text{If } \kappa_{\mathcal{A}}|1-\hat{a}_i| + \kappa_{\mathcal{Y}}|-1-\hat{y}_i| \le \rho: \\ \quad 0 \le \mu'_{1,-1} - \theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,-1) + \nu'_i, \\ \text{If } \kappa_{\mathcal{A}}|0-\hat{a}_i| + \kappa_{\mathcal{Y}}|-1-\hat{y}_i| \le \rho: \\ \quad 0 \le \mu'_{0,-1} - \theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,-1) + \nu'_i, \end{array}\right\} \\
& \qquad\qquad \forall i\in[N], \\
& \frac{1}{N}\sum_{i\in[N]}\nu'_i + \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{ay}\mu'_{ay} - \theta'(1-\gamma) \le \eta,
\end{aligned}
\tag{17}
$$

where M is the big-M constant.

Reformulation (17) involves $3N$ binary variables and $6N$ big-M constraints. When $\|\cdot\|$ is either a 1-norm or an ∞-norm on $\mathbb{R}^d$, Problem (17) is a mixed binary linear optimization problem; when $\|\cdot\|$ is the Euclidean norm, Problem (17) becomes a mixed binary second-order cone optimization problem.

6. Numerical Experiments

In this section, we present the numerical experiments and examine the performance of different distributionally robust fair classifiers. Except for the DOB+ method (Donini et al. 2018) that is solved by an *sklearn* built-in solver and the DRFLR method (Taskesen et al. 2020) that is solved by MOSEK 10.0, all other optimization problems are implemented in Python 3.11 with package CVXPY 1.3.2 and solved by Gurobi 10.0.3 (Gurobi Optimization, LLC 2023). All experiments were run on a 2.2-GHz Intel Core i7 CPU laptop with 8 GB RAM.

6.1. Synthetic Experiments

In the first part of numerical experiments, we will use a synthetic data set to visually illustrate the performance of different fairness measures and the effect of introducing the DRO scheme. The data-generating distribution and the procedure of constructing this synthetic data set are presented in Online Appendix E. We use stratified sampling to obtain $N = 25$ samples as the training set and depict the classification hyperplanes determined by the ε-DRFC model (8) and the HDRFC model (12) on it. For each of the classifiers, we will plot three variants. The ε-C and HC classification hyperplanes are obtained by dropping the fairness constraints and setting the Wasserstein radius to zero. The ε-FC and the HFC classifiers include the fairness constraint but still without robustness consideration. The ε-DRFC and HDRFC models include both the fairness constraints and the robustification scheme. We choose the ground cost of the form (9) with $\|\cdot\|_*$ being the l_1-norm and $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$.

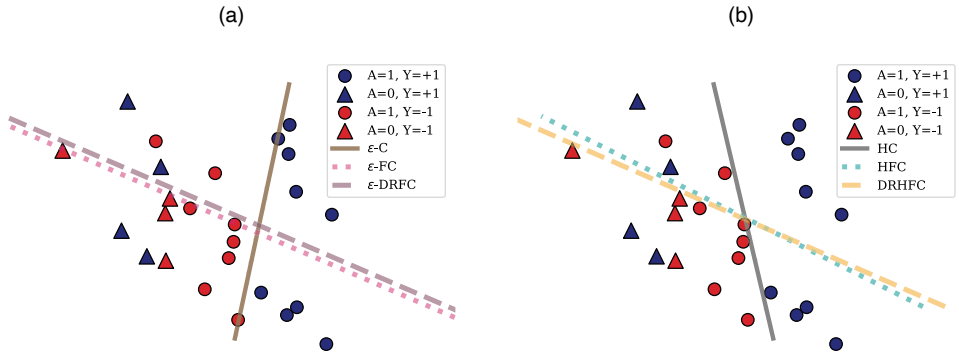
We first demonstrate how the unfairness constraints and the robustification influence the classifiers. The sensitive attribute A (represented by circles and triangles) is correlated with the feature X_1 on the horizontal axis. In Figure 2, (a) and (b), all of the four fairness aware classifiers (ε-FC, ε-DRFC, HFC, HDRFC) assign lower absolute value for the weight w_1 corresponding to feature X_1. Visually, this shift is reflected by the hyperplane determined by them becoming more horizontal compared with that of ε-C and HC. Moreover, by being robust, the ε-DRFC model shifts its hyperplane a bit higher to hedge against potential violations from disturbances, and the HDRFC model becomes even more horizontal to reduce the dependence of the classifiers on X_1.

