arXiv:2103.06828v1 [cs.LG] 11 Mar 2021

WASSERSTEIN ROBUST SUPPORT VECTOR MACHINES WITH FAIRNESS CONSTRAINTS

YIJIE WANG, VIET ANH NGUYEN, AND GRANI A. HANASUSANTO

Abstract. We propose a distributionally robust support vector machine with a fairness constraint that encourages the classifier to be fair in view of the equality of opportunity criterion. We use a type-∞ Wasserstein ambiguity set centered at the empirical distribution to model distributional uncertainty and derive an exact reformulation for worst-case unfairness measure. We establish that the model is equivalent to a mixed-binary optimization problem, which can be solved by standard off-the-shelf solvers. We further prove that the expectation of the hinge loss objective function constitutes an upper bound on the misclassification probability. Finally, we numerically demonstrate that our proposed approach improves fairness with negligible loss of predictive accuracy.

1. Introduction

Machine learning algorithms is increasingly deployed to support consequential decision-making processes, from deciding which applicants will receive the job offers [41, 18], loans [11, 56], university enrollments [15, 35], or to medical intervention [57, 50]. Even though these algorithms can effectively solve large-scale problems, they may not be entirely objective and can be susceptible to amplify human biases. For example, it was found that the hiring recommendation system of Amazon AI discriminated against female candidates for technical positions [18]. Similarly, Google's ad-targeting algorithm had recommended higher-paying executive jobs more often to male than to female candidates [19]. It has also been shown that an algorithm used by the US justice system to predict future criminals is significantly biased against African Americans—where it falsely flags black defendants as future criminals at almost twice the rate of white defendants [3].

The amplification of human bias caused by algorithms has sparked the emerging field of algorithmic fairness. Strategies to promote fairness in machine learning can be divided into three main categories. The first category includes proposals to *pre-process* the training data before solving a plain-vanilla machine learning problem [14, 28, 24, 36, 42, 54, 69]. The second category includes *post-processing* approaches applied to a pre-trained classifier in order to increase its fairness properties while retaining to the largest extent as possible the predictive power of the learned algorithms [17, 21, 29, 45]. The third category of strategies aims to enforce fairness in the training process by modelling explicitly fairness constraints to the learning problem [20, 45, 62, 64, 66, 68], by penalizing discrimination using fairness-driven regularization terms [4, 34, 37, 38] or by (approximately) penalizing any mismatches between the true positive rates and the false negative rates across different groups [5]. Adversarial training to promote algorithmic fairness have also been used and shown to deliver promosing results [23, 26, 31, 39, 43, 52, 65, 70].

The method we propose in this paper can be viewed as an adversarial approach pertaining to the third category. More specifically, we consider the training problem of support vector machines (SVM), which is arguably one of the most popular classification methods in the statistical learning literature [32]. SVM aims to establish a deterministic relationship between a feature vector $X \in \mathcal{X} = \mathbb{R}^d$ and a binary response, or label, variable $Y \in \mathcal{Y} = \{-1, 1\}$. Without any loss of generality, we associate the positive response with the "advantaged" outcome, such as "being hired" or "receiving a loan approval." We also assume that there is

Date: March 12, 2021.

The authors are with the Graduate Program in Operations Research and Industrial Engineering, University of Texas at Austin (`yijie-wang, grani.hanasusanto@utexas.edu`) and the Department of Management Science and Engineering, Stanford University (`viet-anh.nguyen@stanford.edu`).

a single 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 from the unprivileged individuals. Throughout this paper, we assume that we possess a training data set containing N samples of the form $\{(\hat{x}_i, \hat{a}_i, \hat{y}_i)\}_{i=1}^N$, and these samples are generated independently from a single data-generating probability distribution. In the SVM setting, a classifier $\mathcal{C} : \mathcal{X} \to \mathcal{Y}$ is parameterized by a slope parameter $w \in \mathbb{R}^d$ and an offset $b \in \mathbb{R}$, and the classification output is determined through an indicator function of the form

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

Throughout, we consider the privileged learning setting [60, 51] in which the sensitive information is only available at the training stage but not at the testing stage. Thus, it is reasonable to consider classifiers $\mathcal{C}$ that does not take the sensitive attribute as input. An SVM classifier $\mathcal{C}$ in this form also satisfies the requirement of disparate treatment.

SVM is obtained by determining the parameters $(w, b) \in \mathbb{R}^{d+1}$ that solves the tractable convex optimization problem

$$\min_{w,b} \ \frac{1}{N} \sum_{i=1}^N \ell_{(\hat{x}_i, \hat{y}_i)}(w, b), \quad \ell_{(\hat{x}_i, \hat{y}_i)}(w, b) = \max\{0, 1 - \hat{y}_i(w^\top \hat{x}_i + b)\}, \tag{1}$$

which minimizes the *hinge loss* applied to the training data. To make SVM fair, we can incorporate a measure of fairness into problem (1), either in the form of a constraint or in the form of an objective regularization. There are a plethora of fairness measure that we can utilize to promote fairness in this case, including the equality of opportunity [30], demographic parity [13], and equalized odds [30, 67] among many others. We refer the reader to the references [6, 16, 17, 44] for comprehensive treatments of fairness in machine learning in general and in the classification problem in particular.

The existing notions of fairness proposed in the literature are both conceptually and computationally difficult to evaluate in practice. Conceptually, these fairness notions necessitate precise knowledge about the joint probability distribution that governs (X, A, Y). This distribution is never available to the decision-makers and is typically estimated using the empirical distribution generated from the imbalanced—and possibly biased—historical observations. While the empirical-based methods may work well on the observed data set, they often fail to yield complete 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. On the other hand, even if the true underlying distribution is available, computing the fairness of the decision is generically intractable (#P-hard [22]) because it involves evaluating a multi-dimensional integration (e.g., computing the probability of getting hired conditionally on being an unprivileged person).

Fundamentally, promoting fairness in machine learning algorithms needs to balance among conflicting objectives including predictive accuracy, fairness, computational efficiency, while at the same time having to deal with mismatches between training and test data. In this paper, we endeavor to explore the trade-offs between these objectives using a modern methodology for decision-making under uncertainty called *distributionally robust optimization (DRO)*. DRO is a fundamentally truthful method that does not make any distributional assumption about the features, the 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 in view of the most adverse distribution from within the prescribed distribution set. This approach thus yields a fair decision that also ascertains provable guarantees on the out-of-sample data sets.

Contributions. The contributions of this paper can be summarized as follows.

- **Worst-case equality of opportunity:** We propose a new unfairness criterion obtained from robustifying the classical equality of opportunity notion. Specifically, we employ the data-driven distributionally robust model and consider the worst-case unfairness under the most adverse distribution from within the type-∞ Wasserstein ambiguity set. If the radius of the ambiguity set vanishes to zero, our formulation recovers the unfairness measure evaluated at the empirical distribution. As such, our proposed conservative estimate can principally be leveraged as a regularization of the empirical-based method.
- **Distributionally robust fair SVM:** We incorporate the unfairness criterion into the distributionally robust SVM model and establish that the resulting problem is equivalent to a concise mixed-binary optimization program. Experimental tests demonstrate that our proposed classifiers generate a marked improvement in terms of fairness, with a negligible loss of predictive accuracy.
- **Relationship between SVM and misclassification probability:** We prove that for any fixed data-generating distribution the expectation of the hinge loss function constitutes an upper bound on the misclassification probability. To the best of our knowledge, we are the first to reveal this salient relationship between these two quantities. In conjunction with the concentration bounds for the type-∞ Wasserstein ambiguity set, this result implies that the resulting classifiers can deliver strong out-of-sample accuracy and fairness guarantees.

In general, adding a bound on the unfairness measures leads to a non-convex constraint. Convex approximations were first deployed to enhance model scalability [1, 20], and recently, exact reformulation of unfairness measures were also developed for the purpose of quantifying unfairness exactly [64]. In this paper, we take the latter approach and we aim to quantify the unfairness measures in an exact manner.

Our paper belongs to an emerging class of fairness-aware distributionally robust algorithms. Previously, a repeated loss minimization model with a χ^2-divergence ambiguity set is considered in [31]. Alternatively, [52] embed the fairness constraint in the ambiguity set and propose a robust classification model, while robust fairness constraints based on a total variation ambiguity set is described in [61]. Wasserstein distributionally robust classification is also proposed to promote individual fairness [65], or to train a log-probabilistic fair logistic classifier [59]. Our paper is also closely related to the literature on Wasserstein min-max statistical learning, which connects to various forms of regularization (e.g., norm [9, 55]; shrinkage [48]). Our formulation considers adversarial perturbations based on the Wasserstein distance [10, 25, 33, 40, 46]. In particular, the type-∞ Wasserstein distance [27] is recently applied in distributionally robust formulations [7, 8, 49, 63].

The paper is organized as follows. Section 2 delineates our fairness-aware distributionally robust SVM training problem. Section 3 and 4 provide the finite-dimensional, binary optimization reformulation of the training problem for two instances of the ground transportation with/without absolute trust on the sensitive attribute and label. Section 5 discusses some performance guarantee and Section 6 reports on the numerical experiments.

Notations. For any set $\mathcal{S}$, we use $\mathcal{M}(\mathcal{S})$ to denote the set of probability measures supported on $\mathcal{S}$. 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.

2. Fairness-aware Distributionally Robust Support Vector Machines

Throughout this section, we focus on promoting fairness in SVM with respect to the criterion of equal opportunity, or also known as, equality of opportunity [29]. This criterion is formally defined as follows.

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

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

We say that a classifier is *trivial* if it is parametrized by $(w, b) = (0, 0) \in \mathbb{R}^{d+1}$. In this case, $\mathcal{C}(x) = 1$ for any input $x \in \mathcal{X}$. It is easy to verify that the trivial classifier is also fair with respect to any possible distribution $\mathbb{Q}$. Our goal in this paper is to search for a *non-trivial* classifier that strikes a balance between promoting fairness and achieving superior predictive power. To this end, suppose that $\mathbb{P} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$ is the data-generating distribution of the joint random vector (X, A, Y). The fair support vector machine (SVM) solves the stochastic optimization problem

$$\begin{array}{ll} \min & \mathbb{E}_{\mathbb{P}}[\max\{0, 1 - Y(w^\top X + b)\}] \\ \text{s.t.} & w \in \mathbb{R}^d, \ b \in \mathbb{R} \\ & \mathbb{P}(w^\top X + b \geq 0 | A = 1, Y = 1) = \mathbb{P}(w^\top X + b \geq 0 | A = 0, Y = 1). \end{array} \tag{2}$$

The objective function of (2) minimizes the expected hinge loss, while the constraint of (2) imposes that the linear classifier (w, b) satisfies the equal opportunity criterion with respect to $\mathbb{P}$. Unfortunately, the data-generating distribution $\mathbb{P}$ is elusive to the decision maker. Even if $\mathbb{P}$ is known, the probabilistic program (2) is computationally intractable.[1] In a data-driven setting, we assume that we can access to N training samples generated from $\mathbb{P}$. Let $\hat{\mathbb{P}}$ be the empirical distribution supported on $\{(\hat{x}_i, \hat{a}_i, \hat{y}_i)\}_{i=1}^N$. We will construct an ambiguity set around $\hat{\mathbb{P}}$ using the Wasserstein distance.