We then assess the unfairness and accuracy scores on the training and testing sets. The unfairness score is evaluated by the absolute EO unfairness measure:

$$
\begin{aligned}
|\mathbb{U}(\boldsymbol{w},b,\mathbb{Q})| \triangleq\ & |\mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1 | A = 1, Y = 1) - \mathbb{Q}(\mathcal{C}(\boldsymbol{X}) \\
& = 1 | A = 0, Y = 1)|.
\end{aligned}
$$

Compared with the EO unfairness measure, the absolute EO unfairness measure is more suitable for performance

Figure 2. (Color online) Classification Hyperplanes (Dashed) Obtained by Different Approaches



Notes. Color encodes the labels and shape encodes the sensitive attributes. (a) Classification hyperplanes obtained by the mixed binary conic model. (b) Classification hyperplanes obtained by the convex model. (a) ε-DRFC. (b) HDRFC.

evaluation because it reflects if the trained fair classifier overdampens the classifier's performance on the privileged group in the test set. From Table 2, we observe that all the fairness aware classifiers reduce the unfairness score with a moderate cost of accuracy. In addition, although including robustness yields identical scores in the in-sample test, the out-of-sample performances improve significantly. With the distributionally robust model, the generated classifiers slightly deviate from the nonrobust classifiers to hedge against possible noises from the observed training samples, making the decisions more stable in the unseen testing set.

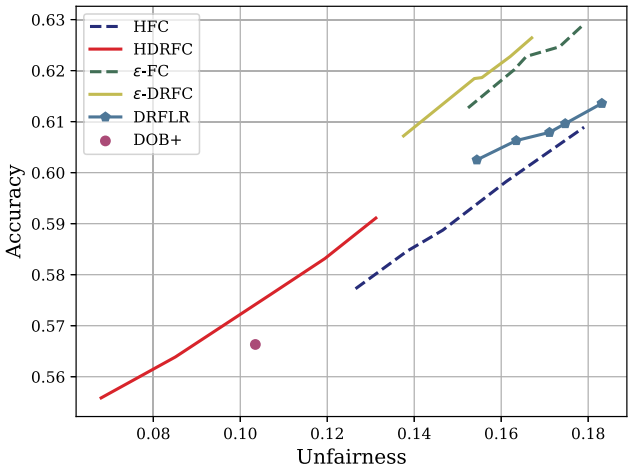
In the second set of synthetic experiments, we compare the performance of our models against the DOB+ (Donini et al. 2018) and DRFLR (Taskesen et al. 2020). The DOB+ model is the state-of-the-art method in deterministic linear classification. It minimizes the empirical hinge loss in the objective function and adopts a *linear-loss–based* unfairness measure to approximate the EO unfairness measure in the constraint. The DRFLR model is a distributionally robust logistic regression model. It minimizes the empirical *log-loss* and a fairness-driven regularization term in the objective function. Specifically, the paper proposes a *log-probabilistic equalized opportunities* unfairness measure, which is a convex approximation of the EO unfairness measure, as the fairness-driven regularization term. The DRFLR model is also considered the state-of-the-art method in distributionally robust fair logistic regression.

We plot the Pareto frontiers of the ε-FC, ε-DRFC, HFC, and HDRFC against those of DOB+ and DRFLR in Figure 3. We draw 200 samples from the well-known COMPAS data set (Brennan et al. 2009) and then separate them into a group of 50 samples used for the training, whereas the remaining 150 samples are used as the test set. For the ε-FC and ε-DRFC models, we examine the models with different values of the unfairness controlling parameter η on [0.05, 0.25] with five equidistant points. Similarly, we examine the HFC and HDRFC models with ζ on [1.2, 1.6] with five equidistant points, and the DRFLR model with η_f on $[0.1, \min\{\hat{p}_{01}, \hat{p}_{11}\}]$ as the DRFLR model admits tractable reformulations only if $\eta_f \leq \min\{\hat{p}_{01}, \hat{p}_{11}\}$. We fix the Wasserstein radius of the ε-DRFC and HDRFC models to 0.1 and the radius of the DRFLR model to 0.005. Because the authors of the DOB+ method argue that zero is a reasonable selection for the unfairness controlling parameter in their model, and their code is implemented under this prerequisite, to be consistent with their paper, we fix this parameter for the DOB+ method in our experiment. The hyperparameter C of the DOB+ method is chosen from $[10^{-1}, 10^{1}]$ by cross-validation using the authors'

Table 2. Predictive Accuracy and Unfairness on Training and Test Data for the Synthetic Experiment

Classifier	Train accuracy	Train unfairness ($\lvert\mathbb{U}\rvert$)	Test accuracy	Test unfairness ($\lvert\mathbb{U}\rvert$)
ε-C	84.00%	1.000	72.92%	0.9303
ε-FC	68.00%	0.056	58.56%	0.3560
ε-DRFC	68.00%	0.056	57.33%	0.3316
HC	84.00%	1.000	70.76%	0.9269
HFC	68.00%	0.056	56.92%	0.3284
HDRFC	68.00%	0.056	57.23%	0.3984

Figure 3. (Color online) Out-of-Sample Unfairness Accuracy Pareto Frontiers for Different Approaches



code. The described procedure is repeated 50 times independently, and the results are averaged over 50 trials.