Definition 2.2 (Wasserstein distance)**.** Let c be a metric on Ξ. The type-p $(1 \leq p < +\infty)$ Wasserstein distance between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is defined as

$$\mathbb{W}_p(\mathbb{Q}_1, \mathbb{Q}_2) \triangleq \inf\left\{ \left(\mathbb{E}_\pi[c(\xi_1, \xi_2)^p]\right)^{\frac{1}{p}} : \pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2) \right\},$$

where $\Pi(\Xi \times \Xi)$ is the set of all probability measures on $\Xi \times \Xi$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively. The type-∞ Wasserstein distance is defined as the limit of $\mathbb{W}_p$ as p tends to ∞ and amounts to

$$\mathbb{W}_\infty(\mathbb{Q}_1, \mathbb{Q}_2) \triangleq \inf\left\{ \operatorname*{ess\,sup}_\pi \left\{ c(\xi_1, \xi_2) : (\xi_1, \xi_2) \in \Xi \times \Xi \right\} : \pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2) \right\}.$$

We let $\Xi = \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ be the joint outcome space of the covariate, the sensitive attribute and the label. The ground metric on Ξ is supposed to be separable, that means c can be written as a sum of three components as

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

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

$$\mathbb{B}(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) : \begin{array}{l} \mathbb{W}_\infty(\mathbb{Q}, \hat{\mathbb{P}}) \leq \rho \\ \mathbb{Q}(A = a, Y = y) = \hat{p}_{ay} \quad \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\}, \tag{3}$$

which is a neighborhood around the empirical distribution $\hat{\mathbb{P}}$. Intuitively, $\mathbb{B}(\hat{\mathbb{P}})$ contains all the distributions of (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 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 made 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 fairness [59]. Indeed, conditional expectation is typically a non-linear function of the probability measure. However, when confining inside the set $\mathbb{B}(\hat{\mathbb{P}})$, we have

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

which becomes linear in $\mathbb{Q}$ and conveniently simplifies the problem.

[1] Formally, the problem of computing the probability of an event involving multiple random variables belongs to the complexity class #P-hard [22]—which is perceived to be 'harder' than the class NP-hard.

Equipped with the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$, we can consider the fairness-aware distributionally robust SVM model

$$\begin{array}{ll} \min & \sup\limits_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(w^\top X + b)\}] \\ \text{s.t.} & w \in \mathbb{R}^d, b \in \mathbb{R} \\ & \sup\limits_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} |\mathbb{Q}(w^\top X + b \geq 0 | A = 1, Y = 1) - \mathbb{Q}(w^\top X + b \geq 0 | A = 0, Y = 1)| \leq \eta. \end{array} \tag{4}$$

The constraint of problem (4) depends on a tolerance $\eta \in \mathbb{R}_+$: it requires that the difference between the correct positive classification rate in two groups $A = 0$ and $A = 1$ to be uniformly bounded below η. It is easy to verify that the trivial classifier with $(w, b) = (0, 0)$ is feasible for (4) with an objective value of 1.

For 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 $\mathbb{Q}_{ay}(X \in \cdot) = \mathbb{Q}(X \in \cdot | A = a, Y = y)$. For mathematical rigor, we consider the following modification of the problem. Define the *un*fairness measure $\mathbb{U}$

$$\mathbb{U}(w, b, \mathbb{Q}) \triangleq \max \left\{ \begin{array}{l} \mathbb{Q}_{01}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{11}(w^\top X + b \geq 0), \\ \mathbb{Q}_{11}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{01}(w^\top X + b \geq 0) \end{array} \right\} \tag{5}$$

for some strictly positive value $\varepsilon \in \mathbb{R}_{++}$ and consider the conservative fairness-aware distributionally robust SVM problem

$$\begin{array}{ll} \min & \sup\limits_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(w^\top X + b)\}] \\ \text{s.t.} & w \in \mathbb{R}^d, b \in \mathbb{R} \\ & \sup\limits_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{U}(w, b, \mathbb{Q}) \leq \eta. \end{array} \tag{6}$$

By definition of $\mathbb{U}$, we find

$$|\mathbb{Q}(w^\top X + b \geq 0 | A = 1, Y = 1) - \mathbb{Q}(w^\top X + b \geq 0 | A = 0, Y = 1)| \leq \mathbb{U}(w, b, \mathbb{Q}),$$

for every possible value of the classifier parameter (w, b) and any distribution $\mathbb{Q}$. As a consequence, problem (6) is a conservative approximation of problem (4). This implies that the optimal solution $(w^\star, b^\star)$ of problem (6) is also feasible for problem (4). The next result shows that problem (6) is well-defined, in the sense that its feasible set contains a non-trivial classifier.

Lemma 2.3 (Feasibility)**.** For any $\eta \in \mathbb{R}_+$, there exists a non-trivial classifier that is feasible for problem (6).

Proof of Lemma 2.3. It suffices to show that problem (6) is feasible for $\eta = 0$. Let us consider a hyperplane parameterized by $(w_s, b_s) \in \mathbb{R}^{d+1}$ such that $w_s^\top x + b_s > 0$ for all $x \in \mathbb{X}$, where $\mathbb{X}$ is defined as in Lemma 8.1. Because $\mathbb{X}$ is compact and convex, the existence of the hyperplane (w_s, b_s) is a direct result of the separating hyperplane theorem [12, §2.5.1]. In this case, one can verify that for any $\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})$, we have

$$\begin{aligned} \mathbb{Q}_{11}(w_s^\top X + b_s > -\varepsilon) - \mathbb{Q}_{01}(w_s^\top X + b_s \geq 0) = 1 - 1 = 0 \\ \mathbb{Q}_{01}(w_s^\top X + b_s > -\varepsilon) - \mathbb{Q}_{11}(w_s^\top X + b_s \geq 0) = 1 - 1 = 0, \end{aligned}$$

where $\mathbb{Q}_{ay}$ are the conditional distributions of $X | A = a, Y = y$. This implies that

$$\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{U}(w_s, b_s, \mathbb{Q}) \leq 0,$$

and thus (w_s, b_s) is feasible for problem (6) at $\eta = 0$. This completes the proof. □

For any $\varepsilon \in \mathbb{R}_{++}$, problem (6) is easier to solve compared to, and will be the focus of this paper.

3. Training with Absolute Trust in Sensitive Attributes and Labels

Throughout this section, we use the following ground metric

$$c\big((x', a', y'), (x, a, y)\big) = \|x - x'\| + \infty|a - a'| + \infty|y - y'| \tag{7}$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^d$. Notice that in this setting, we have 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 (7), a simple modification of the proof of [58, 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 $(a, y) \in \mathcal{A} \times \mathcal{Y}$. As a consequence, the marginal constraint in the definition of 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 [58] and to train fair logistic classifier [59].

Next, we present the main result of this section which asserts that the conservative fairness-aware DR-SVM training problem (6) can be reformulated as a mixed binary optimization problem.

Theorem 3.1 (Reformulation)**.** Suppose that the ground metric is prescribed using (7), then the conservative fairness-aware DR-SVM problem (6) is equivalent to the conic mixed binary optimization problem

$$\begin{array}{lll}
\min & \displaystyle\frac{1}{N}\sum_{i=1}^{N} t_i & \\
\text{s.t.} & w \in \mathbb{R}^d,\ b \in \mathbb{R},\ t \in \mathbb{R}^N_+,\ \lambda^0 \in \{0,1\}^N,\ \lambda^1 \in \{0,1\}^N & \\
& -\hat{y}_i(w^\top \hat{x}_i + b) + \rho\|w\|_* \le t_i - 1 & \forall i \in [N] \\
& \left.\begin{array}{l}
\displaystyle\frac{1}{N}\left(\hat{p}_{a1}^{-1}\sum_{i\in\mathcal{I}_{a1}}\lambda_i^a + \hat{p}_{a'1}^{-1}\sum_{i\in\mathcal{I}_{a'1}}\lambda_i^a - \hat{p}_{a'1}^{-1}|\mathcal{I}_{a'1}|\right) \le \eta \\
w^\top \hat{x}_i + \rho\|w\|_* + b + \varepsilon \le M\lambda_i^a \quad \forall i \in \mathcal{I}_{a1} \\
-w^\top \hat{x}_i + \rho\|w\|_* - b \le M\lambda_i^a \qquad \forall i \in \mathcal{I}_{a'1}
\end{array}\right\} & \forall (a, a') \in \{(0,1),(1,0)\},
\end{array} \tag{8}$$

where M is the big-M parameter.

For notational simplicity, we present the reformulation (8) with $2N$ binary variables. A closer investigation into problem (8) reveals that it suffices to use $2|\mathcal{I}_1|$ 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 a 1-norm or an ∞-norm on $\mathbb{R}^d$, problem (8) is a linear mixed binary optimization problem. If $\|\cdot\|$ is an Euclidean norm, problem (8) becomes a second-order cone, mixed binary optimization problem. Both types of problems can be solved using off-the-shelf solvers such as MOSEK [47].

For the remainder of this section, we will provide the proof for Theorem 3.1. This proof relies on the following auxiliary result.

Lemma 3.2. Fix any index set $\mathcal{K} \subseteq \{1, \ldots, N\}$, a radius $\rho \in \mathbb{R}_+$, a classifier $(w, b) \in \mathbb{R}^{d+1}$ and a collection of samples $\{\hat{x}_k\}_{k\in\mathcal{K}}$. For any $\varepsilon \in \mathbb{R}$, we have

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

where M is the big-M constant.

Proof of Lemma 3.2. Using an epigraphical formulation of each supremum term, we find

$$\sum_{k\in\mathcal{K}} \sup_{x_k:\|x_k-\hat{x}_k\|\leq\rho} \mathbb{I}(w^\top x_k + b > \varepsilon) = \left\{\begin{array}{ll} \min & \sum_{k\in\mathcal{K}} \lambda_k \\ \text{s.t.} & \lambda \in \{0,1\}^N \\ & \sup_{x_k:\|x_k-\hat{x}_k\|\leq\rho} \mathbb{I}(w^\top x_k + b > \varepsilon) \leq \lambda_k \quad \forall k \in \mathcal{K} \end{array}\right.$$
$$= \left\{\begin{array}{ll} \min & \sum_{k\in\mathcal{K}} \lambda_k \\ \text{s.t.} & \lambda \in \{0,1\}^N \\ & \max_{x_k:\|x_k-\hat{x}_k\|\leq\rho} w^\top x_k + b - \varepsilon \leq M\lambda_k \quad \forall k \in \mathcal{K}, \end{array}\right.$$

where M is the big-M constant. The dual norm definition implies that

$$\sup_{x_k:\|x_k-\hat{x}_k\|\leq\rho} w^\top x_k = w^\top \hat{x}_k + \rho\|w\|_*,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ on $\mathbb{R}^d$. This completes the proof. □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. 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

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

where $\operatorname{supp}(\pi_i)$ denotes the support of the probability measure π_i [2, Page 441]. We first provide the reformulation for the objective function of (6). For any $(w,b)\in\mathbb{R}^{d+1}$, we have

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

where the last equality follows from an epigraphical reformulation and from the properties of the dual norm.

Next, we provide the reformulation for the objective function of (6). For any $(w,b)\in\mathbb{R}^{d+1}$, we can rewrite the worst-case unfairness value as

$$\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{U}(w,b,\mathbb{Q}) = \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \max \left\{ \begin{array}{l} \mathbb{Q}_{01}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{11}(w^\top X + b \geq 0), \\ \mathbb{Q}_{11}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{01}(w^\top X + b \geq 0) \end{array} \right\}$$
$$= \max \left\{ \begin{array}{l} \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}_{01}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{11}(w^\top X + b \geq 0), \\ \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}_{11}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{01}(w^\top X + b \geq 0) \end{array} \right\}.$$

Define the following index sets $\mathcal{I}_{a1} = \{i \in [N] : \hat{a}_i = a, \hat{y}_i = 1\}$ $\forall a \in \mathcal{A}$. Intuitively, $\mathcal{I}_{a1}$ contains the indices of the samples with sensitive attribute a and label 1. Fix any pair $(a, a') \in \{(0,1),(1,0)\}$, we have

$$
\begin{aligned}
&\sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}(w^\top X + b > -\varepsilon | A = a, Y = 1) - \mathbb{Q}(w^\top X + b \geq 0 | A = a', Y = 1) \\
= &\sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\hat{p}_{a1}^{-1} \mathbb{I}(w^\top X + b > -\varepsilon) \mathbb{1}_{(a,1)}(A, Y) - \hat{p}_{a'1}^{-1} \mathbb{I}(w^\top X + b \geq 0) \mathbb{1}_{(a',1)}(A, Y)] \\
= &\frac{1}{N} \left(\hat{p}_{a1}^{-1} \sum_{i \in \mathcal{I}_{a1}} \sup_{x_i : \|x_i - \hat{x}_i\| \leq \rho} \mathbb{I}(w^\top x_i + b > -\varepsilon) - \hat{p}_{a'1}^{-1} \sum_{i \in \mathcal{I}_{a'1}} \inf_{x_i : \|x_i - \hat{x}_i\| \leq \rho} \mathbb{I}(w^\top x_i + b \geq 0) \right) \\
= &\frac{1}{N} \left(\hat{p}_{a1}^{-1} \sum_{i \in \mathcal{I}_{a1}} \sup_{x_i : \|x_i - \hat{x}_i\| \leq \rho} \mathbb{I}(w^\top x_i + b > -\varepsilon) - \hat{p}_{a'1}^{-1} \big(|\mathcal{I}_{a'1}| - \sum_{i \in \mathcal{I}_{a'1}} \sup_{x_i : \|x_i - \hat{x}_i\| \leq \rho} \mathbb{I}(w^\top x_i + b < 0) \big) \right) \\
= &\frac{1}{N} \left(\hat{p}_{a1}^{-1} \sum_{i \in \mathcal{I}_{a1}} \sup_{x_i : \|x_i - \hat{x}_i\| \leq \rho} \mathbb{I}(w^\top x_i + b > -\varepsilon) + \hat{p}_{a'1}^{-1} \sum_{i \in \mathcal{I}_{a'1}} \sup_{x_i : \|x_i - \hat{x}_i\| \leq \rho} \mathbb{I}(w^\top x_i + b < 0) - \hat{p}_{a'1}^{-1} |\mathcal{I}_{a'1}| \right) \\
= &\left\{
\begin{array}{lll}
\min & \frac{1}{N} \left(\hat{p}_{a1}^{-1} \sum_{i \in \mathcal{I}_{a1}} \lambda_i^a + \hat{p}_{a'1}^{-1} \sum_{i \in \mathcal{I}_{a'1}} \lambda_i^a - \hat{p}_{a'1}^{-1} |\mathcal{I}_{a'1}| \right) & \\
\text{s.t.} & \lambda^a \in \{0,1\}^N & \\
& w^\top \hat{x}_i + \rho \|w\|_* + b + \varepsilon \leq M \lambda_i^a & \forall i \in \mathcal{I}_{a1} \\
& -w^\top \hat{x}_i + \rho \|w\|_* - b \leq M \lambda_i^a & \forall i \in \mathcal{I}_{a'1},
\end{array}
\right.
\end{aligned}
$$

where the last equality follows by applying Lemma 3.2 twice and by noticing that $\mathcal{I}_{a1} \cap \mathcal{I}_{a'1} = \emptyset$. Setting the optimal value of the above minimization problem to be less than η completes the proof. $\square$

4. Training with General Ground Metric

Throughout this section, we use the general ground metric

$$
c\big((x', a', y'), (x, a, y)\big) = \|x - x'\| + \kappa_{\mathcal{A}} |a - a'| + \kappa_{\mathcal{Y}} |y - y'| \tag{9}
$$

for some finite values of $\kappa_{\mathcal{A}}$ and $\kappa_{\mathcal{Y}}$. At the same time, we will consider in this section a more general definition of the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$. To this end, we first observe that the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ can be re-expressed as[2]

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

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

$$
\mathcal{B}_\gamma(\hat{\mathbb{P}}) \triangleq \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) :
\begin{array}{ll}
\exists \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \quad \forall i \in [N] : & \\
\mathbb{Q} = N^{-1} \sum_{i \in [N]} \pi_i & \\
\mathbb{W}_\infty(\pi_i, \delta_{(\hat{x}_i, \hat{a}_i, \hat{y}_i)}) \leq \rho & \forall i \in [N] \\
\mathbb{Q}(A = a, Y = y) = \hat{p}_{ay} & \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \\
\sum_{i \in [N]} \pi_i(A = \hat{a}_i, Y = \hat{y}_i) \geq (1 - \gamma) N &
\end{array}
\right\}. \tag{10}
$$

Notice that $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ differs from $\mathbb{B}(\hat{\mathbb{P}})$ solely on the basis of 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

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

[2]A formal proof can be found in Lemma 8.3.

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

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

The latter constraint resembles the case considered in Section 3 with absolute trust in the sensitive attribute and the label. Any value $\gamma \in (0, 1)$ thus can be thought of as an interpolation of the robustness condition between these two above-mentioned extreme cases.

We consider in this section the modified problem

$$\begin{array}{ll} \min & \sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(w^\top X + b)\}] \\ \text{s.t.} & w \in \mathbb{R}^d, b \in \mathbb{R} \\ & \sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{U}(w, b, \mathbb{Q}) \leq \eta. \end{array} \tag{11}$$

It is easy to show, by modifying Lemma 2.3 and leveraging Corollary 8.2, that the feasible set of problem (11) is non-empty, and hence problem (11) is well-defined. We now present the main result of this section, which provides the reformulation for problem (11).

Theorem 4.1. (Reformulation) For any $\gamma \in (0, 1)$, problem (11) is equivalent to the following mixed binary conic program

$$\begin{array}{ll} \inf & \frac{1}{N}\sum_{i\in[N]} \nu_i + \sum_{(\bar{a},\bar{y})\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{\bar{a}\bar{y}}\mu_{\bar{a}\bar{y}} - \theta(1-\gamma) \\ \text{s.t.} & \nu \in \mathbb{R}^N,\ \theta \in \mathbb{R}_+,\ \mu \in \mathbb{R}^{2\times 2} \\ & \nu^a \in \mathbb{R}^N,\ \theta^a \in \mathbb{R}_+,\ \mu^a \in \mathbb{R}^{2\times 2}, \lambda_a^a \in \{0,1\}^N,\ \lambda_{a'}^a \in \{0,1\}^N \qquad \forall (a, a') \in \{(0,1),(1,0)\} \\ & \left.\begin{array}{l} \text{If } \kappa_{\mathcal{A}}|a - \hat{a}_i| + \kappa_{\mathcal{Y}}|y - \hat{y}_i| \leq \rho: \\ \quad 0 \leq \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a, y) + \nu_i \\ \quad 1 - yb - yw^\top\hat{x}_i + (\rho - \kappa_{\mathcal{A}}|a - \hat{a}_i| - \kappa_{\mathcal{Y}}|y - \hat{y}_i|)\|w\|_* \\ \qquad\qquad \leq \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a, y) + \nu_i \end{array}\right\} \quad \forall i \in [N],\ (a, y) \in \mathcal{A}\times\mathcal{Y} \\ & \left.\begin{array}{l} \text{If } \kappa_{\mathcal{A}}|a - \hat{a}_i| + \kappa_{\mathcal{Y}}|1 - \hat{y}_i| \leq \rho: \\ \quad \hat{p}_{a1}^{-1}\lambda_{ai}^a \leq \mu_{a,1}^a - \theta^a\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a, 1) + \nu_i^a \\ \quad w^\top\hat{x}_i + (\rho - \kappa_{\mathcal{A}}|a - \hat{a}_i| - \kappa_{\mathcal{Y}}|1 - \hat{y}_i|)\|w\|_* + b + \varepsilon \leq M\lambda_{ai}^a \\ \text{If } \kappa_{\mathcal{A}}|a' - \hat{a}_i| + \kappa_{\mathcal{Y}}|1 - \hat{y}_i| \leq \rho: \\ \quad \hat{p}_{a'1}^{-1}(\lambda_{a'i}^a - 1) \leq \mu_{a'1}^a - \theta^a\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a', 1) + \nu_i^a \\ \quad -w^\top\hat{x}_i + (\rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}|1 - \hat{y}_i|)\|w\|_* - b \leq M\lambda_{a'i}^a \\ \text{If } \kappa_{\mathcal{A}}|a - \hat{a}_i| + \kappa_{\mathcal{Y}}|-1 - \hat{y}_i| \leq \rho: \\ \quad 0 \leq \mu_{a,-1}^a - \theta^a\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a, -1) + \nu_i^a \\ \text{If } \kappa_{\mathcal{A}}|a' - \hat{a}_i| + \kappa_{\mathcal{Y}}|-1 - \hat{y}_i| \leq \rho: \\ \quad 0 \leq \mu_{a',-1}^a - \theta^a\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a', -1) + \nu_i^a. \end{array}\right\} \quad \forall (a, a') \in \{(0,1),(1,0)\},\ \forall i \in [N] \\ & \frac{1}{N}\sum_{i\in[N]} \nu_i^a + \sum_{(\bar{a},\bar{y})\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{ay}\mu_{\bar{a}\bar{y}}^a - \theta^a(1-\gamma) \leq \eta \quad \forall a \in \mathcal{A}, \end{array} \tag{12}$$

where M is the big-M constant.