Figure 3 visualizes the Pareto frontiers of six fairness aware models in the out-of-sample test, where the dashed lines represent the nonrobust models (ε-FC and HFC), the solid lines represent the distributionally robust models (ε-DRFC and HDRFC), and the dotted-solid line represents the DRFLR model. The ε-FC and ε-DRFC models benefited from their tight conservative approximation reformulation and dominate the DRFLR and HFC models across all unfairness scores. Additionally, the HDRFC model performs better than the DOB+ method, and because of its excellent scalability, it is more suitable for practical problems. Finally, all robust models (ε-DRFC and HDRFC) outperform their nonrobust counterparts (ε-FC and HFC), demonstrating the advantages of our proposed robustification schemes.

6.2. Experiments with Real Data

We then assess the performance of the ε-DRFC and HDRFC models and demonstrate their superior performances compared with competitive benchmarks. The experiments focus on five publicly available data sets (German, Adult, Drug, COMPAS, and Arrhythmia). The German Credit Risks data set classifies people described by a set of attributes as good or bad credit risks. The data are collected from 1,000 individuals, and we consider age (converted to binary values of "less than or equal to 30 years old" or "greater than 30 years old") as the sensitive attribute. The Adult data set is also relevant to candidates scoring in loan audits, where the prediction task is to determine if a person's annual income exceeds $50,000. It contains 13 features concerning demographic characteristics of 45,222 instances, and we consider gender as the sensitive attribute. The Drug and COMPAS data sets concern criminal assessment: The Drug data set includes 12 features of 1,885 respondents, and the objective is to predict whether a respondent has ever used heroin or not. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm judges and parole officers use to score criminal defendants' likelihood of reoffending. The data set contains variables used by the COMPAS algorithm in scoring defendants and their outcomes within two years of the decision for more than 10,000 criminal defendants. In both data sets, we consider ethnicity (Black versus non-Black) as the sensitive attribute. The Arrhythmia data set is related to medical interventions, where the task is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in 1 of the 16 groups. In our experiment, we consider gender as the sensitive attribute and reset the task with the binary classification between normal arrhythmia and 15 other arrhythmia classes.

A summary of these five data sets is presented in Table 3. Although the Adult data set has already been divided into the training and testing sets, we randomly select two-thirds of the samples for training and keep the rest of the data for testing in all other four data sets. It is worth noting that the Adult and Drug data sets contain many more negative samples, indicating that they are imbalanced.

Table 3. Data Sets Statistics and Their Sensitive Feature

Data set	Features d	Sensitive attribute A	Number of samples	Positive (+) vs. negative (−)
German	19	Age	1,000	40.0% : 60.0%
Adult	12	Gender	32,561, 12,661	24.9% : 75.1%
Drug	11	Ethnicity	1,885	21.9% : 78.1%
COMPAS	10	Ethnicity	6,172	47.0% : 53.0%
Arrhythmia	279(12)	Gender	452	54.2% : 45.8%

Notes. Age considers age groups greater than 30 years old and less than or equal to 30 years old. Gender considers the two groups male and female. Ethnicity considers the ethnic groups white and other ethnic groups. The adult data set has preassigned training and test sets. The last column represents the proportion of positive and negative samples in each data set.