The reformulation (12) involves $4N$ binary variables. However, because the constraints of problem (12) are contingent, the empirical number of binary variables is smaller than $4N$. Problem (8) is a linear mixed binary optimization problem whenever if $\|\cdot\|$ is either a 1-norm or an ∞-norm on $\mathbb{R}^d$. In the remainder of this section, we will provide the proof of Theorem 4.1. This proof leverages the following duality result.

Lemma 4.2. (Strong duality) Let $\phi : \mathcal{X}\times\mathcal{A}\times\mathcal{Y} \to \mathbb{R}$ be a Borel measurable loss function. Then for any $\gamma \in (0, 1)$, the semi-infinite program

$$\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\phi(X, A, Y)] \tag{13a}$$

admits the following dual form

$$
\begin{array}{ll}
\inf & \frac{1}{N}\sum_{i\in[N]}\nu_i + \sum_{a\in\mathcal{A},y\in\mathcal{Y}}\hat{p}_{a,y}\mu_{a,y} - \theta(1-\gamma) \\
\text{s.t.} & \nu\in\mathbb{R}^N, \theta\in\mathbb{R}_+, \mu\in\mathbb{R}^{2\times 2} \\
& \sup\limits_{x:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|a-\hat{a}_i|-\kappa_{\mathcal{Y}}|y-\hat{y}_i|} \phi(x,a,y) \le \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y) + \nu_i \quad \forall i\in[N], (a,y)\in\mathcal{A}\times\mathcal{Y},
\end{array} \tag{13b}
$$

where the supremum value is considered to be $-\infty$ if the corresponding feasible set is empty.

Proof of Lemma 4.2. Using the definition of the type-∞ Wasserstein distance, we can re-express the ambiguity set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ as

$$
\mathcal{B}_\gamma(\hat{\mathbb{P}}) = \left\{ \mathbb{Q}\in\mathcal{M}(\mathcal{X}\times\mathcal{A}\times\mathcal{Y}) : \begin{array}{l}
\exists \pi_i\in\mathcal{M}(\mathcal{X}\times\mathcal{A}\times\mathcal{Y})\ \forall i\in[N] \text{ such that } \mathbb{Q} = \frac{1}{N}\sum_{i\in[N]}\pi_i \\
N^{-1}\sum_{i=1}^N \pi_i(A=a,Y=y) = \hat{p}_{ay} \quad \forall (a,y)\in\mathcal{A}\times\mathcal{Y} \\
\|x-\hat{x}_i\| + \kappa_{\mathcal{A}}|a-\hat{a}_i| + \kappa_{\mathcal{Y}}|y-\hat{y}_i| \le \rho \quad \forall (x,a,y)\in\mathrm{supp}(\pi_i) \quad \forall i\in[N] \\
\sum_{i\in[N]}\pi_i(A=\hat{a}_i, Y=\hat{y}_i) \ge (1-\gamma)N
\end{array} \right\}.
$$

The worst-case expected loss can now be written as

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\phi(X,A,Y)] \\
&= \left\{ \begin{array}{lll}
\sup & N^{-1}\sum_{i\in[N]} \mathbb{E}_{\pi_i}[\phi(X,A,Y)] & \\
\text{s.t.} & \pi_i\in\mathcal{M}(\mathcal{X}\times\mathcal{A}\times\mathcal{Y}) & \forall i\in[N] \\
& \sum_{i\in[N]}\pi_i(A=a,Y=y) = N\hat{p}_{ay} & \forall (a,y)\in\mathcal{A}\times\mathcal{Y} \\
& \sum_{i\in[N]}\pi_i(A=\hat{a}_i,Y=\hat{y}_i) \ge (1-\gamma)N & \\
& \|x-\hat{x}_i\| + \kappa_{\mathcal{A}}|a-\hat{a}_i| + \kappa_{\mathcal{Y}}|y-\hat{y}_i| \le \rho & \forall (x,a,y)\in\mathrm{supp}(\pi_i) \quad \forall i\in[N]
\end{array} \right.
\end{aligned}
$$

Any $\pi_i\in\mathcal{M}(\mathcal{X}\times\mathcal{A}\times\mathcal{Y})$ can be decomposed as

$$
\pi_i(\mathrm{d}x\times\mathrm{d}a'\times\mathrm{d}y') = \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} \tau_{iay}\pi_{iay}(\mathrm{d}x)\delta_{(a,y)}(\mathrm{d}a'\times\mathrm{d}y'),
$$

where π_{iay} is the conditional distribution of X given that $(A,Y)=(a,y)$ and the weights $\tau\in\mathbb{R}^{N\times|\mathcal{A}|\times|\mathcal{Y}|}$ satisfy

$$
\tau_{iay}\ge 0 \quad \forall (a,y)\in\mathcal{A}\times\mathcal{Y}, \quad \text{and} \quad \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}}\tau_{iay} = 1 \qquad \forall i\in[N].
$$

Moreover, define the following optimal values

$$
v_{iay} = \sup\{\phi(x,a,y) : x\in\mathcal{X},\ \|x-\hat{x}_i\| \le \rho - \kappa_{\mathcal{A}}|a-\hat{a}_i| - \kappa_{\mathcal{Y}}|y-\hat{y}_i|\} \tag{14}
$$

for each $i\in[N]$ and $(a,y)\in\mathcal{A}\times\mathcal{Y}$. Denote momentarily the feasible set of the above optimization problem as $\mathcal{X}_{iay}$. Notice that $\mathcal{X}_{iay}=\emptyset$ if $\rho-\kappa_{\mathcal{A}}|a-\hat{a}_i|-\kappa_{\mathcal{Y}}|y-\hat{y}_i|<0$ and in this case we set $v_{iay}=-\infty$. By definition, we also have

$$
v_{iay} = \sup_{\pi_{iay}\in\mathcal{M}(\mathcal{X})} \int_{\mathcal{X}_{iay}} \phi(x,a,y)\pi_{iay}(\mathrm{d}x)
$$

whenever $\mathcal{X}_{iay}$ is non-empty. Using this definition of v and by the above decomposition of π, we obtain

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\phi(X,A,Y)] \\
&= \left\{ \begin{array}{lll}
\sup & N^{-1}\sum_{i\in[N]}\sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}}\tau_{iay}v_{iay} & \\
\text{s.t.} & \tau_{iay}\ge 0 & \forall i\in[N], (a,y)\in\mathcal{A}\times\mathcal{Y} \\
& \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}}\tau_{iay} = 1 & \forall i\in[N] \\
& \sum_{i\in[N]}\tau_{iay} = N\hat{p}_{ay} & \forall (a,y)\in\mathcal{A}\times\mathcal{Y} \\
& \sum_{i\in[N]}\tau_{i\hat{a}_i\hat{y}_i} \ge (1-\gamma)N &
\end{array} \right.
\end{aligned}
$$

which is a finite-dimensional linear program. Strong duality result from linear programming asserts that

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\phi(X,A,Y)] \\
&= \left\{ \begin{array}{ll} \inf & \frac{1}{N}\sum_{i\in[N]}\nu_i + \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{ay}\mu_{ay} - \theta(1-\gamma) \\ \text{s.t.} & \nu\in\mathbb{R}^N,\ \mu\in\mathbb{R}^{|\mathcal{A}|\times|\mathcal{Y}|},\ \theta\in\mathbb{R}_+ \\ & v_{iay} \le \nu_i + \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y) \qquad \forall i\in[N], (a,y)\in\mathcal{A}\times\mathcal{Y}. \end{array}\right.
\end{aligned}
$$

Substituting the definition of v into the above optimization problem completes the proof. □

Equipped with the duality result of Lemma 4.2, we now present the proof of Theorem 4.1.

Proof of Theorem 4.1. Notice that the objective function can be written in the form of

$$\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\phi(X,A,Y)],$$

where $\phi(X,A,Y)=\max\{0,1-Y(w^\top X+b)\}$ is the SVM hinge loss. Thus, by Lemma 4.2, we have

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\phi(X,A,Y)] \\
&= \left\{ \begin{array}{ll} \inf & \frac{1}{N}\sum_{i\in[N]}\nu_i + \sum_{(\bar{a},\bar{y})\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{\bar{a}\bar{y}}\mu_{\bar{a}\bar{y}} - \theta(1-\gamma) \\ \text{s.t.} & \nu\in\mathbb{R}^N,\ \theta\in\mathbb{R}_+,\ \mu\in\mathbb{R}^{2\times 2} \\ & \sup\limits_{x:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|a-\hat{a}_i|-\kappa_{\mathcal{Y}}|y-\hat{y}_i|} \max\{0,1-y(w^\top x+b)\} \le \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y) + \nu_i \quad \forall i\in[N], (a,y)\in\mathcal{A}\times\mathcal{Y}. \end{array}\right.
\end{aligned}
$$

The constraint in the above infimum problem is equivalent to

$$
\left.\begin{array}{l} \text{If } \kappa_{\mathcal{A}}|a-\hat{a}_i| + \kappa_{\mathcal{Y}}|y-\hat{y}_i| \le \rho: \\ \quad 0 \le \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y) + \nu_i \\ \quad 1 - yb + \sup\limits_{x:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|a-\hat{a}_i|-\kappa_{\mathcal{Y}}|y-\hat{y}_i|} \{-yw^\top x\} \le \mu_{ay} - \theta\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y) + \nu_i \end{array}\right\} \forall i\in[N],\ (a,y)\in\mathcal{A}\times\mathcal{Y}.
$$

Remind that $\mathcal{Y}=\{-1,+1\}$, thus the dual norm relationship implies that

$$\sup_{x:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|a-\hat{a}_i|-\kappa_{\mathcal{Y}}|y-\hat{y}_i|} \{-yw^\top x\} = -yw^\top\hat{x}_i + (\rho-\kappa_{\mathcal{A}}|a-\hat{a}_i|-\kappa_{\mathcal{Y}}|y-\hat{y}_i|)\|w\|_*,$$

which lead to first set of constraints in the reformulation

Next, we show the derivation for constraints. Recall that the worst-case unfairness measure can be written as

$$\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{U}(w,b,\mathbb{Q}) = \max\left\{ \begin{array}{l} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{Q}_{01}(w^\top X+b>-\varepsilon) - \mathbb{Q}_{11}(w^\top X+b\ge 0), \\ \sup\limits_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{Q}_{11}(w^\top X+b>-\varepsilon) - \mathbb{Q}_{01}(w^\top X+b\ge 0) \end{array}\right\}.$$

Consider a fixed pair of $(a,a')\in\{(0,1),(1,0)\}$. To employ the result of Lemma 4.2, we write the constraints in the form of (13a)

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{Q}(w^\top X+b>-\varepsilon|A=a,Y=1) - \mathbb{Q}(w^\top X+b\ge 0|A=a',Y=1) \\
&= \sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\hat{p}_{a1}^{-1}\mathbb{I}(w^\top X+b>-\varepsilon)\mathbb{1}_{(a,1)}(A,Y) - \hat{p}_{a'1}^{-1}\mathbb{I}(w^\top X+b\ge 0)\mathbb{1}_{(a',1)}(A,Y)] \\
&= \left\{ \begin{array}{ll} \inf & \dfrac{1}{N}\sum\limits_{i\in[N]}\nu_i^a + \sum\limits_{(\bar{a},\bar{y})\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{ay}\mu^a_{\bar{a}\bar{y}} - \theta^a(1-\gamma) \\ \text{s.t.} & \nu^a\in\mathbb{R}^N,\ \theta^a\in\mathbb{R}_+,\ \mu^a\in\mathbb{R}^{2\times 2}, \\ & \sup\limits_{x:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|\bar{a}_i-\hat{a}_i|-\kappa_{\mathcal{Y}}|\bar{y}_i-\hat{y}_i|} \phi_a(x,\bar{a}_i,\bar{y}_i) \le \mu^a_{\bar{a}_i\bar{y}_i} - \theta^a\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(\bar{a}_i,\bar{y}_i) + \nu_i^a \quad \forall(\bar{a}_i,\bar{y}_i)\in\mathcal{A}\times\mathcal{Y}, \quad i\in[N], \end{array}\right.
\end{aligned}
$$

where the second equation relies on the result of Lemma 4.2 by defining

$$\phi_a(X,A,Y) = \hat{p}_{a1}^{-1}\mathbb{I}(w^\top X + b > -\varepsilon)\mathbb{1}_{(a,1)}(A,Y) - \hat{p}_{a'1}^{-1}\mathbb{I}(w^\top X + b \geq 0)\mathbb{1}_{(a',1)}(A,Y).$$

Fix any $i \in [N]$, we now iterate over $(\bar{a}_i, \bar{y}_i)$.

(1) Case 1: $(\bar{a}_i, \bar{y}_i) = (a, 1)$. There is an active constraint if $\kappa_{\mathcal{A}}|a - \hat{a}_i| + \kappa_{\mathcal{Y}}|1 - \hat{y}_i| \leq \rho$, and the semi-infinite constraint is equivalent to

$$\begin{aligned}
&\hat{p}_{a1}^{-1}\mathbb{I}(w^\top x_i + b > -\varepsilon)\mathbb{1}_{(a,1)}(a,1) - \hat{p}_{11}^{-1}\mathbb{I}(w^\top x_i + b \geq 0)\mathbb{1}_{(a',1)}(a,1) \leq \mu_{a1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,1) + \nu_i^a \\
&\hspace{20em} \forall x_i : \|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a - \hat{a}_i| - \kappa_{\mathcal{Y}}|1 - \hat{y}_i| \\
\iff & \sup_{\|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a_- \hat{a}_i| - \kappa_{\mathcal{Y}}|1 - \hat{y}_i|} \hat{p}_{a1}^{-1}\mathbb{I}(w^\top x_i + b > -\varepsilon) \leq \mu_{a1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,1) + \nu_i^a \\
\iff & \begin{cases} \exists \lambda_{ai}^a \in \{0,1\} \\ \hat{p}_{a1}^{-1}\lambda_{ai}^a \leq \mu_{a1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,1) + \nu_i^a \\ w^\top \hat{x}_i + (\rho - \kappa_{\mathcal{A}}|a - \hat{a}_i| - \kappa_{\mathcal{Y}}|1 - \hat{y}_i|)\|w\|_* + b + \varepsilon \leq M\lambda_{ai}^a, \end{cases}
\end{aligned}$$

where the last equation follows from the result of Lemma 3.2.

(2) Case 2: $(\bar{a}_i, \bar{y}_i) = (a', 1)$. There is an active constraint if $\kappa_{\mathcal{A}}|a' - \hat{a}_i| + \kappa_{\mathcal{Y}}|1 - \hat{y}_i| \leq \rho$, and the semi-infinite constraint is equivalent to

$$\begin{aligned}
&\hat{p}_{a1}^{-1}\mathbb{I}(w^\top x_i + b > -\varepsilon)\mathbb{1}_{(a,1)}(a',1) - \hat{p}_{a'1}^{-1}\mathbb{I}(w^\top x_i + b \geq 0)\mathbb{1}_{(a',1)}(a',1) \leq \mu_{a'1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a',1) + \nu_i^a \\
&\hspace{16em} \forall x_i : \|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}|1 - \hat{y}_i| \\
\iff & \sup_{\|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}|y_i - \hat{y}_i|} -\hat{p}_{a'1}^{-1}\mathbb{I}(w^\top x_i + b \geq 0) \leq \mu_{a'1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a',1) + \nu_i^a \\
\iff & -\hat{p}_{a'1}^{-1} \inf_{\|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}|y_i - \hat{y}_i|} \mathbb{I}(w^\top x_i + b \geq 0) \leq \mu_{a'1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a',1) + \nu_i^a \\
\iff & -\hat{p}_{a'1}^{-1}\left(1 - \sup_{\|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}|y_i - \hat{y}_i|} \mathbb{I}(w^\top x_i + b < 0)\right) \leq \mu_{a'1}^a - \theta^a \mathbb{1}_{(\hat{a}',\hat{y}_i)}(a',1) + \nu_i^a \\
\iff & \hat{p}_{a'1}^{-1}\left(\sup_{\|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}|y_i - \hat{y}_i|} \mathbb{I}(w^\top x_i + b < 0) - 1\right) \leq \mu_{a'1}^a - \theta^a \mathbb{1}_{(\hat{a}',\hat{y}_i)}(a',1) + \nu_i^a \\
\iff & \begin{cases} \exists \lambda_{a'i}^a \in \{0,1\} \\ \hat{p}_{a'1}^{-1}(\lambda_{a'i}^a - 1) \leq \mu_{a'1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a',1) + \nu_i^a \\ -w^\top \hat{x}_i + (\rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}|1 - \hat{y}_i|)\|w\|_* - b \leq M\lambda_{a'i}^a. \end{cases}
\end{aligned}$$

(3) Case 3: $(\bar{a}_i, \bar{y}_i) = (a, -1)$. There is an active constraint if $\kappa_{\mathcal{A}}|a - \hat{a}_i| + \kappa_{\mathcal{Y}}|-1 - \hat{y}_i| \leq \rho$, the semi-infinite constraint is equivalent to

$$\begin{aligned}
&\hat{p}_{a1}^{-1}\mathbb{I}(w^\top x_i + b > -\varepsilon)\mathbb{1}_{(a,1)}(a,-1) - \hat{p}_{a'1}^{-1}\mathbb{I}(w^\top x_i + b \geq 0)\mathbb{1}_{(a',1)}(a,-1) \leq \mu_{a,-1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,-1) + \nu_i^a \\
&\hspace{20em} \forall x_i : \|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a - \hat{a}_i| - \kappa_{\mathcal{Y}}| - 1 - \hat{y}_i| \\
\iff & 0 \leq \mu_{a,-1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,-1) + \nu_i^a.
\end{aligned}$$

(4) Case 4: $(\bar{a}_i, \bar{y}_i) = (a', -1)$. There is an active constraint if $\kappa_{\mathcal{A}}|a' - \hat{a}_i| + \kappa_{\mathcal{Y}}| - 1 - \hat{y}_i| \leq \rho$, the semi-infinite constraint is equivalent to

$$\begin{aligned}
&\hat{p}_{a1}^{-1}\mathbb{I}(w^\top x_i + b > -\varepsilon)\mathbb{1}_{(a,1)}(a',-1) - \hat{p}_{a'1}^{-1}\mathbb{I}(w^\top x_i + b \geq 0)\mathbb{1}_{(a',1)}(a',-1) \leq \mu_{a',-1}^a - \theta \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a',-1) + \nu_i \\
&\hspace{20em} \forall x_i : \|x_i - \hat{x}_i\| \leq \rho - \kappa_{\mathcal{A}}|a' - \hat{a}_i| - \kappa_{\mathcal{Y}}| - 1 - \hat{y}_i| \\
\iff & 0 \leq \mu_{a',-1}^a - \theta^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a',-1) + \nu_i^a.
\end{aligned}$$

Notice that at least one of the above four conditions will be satisfied, because when $\bar{a}_i = \hat{a}_i$ and $\bar{y}_i = \hat{y}_i$, we have

$$\kappa_{\mathcal{A}}|a_i - \hat{a}_i| + \kappa_{\mathcal{Y}}|y_i - \hat{y}_i| = 0 \leq \rho$$

for any $\rho \geq 0$. Combining all four cases leads to the second set of constraints.

The last constraint in the reformulation is obtained by setting the optimal value of the dual problem to be less than η for each value of $a \in \mathcal{A}$. This completes the proof. □

5. Performance Guarantees

We now extend our study to a probability minimization perspective, and show that SVM actually serves as a CVaR approximation of the misclassification probability minimization problem, which also implies that SVM provides an upperbound for the misclassification probability.

In the context of linear binary classification problems, we need to find a classifier that maximizes the correct classification probability. For a linear classifier parametrized by $(w, b) \in \mathbb{R}^{d+1}$, we can consider the *correct classfication probability* with respect to the distribution $\mathbb{Q}$ as

$$\mathbb{Q}\left(Y(w^\top X + b) \geq 0\right).$$

Notice that by definition, we consider that any x falling exactly on the hyperplane $w^\top X + b = 0$ is correctly classified irrespective of the true label of x. Complementary, the *misclassification probability* with respect to $\mathbb{Q}$ is defined as

$$\mathbb{Q}\left(Y(w^\top X + b) < 0\right).$$

The next result assert that the probability of misclassification can be upper bounded by the expected hinge loss. Interestingly, we could not manage to locate this result in the existing literature.

Proposition 5.1 (Misclassification rate)**.** For any classifier parametrized by $(w_0, b_0) \in \mathbb{R}^{d+1}$ and any distribution $\mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$, we have

$$\mathbb{Q}(Y(w_0^\top X + b_0) < 0) \leq \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(w_0^\top X + b_0)\}].$$

Proof of Proposition 5.1. Suppose that under $\mathbb{Q}$, we have

$$\mathbb{Q}(Y(w_0^\top X + b_0) \geq 0) = 1,$$

then we have $\mathbb{Q}(Y(w_0^\top X + b_0) < 0) = 0$ and the claim holds true trivially because the expected value on the right hand side is always non-negative.