Table 4. Test Accuracy, F_1 Score, and Different Unfairness Measures (Average ± Standard Deviation) for $N = 100$

Data set	Metric	SVM	DOB+	DRFLR	HDRFC	ε-DRFC
German	Accuracy	**0.71±0.02**	0.70 ± 0.02	0.70 ± 0.01	**0.71±0.01**	**0.71±0.01**
	F_1-score	0.80 ± 0.01	0.81 ± 0.02	0.79 ± 0.02	0.81 ± 0.01	**0.82±0.02**
	Unfairness ($\lvert\mathbb{U}\rvert$)	0.08 ± 0.04	0.05 ± 0.03	0.03 ± 0.02	**0.02±0.01**	0.03 ± 0.01
	Unfairness ($\lvert\mathbb{D}\rvert$)	0.09 ± 0.05	0.06 ± 0.03	0.03 ± 0.01	**0.01±0.01**	0.02 ± 0.01
	Unfairness ($\lvert\Phi\rvert$)	0.11 ± 0.05	0.08 ± 0.03	0.05 ± 0.02	**0.03±0.01**	**0.03±0.02**
COMPAS	Accuracy	**0.63±0.02**	0.59 ± 0.04	0.58 ± 0.03	0.58 ± 0.03	0.62 ± 0.03
	F_1-score	**0.58±0.02**	0.48 ± 0.02	0.46 ± 0.02	0.48 ± 0.01	**0.58±0.02**
	Unfairness ($\lvert\mathbb{U}\rvert$)	0.27 ± 0.05	0.17 ± 0.07	0.16 ± 0.06	**0.15±0.05**	0.17 ± 0.05
	Unfairness ($\lvert\mathbb{D}\rvert$)	0.23 ± 0.13	0.15 ± 0.07	**0.14±0.07**	**0.14±0.06**	0.16 ± 0.07
	Unfairness ($\lvert\Phi\rvert$)	0.29 ± 0.13	0.17 ± 0.07	0.17 ± 0.08	**0.16±0.08**	0.17 ± 0.08
Arrhythmia	Accuracy	**0.65±0.03**	0.63 ± 0.03	0.63 ± 0.02	0.62 ± 0.02	0.61 ± 0.02
	F_1-score	**0.72±0.02**	0.72 ± 0.02	0.71 ± 0.02	0.71 ± 0.01	0.71 ± 0.01
	Unfairness ($\lvert\mathbb{U}\rvert$)	0.22 ± 0.07	0.10 ± 0.08	0.08 ± 0.06	**0.06±0.04**	0.08 ± 0.05
	Unfairness ($\lvert\mathbb{D}\rvert$)	0.26 ± 0.11	0.15 ± 0.08	0.14 ± 0.07	0.11 ± 0.05	**0.10±0.05**
	Unfairness ($\lvert\Phi\rvert$)	0.26 ± 0.10	0.16 ± 0.07	0.14 ± 0.06	**0.12±0.05**	**0.12±0.05**
Adult	Accuracy	**0.79±0.03**	**0.79±0.02**	0.78 ± 0.02	0.72 ± 0.02	0.72 ± 0.02
	F_1-score	0.45 ± 0.02	0.36 ± 0.01	0.34 ± 0.02	0.52 ± 0.03	**0.53±0.02**
	Unfairness ($\lvert\mathbb{U}\rvert$)	0.21 ± 0.11	0.11 ± 0.08	**0.10±0.08**	0.15 ± 0.11	0.14 ± 0.10
	Unfairness ($\lvert\mathbb{D}\rvert$)	0.18 ± 0.13	**0.06±0.04**	0.07 ± 0.04	0.12 ± 0.07	0.12 ± 0.07
	Unfairness ($\lvert\Phi\rvert$)	0.22 ± 0.11	**0.11±0.08**	**0.11±0.08**	0.16 ± 0.11	0.15 ± 0.10
Drug	Accuracy	**0.79±0.03**	0.78 ± 0.02	0.78 ± 0.02	0.70 ± 0.02	0.71 ± 0.02
	F_1-score	0.42 ± 0.04	0.30 ± 0.01	0.26 ± 0.02	0.51 ± 0.02	**0.52±0.02**
	Unfairness ($\lvert\mathbb{U}\rvert$)	0.13 ± 0.08	0.08 ± 0.07	**0.07±0.06**	0.08 ± 0.06	0.09 ± 0.05
	Unfairness ($\lvert\mathbb{D}\rvert$)	0.07 ± 0.05	0.06 ± 0.03	**0.05±0.04**	0.06 ± 0.05	0.08 ± 0.05
	Unfairness ($\lvert\Phi\rvert$)	0.13 ± 0.06	**0.08±0.07**	**0.08±0.06**	0.11 ± 0.05	0.13 ± 0.06

Notes. The best results for each data set are highlighted in bold. For the imbalanced Adult and Drug data sets, we adopt balanced accuracy–driven objectives in the HDRFC and ε-DRFC models.

We formally benchmark the models following a cross-validation, training, and testing procedure. The hyperparameters of the HDRFC, and ε-DRFC models are determined following a cross-validation procedure similar to Donini et al. (2018). We first determine η and ζ under the empirical distribution. In each trial, we solve two plain vanilla classifiers to determine the privileged group and the empirical unfairness scores $\hat{\eta}$ and $\hat{\zeta}$. Then, we set η to half of the empirical EO unfairness score and ζ to half of the empirical Hinge unfairness score. With η and ζ being fixed, we tune the hyperparameters of the ambiguity set (15). We adopt the general ground metric (14) illustrated in Section 5 and set $\kappa_A = 2\kappa_Y = 0.5$, because the difference of the two sensitive attributes is $|a - a'| = 1$, whereas the difference of the two labels is $|y - y'| = 2$. We split the training set into a subtraining set with two-thirds of the samples while keeping the remainder as a subvalidation set. Then, we collect statistics of accuracy and absolute EO unfairness scores for $\rho \in [0.001, 1]$ on a logarithm scale with 30 discretization points evaluated on the subvalidation sets. The maximal value in the grid search for ρ equals $2\kappa_Y + \kappa_A$, which suffices to induce the perturbation on the label Y and the sensitive attribute A. Next, if the radius ρ obtained in the first step is greater than or equal to 0.5, then we fix it and tune γ from $\{0, 0.01, 0.02, \ldots, 0.05\}$; otherwise, we set γ to zero as the radius is less than the cost of perturbing the label Y or the sensitive attribute A. In this case, the ε-DRFC and HDRFC revert to the simplified models discussed in Sections 3 and 4, respectively. Finally, we select the values with the highest (Accuracy $- 0.5 \times$ Unfairness) score from the search grid. Similarly, the tuning parameters of the DOB+ and DRFLR methods are also determined by cross-validation using the authors' code. Next, we evaluate the accuracy and unfairness measures of all classifiers on the test set. We repeat this procedure for $K = 50$ times and report the average accuracy scores and unfairness measures in Table 4.

Table 4 suggests that our proposed HDRFC and ε-DRFC models perform favorably relative to their competitors. They yield low unfairness scores across the three balanced data sets with only a moderate loss in accuracy. For the imbalanced data sets, our proposed methods adopt balanced accuracy and balanced hinge loss as the objective functions. Benefiting from the modified objective functions, our methods perform well in all evaluation metrics: accuracy, F_1-score, and unfairness scores. As a comparison, the DOB+ and DRFLR methods work well for accuracy and unfairness but perform poorly in terms of F_1-score. The reason for getting low F_1 scores is that both methods are accuracy driven, which is a misleading metric for imbalanced data sets. To see this, let us illustrate using the Drug data set. Because the Drug data set contains 78% negative samples, the decision maker can easily design an

accurate (78% accuracy) and fair (0% positive rate for both groups) classifier by assigning all data points to the negative halfspace. However, this classifier is trivial since it does not identify qualified samples. Similar issues arise in other accuracy-driven methods. In the Adult and Drug data sets, the DOB+ and DRFLR models attempt to maintain fairness by naively rejecting most samples. Although the generated classifiers achieve deceivingly high scores in accuracy and unfairness, they inevitably fail in terms of F_1 score. In contrast, by minimizing the balanced misclassification rate and balanced hinge loss, our methods achieve much better F_1 scores in the imbalanced data set with only a slight loss in accuracy.

We also evaluate the performances of the aforementioned methods in terms of other unfairness measures. Specifically, we consider two popular fairness notions called demographic parity (Calders et al. 2009) and equalized odds (Hardt et al. 2016, Zafar et al. 2017). Similar to the EO unfairness measure, we define the *demographic parity unfairness measure* by

$$|\mathbb{D}(\boldsymbol{w}, b, \mathbb{Q})| \triangleq |\mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1|A = 1) - \mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1|A = 0)|,$$

and the *equalized odds unfairness measure* (Bird et al. 2020) by

$$\begin{aligned}|\mathbb{O}(\boldsymbol{w}, b, \mathbb{Q})| \triangleq \max\{&|\mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1|A = 1, Y = 1) \\ &- \mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1|A = 0, Y = 1)|, \\ &|\mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1|A = 1, Y = -1) \\ &- \mathbb{Q}(\mathcal{C}(\boldsymbol{X}) = 1|A = 0, Y = -1)|\}.\end{aligned}$$

Table 4 shows that promoting fairness in terms of equal opportunity will also improve demographic parity and equalized odds fairness scores, at least empirically. In addition, we also observe that the unfairness score using equalized odds always serves as an upper bound to the score using equal opportunity, which coincides with the definition that equal opportunity is a relaxation of the equalized odds criterion.