Consider now the case where $\mathbb{Q}(Y(w_0^\top X + b_0) \geq 0) < 1$. We find

$$\begin{aligned} t \triangleq \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(w_0^\top X + b_0)\}] &\geq \inf_{\beta > 0} \; \mathbb{E}_{\mathbb{Q}}\left[\max\left\{0, 1 - Y\left(\left(\frac{w_0}{\beta}\right)^\top X + \left(\frac{b_0}{\beta}\right)\right)\right\}\right] \\ &= \inf_{\beta > 0} \; \frac{1}{\beta}\mathbb{E}_{\mathbb{Q}}[\max\{0, \beta - Y(w_0^\top X + b_0)\}]. \end{aligned}$$

Suppose momentarily that $t > 0$. The above inequality implies that

$$\inf_{\beta > 0} \left\{-\beta + \frac{1}{t}\mathbb{E}_{\mathbb{Q}}[\max\{0, \beta - Y(w_0^\top X + b_0)\}]\right\} \leq 0. \tag{15a}$$

Notice that condition (15a) is equivalent to

$$\inf_{\beta \in \mathbb{R}} \left\{-\beta + \frac{1}{t}\mathbb{E}_{\mathbb{Q}}[\max\{0, \beta - Y(w_0^\top X + b_0)\}]\right\} \leq 0, \tag{15b}$$

in which the feasible set over β is relaxed to the whole real line. To see this, consider when $\beta < 0$, then we have $-\beta > 0$; meanwhile, $\mathbb{E}_{\mathbb{Q}}[\max\{0, \beta - Y(w_0^\top X + b_0)\}]$ is a non-negative term. Hence, the sum of these two terms must be positive. When $\beta = 0$, we have

$$\mathbb{E}_{\mathbb{Q}}[\max\{0, -Y(w_0^\top X + b_0)\}] = \underbrace{\mathbb{E}_{\mathbb{Q}}[-Y(w_0^\top X + b_0)| - Y(w_0^\top X + b_0) > 0]}_{>0} \underbrace{\mathbb{Q}(-Y(w_0^\top X + b_0) > 0)}_{>0} > 0.$$

by the inseparable assumption. The definition of the Conditional Value-at-Risk (CVaR, [53]) implies that condition (15b) is equivalent to

$$\mathbb{Q}\text{-CVaR}_{1-t}(-Y(w_0^\top X + b_0)) \leq 0.$$

Because CVaR is a conservative approximation of the Value-at-Risk (VaR), we thus find

$$\mathbb{Q}(-Y(w_0^\top X + b_0) \leq 0) \geq 1 - t,$$

which means

$$\mathbb{Q}(Y(w_0^\top X + b_0) < 0) \leq t.$$

To complete the proof, we consider the case where $t = 0$. The definition of t implies that the distribution $\mathbb{Q}$ should thus satisfy

$$\mathbb{Q}(Y(w_0^\top X + b_0) \geq 1) = 1.$$

Nevertheless, the above condition conflicts with the assumption that $\mathbb{Q}(Y(w_0^\top X + b_0) \geq 1) < 1$. Thus, the case where $t = 0$ can be discarded. This completes the proof. □

Proposition 5.1 leads to the following results, the proof is trivial and thus is omitted.

Corollary 5.2 (Worst-case misclassification rate)**.** The following claims hold.

(i) Suppose that $(w_0, b_0) \in \mathbb{R}^{d+1}$ is the optimal solution of problem (6) with corresponding optimal value $v^\star$, then

$$\mathbb{Q}(Y(w_0^\top X + b_0) < 0) \leq v^\star \qquad \forall \mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}).$$

(ii) For any $\gamma \in [0, 1]$, suppose that $(w_0, b_0) \in \mathbb{R}^{d+1}$ is the optimal solution of problem (11) with corresponding optimal value $v^\star$, then

$$\mathbb{Q}(Y(w_0^\top X + b_0) < 0) \leq v^\star \qquad \forall \mathbb{Q} \in \mathcal{B}_\gamma(\hat{\mathbb{P}}).$$

An avid reader may be interested in the role of the parameter γ in the set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ defined in (10). Remind that when $\gamma = 1$, we have $\mathcal{B}_\gamma(\hat{\mathbb{P}}) = \mathbb{B}(\hat{\mathbb{P}})$. The next lemma asserts that when $\gamma = 1$ and when ρ is set to sufficiently big, then any SVM classifier will incur a worst-case misclassification rate of 1.

Lemma 5.3 (Uninformative misclassification rate)**.** Suppose that the ground cost is chosen as in (9), $\gamma = 1$ and $\rho > \max\{2\kappa_{\mathcal{Y}}, 0\}$. For any non-trivial classifier parametrized by $(w, b) \in \mathbb{R}^{d+1}$ with $w \neq 0$, we have

$$\sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}(Y(w^\top X + b) < 0) = \sup_{\mathbb{Q} \in \mathcal{B}_1(\hat{\mathbb{P}})} \mathbb{Q}(Y(w^\top X + b) < 0) = 1$$

Proof of Lemma 4.2. Because $2\kappa_{\mathcal{Y}} < \rho$, it is feasible to flip the label of each training sample from $+1$ to -1, and vice versa. Thus, consider the following distribution

$$\mathbb{Q}^\star = \frac{1}{N} \sum_{i: \hat{y}_i(w^\top \hat{x}_i + b) < 0} \delta_{(\hat{x}_i, \hat{a}_i, -\hat{y}_i)} + \frac{1}{N} \sum_{i: \hat{y}_i(w^\top \hat{x}_i + b) > 0} \delta_{(\hat{x}_i, \hat{a}_i, -\hat{y}_i)} + \frac{1}{N} \sum_{i: w^\top \hat{x}_i + b = 0} \delta_{(x_i^\star, \hat{a}_i, \hat{y}_i)},$$

for some $x_i^\star \in \{x \in \mathcal{X} : \|x - \hat{x}_i\| \leq \rho,\ \hat{y}_i(w^\top \hat{x}_i + b) < 0\}$. Notice that $\rho > 0$ and thus $x_i^\star$ exists for all $i \in [N]$. It is easy to verify that $\mathbb{Q}^\star \in \mathbb{B}(\hat{\mathbb{P}})$ and

$$\mathbb{Q}^\star(Y(w^\top X + b) < 0) = 1.$$

Remind that $\mathbb{B}(\hat{\mathbb{P}}) = \mathcal{B}_1(\hat{\mathbb{P}})$, this completes the proof. □

6. Numerical Experiment

In this section we present several numerical experiments to examine the distributionally robust fair SVM model. All optimization problems are implemented in Python 3.7 with package CVXPY 1.1.0 and solved by GUROBI 8.1. The experiments were run on a 2.2GHz Intel Core i7 CPU laptop with 8GB RAM.

6.1. **Synthetic Experiments.** We visualizes the standard SVM (SSVM), fair SVM (SSVM) and distributionally robust fair SVM (DRFSVM) on a toy dataset with 200 samples (50 for training, 150 for testing) and $d = 2$ features. We choose the ground cost with $\| \cdot \|$ being the l_∞-norm and $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$. The Wasserstein radius is set to $\rho = 0.1$ and the upper bound η is chosen to be 0.3.

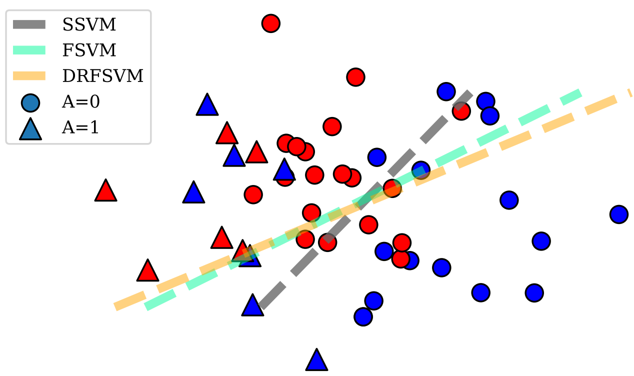


Figure 1. Classification hyperplanes (dashed) obtained by different approaches. Color decodes the labels (blue for +1 labels and red for -1 labels)

Figure 1 visualizes the effects of the unfairness constraint and the robustification on the hyperplane dictated by w and b. Notice that the sensitive attribute A (represented by cirles and triangles) is correlated with the feature X_1 on the horizontal axis. The FSVM and DRFSVM assign lower absolute value for the weight w_1 corresponding to feature X_1. Visually, this shift is reflected by the hyperplane of FSVM becoming more horizontal compared to that of SSVM. The DRFSVM, by being robust, shifts this hyperplane even more horizontal to reduce the dependency of the classifier on X_1.

We next test the unfairness measure and accuracy on the test dataset. The DRFSVM lowers the unfairness from 0.71 to 0.37 at the cost of reducing the accuracy from 66.7% to 6.3%. Details are exhibited in Table 1.

Classifier	Test accuracy	Test unfairness
SSVM	65.68%	0.7092
FSVM	63.27%	0.4414
DRFSVM	62.32%	0.3669

Table 1. Testing accuracy and unfairness on test data for the synthetic experiment.

In the second set of synthetic experiments, we compare the performance of our model against the approach DOB+ [20], which is considered as the state-of-the-art method in fair SVM. We plot the Pareto frontiers of the FSVM and DRFSVM against that of DOB+ in Figure 2. The setup for this experiment follows from the synthetic experiment in [68]. The data-generating probability distribution $\mathbb{P}$ satisfies $\mathbb{P}(Y = 1) = \mathbb{P}(Y =$

$-1) = 0.5$, while the conditional distribution of the 2-dimensional feature vectors are set as the following Gaussian distributions

$$X|Y = 1 \sim \mathcal{N}([2;2],[5,1;1,5], \qquad X|Y = -1 \sim \mathcal{N}([-2;-2],[10,1;1,3].$$

Next, we generate sensitive feature for each sample x from a Bernoulli distribution

$$\mathbb{P}(A = 1|X = x') = \frac{pdf(x'|Y = 1)}{pdf(x'|Y = 1) + pdf(x'|Y = -1)},$$

where $x' = [\cos(\pi/4), \sin(\pi/4); \sin(\pi/4), \cos(\pi/4)]x$ is a rotated value of the feature vector x and $pdf(\cdot|Y = y)$ is the Gaussian probability density function of $X|y = y$.

We then draw 200 samples from the data generating distribution $\mathbb{P}$, and then separate them into a group of 50 samples used for the training, while the remaining 150 samples are used as the test set. For the FSVM and DRFSVM model, we examine the models with different values of the unfairness controlling parameter η on $[0, 0.1]$ with 5 equidistant points. We fix the Wasserstein radius of the DRFSVM model to 0.1. Since the authors of the DOB+ method argue that $\eta = 0$ is a reasonable selection for the unfairness controlling parameter and their code is implemented under this prerequisite, to be consistent with their paper, we fix this parameter for 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' code. The described procedure is repeated 100 times independently, and the results are averaged over 100 trials.

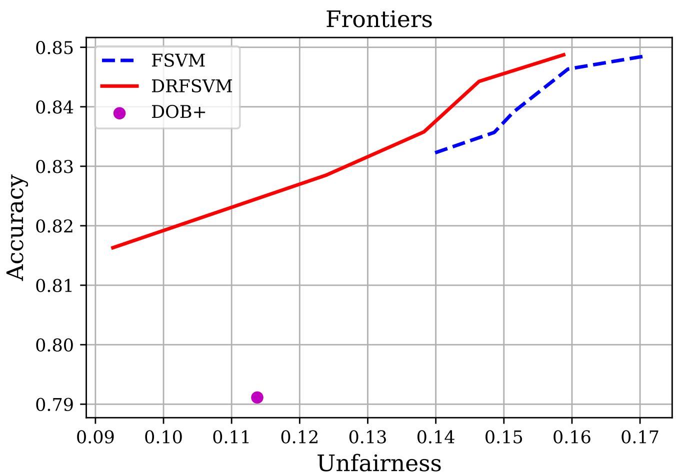


FIGURE 2. Unfairness-accuracy Pareto frontiers for different approaches.