6.3. Solution Time

We report the running time of different methods on six data sets (German, Adult, Drug, COMPAS, Arrhythmia, and Synthetic) with the sample size varying from 25 to 1,000. We set the unfairness controlling parameters $\eta = 0.1$ for ε-FC and ε-DRFC, $\zeta = 1.1$ for HDRFC, $\eta_f = \min\{\hat{p}_{01}, \hat{p}_{11}\}/2$ for DRFLR, Wasserstein radius $\rho = 0.5$ for all distributionally robust models. The ε-DRFC and HDRFC are trained with the general ground metric with

Table 5. Running Time (s) of Different Methods

		Sample size N					
Data set	Classifier	25	50	100	250	500	1,000
German	ε-DRFC	1.87	2.36	13.66	148.25	3,432.71	/
	HDRFC	0.02	0.02	0.04	0.09	0.13	0.15
	DOB+	0.02	0.03	0.07	0.14	0.32	0.41
	DRFLR	2.53	3.61	8.70	20.18	44.32	80.62
Adult	ε-DRFC	3.56	17.91	265.47	/	/	/
	HDRFC	0.02	0.03	0.05	0.10	0.13	0.16
	DOB+	0.02	0.03	0.08	0.16	0.33	0.47
	DRFLR	3.02	3.79	8.01	21.59	40.52	85.90
Drug	ε-DRFC	4.27	26.73	1,072.14	/	/	/
	HDRFC	0.02	0.03	0.05	0.08	0.11	0.15
	DOB+	0.03	0.03	0.07	0.16	0.21	0.34
	DRFLR	2.58	3.66	7.13	20.57	43.69	93.42
COMPAS	ε-DRFC	1.42	3.72	15.26	122.49	3,749.28	/
	HDRFC	0.03	0.04	0.05	0.11	0.14	0.18
	DOB+	0.02	0.03	0.02	0.15	0.18	0.17
	DRFLR	3.04	4.10	7.48	19.32	45.64	91.57
Arrhythmia	ε-DRFC	2.31	4.58	22.40	626.50		
	HDRFC	0.03	0.06	0.24	0.39		
	DOB+	0.04	0.05	0.17	0.32		
	DRFLR	2.12	4.37	11.59	21.28		
Synthetic	ε-DRFC	0.43	0.91	2.78	10.57	147.84	4,741.72
	HDRFC	0.12	0.01	0.02	0.03	0.05	0.08
	DOB+	0.09	0.16	0.19	0.29	0.38	0.59
	DRFLR	2.56	3.84	8.02	19.91	42.54	91.70

Notes. The Arrhythmia data set only contains 452 examples. Hence, we examine its performance up to $N = 250$. The / symbol represents that the solver fails to achieve optimality within 7,200 seconds.

$\gamma = 0.01$, $\kappa_A = 2\kappa_Y = 0.5$. All results are averaged over 10 independent trials.

Table 5 suggests that the ε-DRFC model is applicable to moderate-size problems. However, it encounters computational difficulties at large sample sizes. In addition, the model becomes even more computationally intensive for the imbalanced Adult and Drug data sets. The DRFLR model involves solving an exponential cone program, and we use the popular cone optimization solver MOSEK (MOSEK ApS 2024) to solve the problem. We observe that the DRFLR model is less efficient than the linear-program-based method HDRFC and the gradient-descent-based method DOB+. The sample size is the factor that affects the running time the most because the number of variables and constraints is proportional to the sample size. Compared with the ε-DRFC and DRFLR models, the HDRFC and the DOB+ methods are more efficient across all data sets. For all sample sizes, these methods can be solved in one second. Therefore, this result suggests that the HDRFC model is more suitable for large instances.

7. Concluding Remarks

In this paper, we developed a new principled approach to fair classification by incorporating the equality of opportunity criterion as a constraint and robustifying the resulting optimization problem using the framework of Wasserstein min-max learning. We use the type-∞ Wasserstein ambiguity set, which enables a scalable conic programming reformulation with attractive statistical performance guarantees. Our proposed model can also handle problem instances with noisy adversarially sensitive attributes and labels.

Because the original problem cannot be reformulated exactly, we propose a tight conservative Approximation (8) that is amenable to a mixed binary linear programming reformulation. To the best of our knowledge, this is the first approximation that enables decision-makers to control the worst-case EO unfairness measure explicitly using a constraint formulation. However, this reformulation is not as efficiently solvable due to the number of binary variables growing polynomially with the number of data samples. To address this issue, we further approximate both the objective function and the unfairness measure using the hinge loss function to obtain a convex continuous approximation. We find that the hinge-loss–based distributionally robust fairness aware model performs favorably compared with the state-of-the-art method DOB+ and DRFLR in the numerical experiments.

References

Agarwal A, Beygelzimer A, Dudik M, Langford J, Wallach H (2018) A reductions approach to fair classification. *Internat. Conf. Machine Learn.* (PMLR, New York), 60–69.

Aliprantis CD, Border KC (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide* (Springer, Berlin).

Angwin J, Larson J, Mattu S, Kirchner L (2022) Machine bias. *Ethics of Data and Analytics* (Auerbach Publications, Boca Raton, FL), 254–264.

Baardman L, Boroujeni SB, Cohen-Hillel T, Panchamgam K, Perakis G (2023) Detecting customer trends for optimal promotion targeting. *Manufacturing Service Oper. Management* 25(2):448–467.

Bach F (2021) *Learning Theory from First Principles*. Draft of a book, version of, September 6, 2021, https://www.di.ens.fr/~fbach/ltfp_book.pdf.

Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, et al. (2018) AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Preprint, submitted October 3, https://arxiv.org/abs/1810.01943.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 50(1):3–44.

Bertsimas D, Shtern S, Sturt B (2018) A data-driven approach to multistage linear optimization. *Optimization Online* (November 3), https://optimization-online.org/2018/11/6907/.

Bertsimas D, Shtern S, Sturt B (2022) Two-stage sample robust optimization. *Oper. Res.* 70(1):624–640.

Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, et al. (2020) Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report, Microsoft, Redmond, WA.

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.

Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probabilities* 56(3):830–857.

Bose I, Mahapatra RK (2001) Business data mining—A machine learning perspective. *Inform. Management* 39(3):211–225.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice Behav.* 36(1):21–40.

Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. *Proc. IEEE Internat. Conf. Data Mining Workshops* (IEEE, Piscataway, NJ), 13–18.

Chang L (2006) Applying data mining to predict college admissions yield: A case study. *New Directions Institutional Res.* 131:53–68.

Chapelle O, Teo C, Le Q, Smola A (2008) Tighter bounds for structured estimation. Koller D, Schuurmans D, Bengio Y, Bottou L, eds. *Adv. Neural Inform. Processing Systems*, vol. 21 (Curran Associates, Inc., Red Hook, NY).

Chen X, Owen Z, Pixton C, Simchi-Levi D (2022) A statistical learning approach to personalization in revenue management. *Management Sci.* 68(3):1923–1937.

Chouldechova A, Roth A (2020) A snapshot of the frontiers of fairness in machine learning. *Comm. ACM* 63(5):82–89.

Consumer Financial Protection Bureau (2013) CFPB and DOJ order ally to pay $80 million to consumers harmed by discriminatory auto loan pricing. Accessed October 21, 2020, https://www.consumerfinance.gov/about-us/newsroom/cfpb-and-doj-order-ally-to-pay-80-million-to-consumers-harmed-by-discriminatory-auto-loan-pricing.

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 797–806.

Dastin J (2022) Amazon scraps secret AI recruiting tool that showed bias against women. *Ethics of Data and Analytics* (Auerbach Publications, Boca Raton, FL), 296–299.

Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proc. Privacy Enhancing Tech.* 2015(1):92–112.

Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M (2018) Empirical risk minimization under fairness constraints. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 31 (Curran Associates, Inc., Red Hook, NY). 2791–2801.

Dua D, Graff C (2017) UCI machine learning repository. Accessed October 15, 2020, http://archive.ics.uci.edu/ml.

Dyer ME, Frieze AM (1988) On the complexity of computing the volume of a polyhedron. *SIAM J. Comput.* 17(5):967–974.

Gao R, Kleywegt AJ (2023) Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* 48(2):603–655.

Givens C, Shortt R (1984) A class of Wasserstein metrics for probability distributions. *Michigan Math. J.* 31(2):231–240.

Golrezaei N, Nazerzadeh H, Rusmevichientong P (2014) Real-time optimization of personalized assortments. *Management Sci.* 60(6): 1532–1551.

Gurobi Optimization, LLC (2023) *Gurobi Optimizer Reference Manual* (Gurobi Optimizer, Beaverton, OR).

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 29 (Curran Associates, Inc., Red Hook, NY), 3315–3323.

Hashimoto T, Srivastava M, Namkoong H, Liang P (2018) Fairness without demographics in repeated loss minimization. *Proc. 35th Internat. Conf. Machine Learn.*, 1929–1938.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning* (Springer, Berlin).