Figure 2 shows that the DRFSVM solution outperforms the non-robust FSVM in the out-of-sample test. Compared to the DOB+ solution (purple dot), the DRFSVM achieves 3% higher accuracy of classification. At the extreme, DRFSVM achieves the lowest out-of-sample unfairness measure of 0.09, while the DOB+ and FSVM model only reach the unfairness value at 0.11 and 0.14, respectively.

6.2. **Experiments with Real Data.** We then assess the performance of our proposed DRFSVM model and demonstrate its superior performance on four publicly available datasets (Adult, Drug, COMPAS, Arrhythmia). A brief summary of these four datasets is presented in Table 2. While the Adult dataset has already been divided into the training and testing sets, we randomly select 2/3 samples for training and keep the rest of the data for testing in all other three datasets.

We first visualize the performance by plotting the Pareto frontiers of the FSVM, DRFSVM, and DOB+ models on the real datasets. We stratify 50 data points from the training set to train the models. We test the model with different values of η on $[0, 0.2]$ with 5 equidistant points. For DRFSVM, we set $\rho = 0.1$. All

Dataset	Features d	Sensitive Attribute A	Number of samples
Adult	12	Gender	32561, 12661
Drug	11	Ethnicity	1885
COMPAS	10	Ethnicity	6172
Arrhythmia	279(15)	Gender	452

TABLE 2. Datasets statistics and their sensitive feature. Gender considers the two groups as male and female; ethnicity considers the ethnic groups white and other ethnic groups. The adult dataset has pre-assigned training and test sets.

results are averaged over 50 independent trials. Figure 3 shows that the DRFSVM model admits the lowest out-of-sample unfairness in all the datasets. Moreover, the DRFSVM model can achieve the same level of accuracy as the DOB+ method with a smaller unfairness measure, which implies that the DRFSVM model can outperform the DOB+ method both in accuracy score and unfairness measure.

Next, we formally benchmark the models following a cross-validation, training, and testing procedure. The hyperparameter of DRFSVM, i.e., the radius of the Wasserstein ball, is determined in the crossing-validation procedure similar to [20]. We first split the training set into a sub-training set with $N = 150$ samples and keep the remaining samples as a sub-validation set. Then we collect statistics (i.e., accuracy score, unfairness measure) of $\rho \in [5 \cdot 10^{-3}, 5]$ on a logarithm searching grid with 30 discretization points based on the sub-training and sub-validation sets. Notice that the maximal value in the grid search for ρ equals to $2\kappa_{\mathcal{Y}}$, which is sufficient to induce the perturbation on both the label Y and the sensitive attribute A. This process is repeated $K_1 = 5$ times, and the average accuracy and unfairness are recorded for each candidate value. Finally, we select the value with the highest (Accuracy $-$ Unfairness) score from the list. Similarly, the tuning parameter C of the DOB+ method is also determined by cross-validation using the author's code.

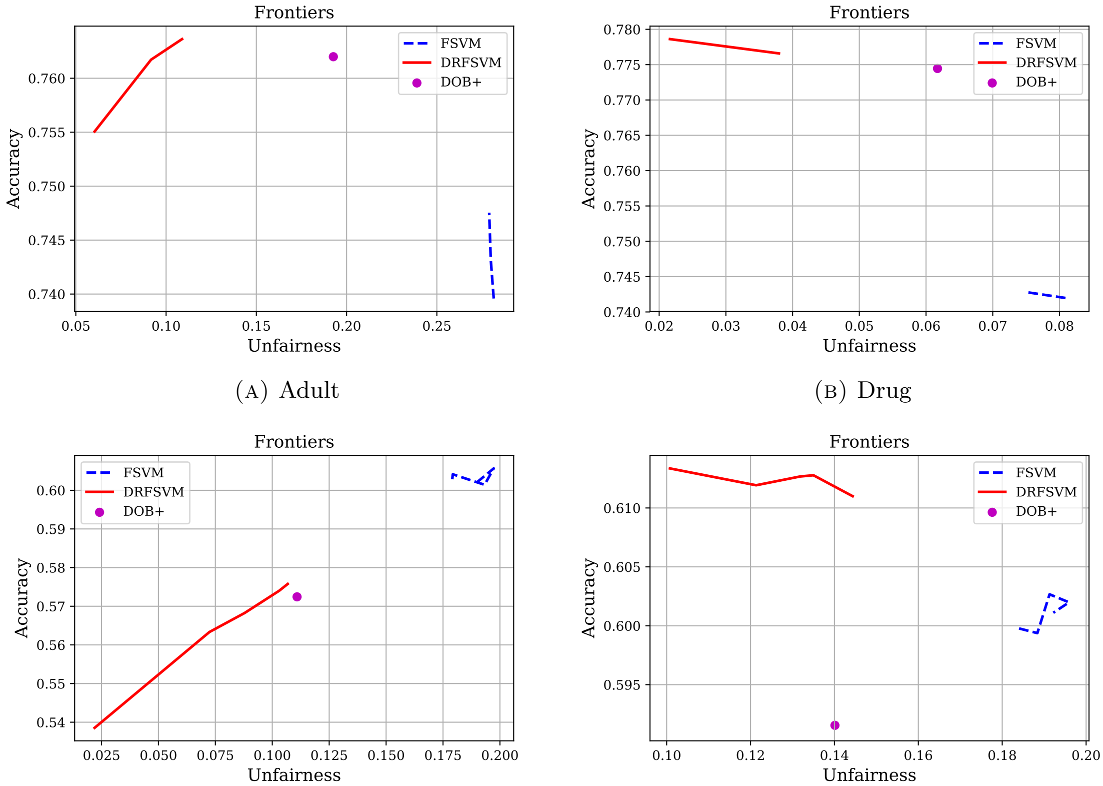
With the hyperparameters obtained from cross-validation, we now retrain the four classifiers using another random draw of N=150 samples from the training set. We set $\eta = 0.1$ for FSVM and DRFSVM, and $\kappa_{\mathcal{A}} = 2.5$, $\kappa_{\mathcal{Y}} = 2.5$, $\gamma = 0.01$ for DRFSVM. The DOB+ method is computed using the authors' code. The accuracy and unfairness measures of all classifiers are then evaluated on the testing set. We repeat this process for $K_2 = 100$ times and report the average accuracy scores and unfairness measures on Table 3.

Table 3 suggests that our DRFSVM classifier performs favorably relative to its competitors: it yields the lowest unfairness score across all four datasets with only a moderate loss in accuracy.

Dataset	Metric	SVM	FSVM	DOB+	DRFSVM
Adult	Accuracy	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$	0.78 ± 0.02	$\mathbf{0.79 \pm 0.02}$
	Unfairness	0.16 ± 0.10	0.14 ± 0.10	0.09 ± 0.08	$\mathbf{0.05 \pm 0.06}$
Drug	Accuracy	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.79 \pm 0.02}$	0.78 ± 0.02	$\mathbf{0.79 \pm 0.02}$
	Unfairness	0.15 ± 0.07	0.14 ± 0.07	0.10 ± 0.08	$\mathbf{0.07 \pm 0.08}$
COMPAS	Accuracy	$\mathbf{0.63 \pm 0.02}$	$\mathbf{0.63 \pm 0.02}$	0.57 ± 0.02	0.53 ± 0.00
	Unfairness	0.21 ± 0.04	0.21 ± 0.05	0.10 ± 0.05	$\mathbf{0.00 \pm 0.00}$
Arrhythmia	Accuracy	$\mathbf{0.65 \pm 0.03}$	0.64 ± 0.03	0.62 ± 0.04	0.61 ± 0.02
	Unfairness	0.22 ± 0.09	0.21 ± 0.07	0.11 ± 0.07	$\mathbf{0.06 \pm 0.06}$

TABLE 3. Test accuracy and unfairness (average $\pm$ standard deviation) for $N = 150$. The best results for each dataset is highlighted in bold.

6.3. **Solution time.** We now report the running time of different methods on 5 datasets (Adult, Drug, COMPAS, Arrhythmia, and Synthetic) with the sample size varying from 50 to 1000. We set the unfairness



(A) Adult (B) Drug

(C) COMPAS (D) Adult

FIGURE 3. Unfairness-accuracy Pareto frontiers of the real datasets.

controlling parameter $\eta = 0.1$, Wasserstein radius $\rho = 0.1$, and assume all samples are correctly labeled. All results are averaged over 10 independent trials.

Table 4 suggests that the DRFSVM model is applicable to moderate-size problems, where all datasets with the size of 1000 can be solved in about 10 minutes. The sample size is the factor that most affects the runtime because the number of binary variables increases proportionally with the size of the dataset. The speed of FSVM is almost the same as DRFSVM, as both of them involve solving a conic mixed-binary program. The DOB+ method achieves the highest efficiency across all datasets. For most datasets, the DOB+ method is solved in a few seconds, and only in the Arrhythmia dataset, the DOB+ method takes about 50 seconds to achieve optimality. Because the datasets have different feature size d, we also observe that the running times of high-dimension datasets (e.g., Arrhythmia) are generally greater than that of low-dimension datasets (e.g., the synthetic dataset).

7. Concluding remarks

We develope a new principled approach to fair SVM 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 utilize the type-∞ Wassertein ambiguity set, which enables a more scalable mixed-binary linear programming reformulation while providing the same statistical performance guarantees as the models based on the standard Wasserstein ambiguity sets. Our proposed model is general and can even handle problem instances with noisy and adversarial sensitive attributes and labels.

In this paper, we delineated for the first time the relationship between the SVM objective and the misclassification probability of the resulting classifier. Our theoretical result implies that the proposed fair classification

		Sample size N					
Dataset	Classifier	50	100	250	500	750	1000
Adult	DRFSVM	0.40	3.24	4.02	93.28	246.24	479.72
	FSVM	0.15	2.65	6.82	85.80	230.94	409.68
	DOB+	0.02	0.03	0.07	0.16	0.35	0.43
Drug	DRFSVM	0.32	4.32	13.34	127.93	234.08	448.27
	FSVM	0.30	3.87	36.39	145.37	233.70	421.48
	DOB+	0.02	0.03	0.07	0.15	0.19	0.30
COMPAS	DRFSVM	0.31	2.04	52.34	139.38	348.08	687.47
	FSVM	0.52	1.59	26.39	113.95	347.47	786.70
	DOB+	0.02	0.03	0.02	0.15	0.18	0.16
Arrhythmia	DRFSVM	0.06	18.81	473.19			
	FSVM	0.55	16.06	404.18			
	DOB+	3.49	18.19	54.30			
Synthetic	DRFSVM	0.02	0.30	2.12	82.41	184.39	244.96
	FSVM	0.47	0.54	1.52	71.43	143.54	223.21
	DOB+	0.18	0.05	0.34	1.85	0.71	1.97

TABLE 4. Running time (in seconds) of different methods. The Arrhythmia dataset only contains 452 examples, hence we only examine its performance up to $N = 250$.

model constitutes a conservative approximation of the actual model which seeks a classifier that minimizes the misclassification probability. We remark that this exact model is also amenable to a mixed-binary linear programming reformulation. However, preliminary experimental results indicate that the reformulation is not as efficiently solvable as the plain-vanilla model that minimizes the hinge loss objective. Thus, in the future, we plan to develop fast, tailored algorithms for the reformulation model.