Ho-Nguyen N, Wright SJ (2023) Adversarial classification via distributional robustness with Wasserstein ambiguity. *Math. Programming* 198(2):1411–1447.

Jacobson T, Roszbach K (2003) Bank lending policy, credit scoring and value-at-risk. *J. Bank. Finance* 27(4):615–633.

Jeroslow RG (1987) Representability in mixed integer programming, I: Characterization results. *Discrete Appl. Math.* 17(3):223–243.

Kabakchieva D (2013) Predicting student performance by using data mining methods for classification. *Cybernetics Inform. Tech.* 13(1):61–72.

Kuhn D, Mohajerin Esfahani P, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research* (INFORMS, Catonsville, MD), 130–166.

Liittschwager J, Wang C (1978) Integer programming solution of a classification problem. *Management Sci.* 24(14):1515–1525.

Lohr S (2013) Big data, trying to build better workers. *The New York Times* (April 21).

Mak H-Y, Rong Y, Zhang J (2014) Appointment scheduling with limited distributional information. *Management Sci.* 61(2):316–334.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM Comput. Surveys (CSUR)* 54(6):1–35.

Mišić VV, Perakis G (2020) Data analytics in operations management: A review. *Manufacturing Service Oper. Management* 22(1):158–169.

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1–2):115–166.

Monahan J, Skeem JL (2016) Risk assessment in criminal sentencing. *Annu. Rev. Clinical Psych.* 12:489–513.

MOSEK ApS (2024) Mosek optimizer API for python. Version 10. Accessed April 11, 2024, https://docs.mosek.com/10.1/pythonapi.pdf.

Nguyen VA, Kuhn D, Mohajerin Esfahani P (2022) Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Oper. Res.* 70(1):490–515.

Nguyen VA, Zhang F, Blanchet J, Delage E, Ye Y (2020) Distributionally robust local non-parametric conditional estimation. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 15232–15242.

Obermeyer Z, Emanuel EJ (2016) Predicting the future—Big data, machine learning, and clinical medicine. *New England J. Medicine* 375(13):1216.

Quadrianto N, Sharmanska V (2017) Recycling privileged learning and distribution matching for fairness. Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 30 (Curran Associates, Inc., Red Hook, NY), 677–688.

Rezaei A, Fathony R, Memarrast O, Ziebart B (2020) Fairness for robust log loss classification. *Proc. AAAI Conf. Artificial Intelligence.*

Rudin W (1964) *Principles of Mathematical Analysis*, vol. 3 (McGraw-Hill, New York).

Samorani M, Harris SL, Blount LG, Lu H, Santoro MA (2022) Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manufacturing Service Oper. Management* 24(6):2825–2842.

Shafieezadeh-Abadeh S, Kuhn D, Mohajerin Esfahani P (2019) Regularization via mass transportation. *J. Machine Learn. Res.* 20(103): 1–68.

Shalev-Shwartz S, Ben-David S (2014) *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, Cambridge, UK).

Shaw MJ, Gentry JA (1988) Using an expert system with inductive learning to evaluate business loans. *Financial Management* 17(3):45–56.

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8(1):68–74.

Taskesen B, Blanchet J, Kuhn D, Nguyen VA (2021) A statistical test for probabilistic fairness. *Proc. ACM Conf. Fairness Accountability Transparency* (ACM, New York), 648–665.

Taskesen B, Nguyen VA, Kuhn D, Blanchet J (2020) A distributionally robust approach to fair classification. Preprint, submitted July 18, https://arxiv.org/abs/2007.09530.

Vapnik V, Vashist A (2009) A new learning paradigm: Learning using privileged information. *Neural Networks* 22(5–6): 544–557.

Wang S, Guo W, Narasimhan H, Cotter A, Gupta M, Jordan M (2020) Robust optimization for fairness with noisy protected groups. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Adv. Neural Inform. Processing Systems* 33 (Curran Associates, Inc., Red Hook, NY), 5190–5203.

Xie W (2020) Tractable reformulations of two-stage distributionally robust linear programs over the type ∞ Wasserstein ball. *Oper. Res. Lett.* 48(4):513–523.

Ye Q, Xie W (2020) Unbiased subdata selection for fair classification: A unified framework and scalable algorithms. Preprint, submitted December 24, https://arxiv.org/abs/2012.12356.

Yurochkin M, Bower A, Sun Y (2020) Training individually fair ML models with sensitive subspace robustness. *Proc. 8th Internat. Conf. Learn. Representations* (Curran Associates, Inc., Red Hook, NY).

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proc. 26th Internat. Conf. World Wide Web* (ACM, New York), 1171–1180.