8. Appendix – Auxiliary Results

Lemma 8.1 (Compactness)**.** The set $\mathbb{B}(\hat{\mathbb{P}})$ defined in (3) is weakly compact and convex. More specifically, there exists a convex, compact set $\mathbb{X} \in \mathcal{X}$ defined as

$$\mathbb{X} = \text{ConvexHull}\big(\{x \in \mathcal{X} : \|x - \hat{x}_i\| \le \rho\}_{i=1}^N\big)$$

such that $\mathbb{Q}(\mathbb{X} \times \mathcal{A} \times \mathcal{Y}) = 1$ for any $\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}})$.

Proof of Lemma 8.1. Because the $\hat{\mathbb{P}}$ is an empirical measure, the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ can be represented as

$$\mathbb{B}(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) : \begin{array}{l} \exists \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \ \forall i \in [N] \text{ such that :} \\ \mathbb{Q} = N^{-1} \sum_{i \in [N]} \pi_i, \\ \|x_i - \hat{x}_i\| + \kappa_{\mathcal{A}} |a_i - \hat{a}_i| + \kappa_{\mathcal{Y}} |y_i - \hat{y}_i| \le \rho \quad \forall (x_i, a_i, y_i) \in \text{supp}(\pi_i) \quad \forall i \in [N] \\ N^{-1} \sum_{i \in [N]} \pi_i(A = a, Y = y) = \hat{p}_{ay} \quad \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\}, \tag{16}$$

where $\text{supp}(\pi_i)$ denotes the support of the probability measure π_i [2, Page 441]. Pick any arbitrary $\mathbb{Q}^0$ and $\mathbb{Q}^1$ from $\mathbb{B}(\hat{\mathbb{P}})$. Associated with $\mathbb{Q}^j$, $j \in \{0, 1\}$ is a collection of conditional probability measures $\{\pi_i^j\} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})^N$ satisfying

$$\begin{cases} \mathbb{Q}^j = N^{-1} \sum_{i \in [N]} \pi_i^j \\ \|x_i - \hat{x}_i\| + \kappa_{\mathcal{A}} |a_i - \hat{a}_i| + \kappa_{\mathcal{Y}} |y_i - \hat{y}_i| \le \rho & \forall (x_i, a_i, y_i) \in \text{supp}(\pi_i^j), \quad \forall i \in [N] \\ N^{-1} \sum_{i \in [N]} \pi_i^j(A = a, Y = y) = \hat{p}_{ay} & \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \end{cases}$$

Consider any convex combination $\mathbb{Q}^\lambda = \lambda \mathbb{Q}^1 + (1-\lambda)\mathbb{Q}^0$ for $\lambda \in (0,1)$. It is easy to verify that the measure $\pi_i^\lambda = \lambda \pi_i^1 + (1-\lambda)\pi_i^0$ for any $i \in [N]$ satisfies

$$\begin{cases} \mathbb{Q}^\lambda = N^{-1} \sum_{i\in[N]} \pi_i^\lambda \\ \|x_i - \hat{x}_i\| + \kappa_{\mathcal{A}} |a_i - \hat{a}_i| + \kappa_{\mathcal{Y}} |y_i - \hat{y}_i| \le \rho & \forall (x_i, a_i, y_i) \in \mathrm{supp}(\pi_i^\lambda), \quad \forall i \in [N] \\ N^{-1} \sum_{i\in[N]} \pi_i^\lambda (A = a, Y = y) = \hat{p}_{ay} & \forall (a,y) \in \mathcal{A}\times\mathcal{Y}, \end{cases}$$

where the middle constraint is satisfied by noticing that $\mathrm{supp}(\pi_i^\lambda) = \mathrm{supp}(\pi_i^0) \cup \mathrm{supp}(\pi_i^1)$. This observation implies that $\mathbb{Q}^\lambda \in \mathbb{B}(\hat{\mathbb{P}})$.

Notice that for any feasible measure π_i, we have

$$\mathrm{supp}(\pi_i) \subseteq \{x \in \mathcal{X} : \|x - \hat{x}_i\| \le \rho\} \times \mathcal{A} \times \mathcal{Y},$$

and as a consequence, we have

$$\mathrm{supp}(\mathbb{Q}) \subseteq \bigcup_{i\in[N]} \{x \in \mathcal{X} : \|x - \hat{x}_i\| \le \rho\} \times \mathcal{A} \times \mathcal{Y}.$$

By definition of $\mathbb{X}$, we have $\bigcup_{i\in[N]} \{x \in \mathcal{X} : \|x - \hat{x}_i\| \le \rho\} \subseteq \mathbb{X}$. Because $\mathbb{X}$ is a compact set, the weakly compactness of $\mathbb{B}(\hat{\mathbb{P}})$ follows from Prohorov's theorem. This completes the proof. □

The result of Lemma 8.1 also extends to the ambiguity set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ defined as in (10).

Corollary 8.2 (Compactness)**.** For any $\gamma \in [0,1]$, the set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ defined in (10) is weakly compact and convex. More specifically, there exists a compact set $\mathbb{X} \in \mathcal{X}$ defined as

$$\mathbb{X} = \mathrm{ConvexHull}\big(\{x \in \mathcal{X} : \|x - \hat{x}_i\| \le \rho\}_{i=1}^N\big)$$

such that $\mathbb{Q}(\mathbb{X} \times \mathcal{A} \times \mathcal{Y}) = 1$ for any $\mathbb{Q} \in \mathcal{B}_\gamma(\hat{\mathbb{P}})$.

The proof of Corollary 8.2 follows a similar line of argument as the proof of Lemma 8.1 by noticing that $\sum_{i\in[N]} \pi_i(A = \hat{a}_i, Y = \hat{y}_i) \ge (1-\gamma)N$ is a convex constraint for π_i.

Lemma 8.3 (Reformulation of $\mathbb{B}(\hat{\mathbb{P}})$)**.** The set $\mathbb{B}(\hat{\mathbb{P}})$ defined in (3) can be equivalently written as

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

Proof of Lemma 8.3. Notice that the condition

$$\|x_i - \hat{x}_i\| + \kappa_{\mathcal{A}} |a_i - \hat{a}_i| + \kappa_{\mathcal{Y}} |y_i - \hat{y}_i| \le \rho \quad \forall (x_i, a_i, y_i) \in \mathrm{supp}(\pi_i) \quad \forall i \in [N]$$

is equivalent to the condition

$$\mathrm{W}_\infty(\pi_i, \delta_{(\hat{x}_i, \hat{a}_i, \hat{y}_i)}) \le \rho \qquad \forall i \in [N]$$

by the definition of the type-∞ Wasserstein distance W_∞. Replacing latter condition into (16) finishes the proof. □

References

[1] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[2] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Accessed: 2020-09-14.

[4] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020.

[5] Y. Bechavod and K. Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

[6] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

[7] D. Bertsimas, S. Shtern, and B. Sturt. A data-driven approach for multi-stage linear optimization. *Available from Optimization Online*, 2018.

[8] D. Bertsimas, S. Shtern, and B. Sturt. Two-stage sample robust optimization. *arXiv preprint arXiv:1907.07142*, 2019.

[9] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[10] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[11] I. Bose and R. K. Mahapatra. Business data mining—a machine learning perspective. *Information & management*, 39(3):211–225, 2001.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[13] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[14] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[15] L. Chang. Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, 131:53–68, 2006.

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

[17] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[18] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Accessed: 2020-09-14.

[19] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[20] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[21] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

[22] M. E. Dyer and A. M. Frieze. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5):967–974, 1988.

[23] H. Edwards and A. Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[24] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 259–268, 2015.

[25] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[26] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[27] C. Givens and R. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[28] P. Gordaliza, E. D. Barrio, G. Fabrice, and J.-M. Loubes. Obtaining fairness using optimal transport theory. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2357–2365, 2019.

[29] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pages 3315–3323, 2016.

[30] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[31] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1929–1938, 2018.

[32] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[33] N. Ho-Nguyen and S. J. Wright. Adversarial classification via distributional robustness with Wasserstein ambiguity. *arXiv preprint arXiv:2005.13815*, 2020.

[34] L. Huang and N. Vishnoi. Stable and fair classification. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2879–2890, 2019.

[35] D. Kabakchieva. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.

[36] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[37] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

[38] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, 2011.

[39] H. Kannan, A. Kurakin, and I. Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[40] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[41] S. Lohr. Big data, trying to build better workers. *The New York Times.* Accessed: 2020-09-14.

[42] B. T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 502–510, 2011.

[43] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3384–3393, 2018.

[44] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[45] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

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

[47] MOSEK ApS. *MOSEK Optimizer API for Python 9.2.10*, 2019.

[48] V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*, 2018.

[49] V. A. Nguyen, F. Zhang, J. Blanchet, E. Delage, and Y. Ye. Distributionally robust local non-parametric conditional estimation. In *Advances in Neural Information Processing Systems 33*, 2020.

[50] Z. Obermeyer and E. J. Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[51] N. Quadrianto and V. Sharmanska. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems 30*, pages 677–688, 2017.

[52] A. Rezaei, R. Fathony, O. Memarrast, and B. Ziebart. Fairness for robust log loss classification. In *AAAI Conference on Artificial Intelligence*, 2020.

[53] R. T. Rockafellar and S. Uryasev. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2:21–41, 2000.

[54] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*, pages 10976–10987, 2018.

[55] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[56] M. J. Shaw and J. A. Gentry. Using an expert system with inductive learning to evaluate business loans. *Financial Management*, pages 45–56, 1988.

[57] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74, 2002.

[58] B. Taskesen, J. Blanchet, D. Kuhn, and V. A. Nguyen. A statistical test for probabilistic fairness. *Accepted to ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[59] B. Taskesen, V. A. Nguyen, D. Kuhn, and J. Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

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

[61] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020.

[62] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1920–1953, 2017.

[63] W. Xie. Tractable reformulations of distributionally robust two-stage stochastic programs with ∞-Wasserstein distance. *arXiv preprint arXiv:1908.08454*, 2019.

[64] Q. Ye and W. Xie. Unbiased subdata selection for fair classification: A unified framework and scalable algorithms. *arXiv preprint arXiv:2012.12356*, 2020.

[65] M. Yurochkin, A. Bower, and Y. Sun. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations*, 2020.

[66] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017.

[67] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

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

[69] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[70] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.