## Appendix A. Auxiliary Results and Proofs

**Lemma A.1** (Compactness)**.** The set $\mathbb{B}(\hat{\mathbb{P}})$ defined in (4) is weakly compact and convex. More specifically, there exists a convex, compact set $\mathbb{X} \in \mathcal{X}$ defined as

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

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 A.1.* Because $\hat{\mathbb{P}}$ is an empirical measure, the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ can be represented as

$$\mathbb{B}(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) : \begin{array}{l} \exists \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})\ \forall i \in [N] \text{ such that :} \\ \mathbb{Q} = N^{-1} \sum_{i\in[N]} \pi_i, \\ \|x_i - \hat{x}_i\| + \kappa_{\mathcal{A}} |a_i - \hat{a}_i| + \kappa_{\mathcal{Y}} |y_i - \hat{y}_i| \leq \rho \quad \forall (x_i, a_i, y_i) \in \text{supp}(\pi_i) \quad \forall i \in [N], \\ N^{-1} \sum_{i\in[N]} \pi_i(A = a, Y = y) = \hat{p}_{ay} \quad \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\}, \tag{A1}$$

where $\text{supp}(\pi_i)$ denotes the support of the probability measure $\pi_i$ [1, 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

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

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

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

where the middle constraint is satisfied by noticing that $\text{supp}(\pi_i^\lambda) = \text{supp}(\pi_i^0) \cup \text{supp}(\pi_i^1)$. This observation implies that $\mathbb{Q}^\lambda \in \mathbb{B}(\hat{\mathbb{P}})$. Notice that for any feasible measure $\pi_i$, we have

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

and as a consequence, we have

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

By definition of $\mathbb{X}$, we have $\bigcup_{i\in[N]} \{x \in \mathcal{X} : \|x - \hat{x}_i\| \leq \rho\} \subseteq \mathbb{X}$. 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 A.1 also extends to the ambiguity set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ defined as in (15).

**Corollary A.2** (Compactness)**.** For any $\gamma \in [0, 1]$, the set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ defined in (15) is weakly compact and convex. More specifically, there exists a compact set $\mathbb{X} \in \mathcal{X}$ defined as

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

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 A.2 follows a similar line of argument as the proof of Lemma A.1 by noticing that $\sum_{i\in[N]} \pi_i(A=\hat{a}_i, Y=\hat{y}_i) \geq (1-\gamma)N$ is a convex constraint for $\pi_i$.

**Lemma A.3** (Reformulation of $\mathbb{B}(\hat{\mathbb{P}})$)**.** The set $\mathbb{B}(\hat{\mathbb{P}})$ defined in (4) can be equivalently written as

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

*Proof of Lemma A.3.* Notice that the condition

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

is equivalent to the condition

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

by the definition of the type-$\infty$ Wasserstein distance. Replacing the condition into (A1) finishes the proof. □

## A.1. Auxiliary Results and Proofs of Section 3.

*Proof of Proposition 3.1.* We first define the events $S = \{x : w^\top x + b \geq 0\}$ and $S_\varepsilon = \{x : w^\top x + b > -\varepsilon\}$. As $\varepsilon$ tends to zero, we have $S = \lim_{\varepsilon\to 0} S_\varepsilon$. Moreover, we see that $S_\varepsilon$ is non-increasing as $\varepsilon \to 0$. Thus, for any distribution $\mathbb{Q}$, we have [4, Lemma 5]

$$\mathbb{Q}(S) = \lim_{\varepsilon\to 0} \mathbb{Q}(S_\varepsilon).$$

Plugging this result into the $\varepsilon$-unfairness measure yields

$$\begin{aligned} \lim_{\varepsilon\to 0} \mathbb{U}_\varepsilon(w,b,\mathbb{Q}) &= \lim_{\varepsilon\to 0} \mathbb{Q}_{11}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{01}(w^\top X + b \geq 0) \\ &= \mathbb{Q}_{11}(w^\top X + b \geq 0) - \mathbb{Q}_{01}(w^\top X + b \geq 0) = \mathbb{U}(w,b,\mathbb{Q}). \end{aligned}$$

The proof of the objective function is the same and thus omitted. □

*Proof of Proposition 3.2.* By definition, we find

$$\begin{aligned} \mathbb{U}(w,b,\mathbb{Q}) =& \mathbb{Q}_{11}(w^\top X + b \geq 0) - \mathbb{Q}_{01}(w^\top X + b \geq 0) \\ \leq& \mathbb{Q}_{11}(w^\top X + b > -\varepsilon) - \mathbb{Q}_{01}(w^\top X + b \geq 0) = \mathbb{U}_\varepsilon(w,b,\mathbb{Q}), \end{aligned}$$

for every possible value of the classifier parameter $(w,b)$ and any distribution $\mathbb{Q}$. As a consequence, the feasible region of problem (8) is an *inner* approximation of the feasible region of problem (5) for any $\varepsilon \in \mathbb{R}_{++}$. This implies that the optimal solution $(w^\star, b^\star)$ of problem (8) is also feasible for problem (5).

Furthermore, one can verify that for any $(w,b)\in\mathbb{R}^{d+1}$ and any distribution $\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})$,

$$\mathbb{Q}(Y(w^\top X + b) \leq 0) \leq \sup_{\mathbb{Q}'\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}'(Y(w^\top X + b) \leq 0) \leq \sup_{\mathbb{Q}'\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}'(Y(w^\top X + b) < \varepsilon).$$

Thus, by plugging in the optimal solution $(w^\star, b^\star)$ we have

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

which completes the proof. □

*Proof of Lemma 3.4.* Using an epigraphical formulation of each supremum term, we find

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

$$= \begin{cases} \min & \sum_{k\in\mathcal{K}} \lambda_k \\ \text{s.t.} & \lambda \in \{0,1\}^N, \\ & \max_{x_k:\|x_k-\hat{x}_k\|\le\rho} w^\top x_k + b - \varepsilon \le M\lambda_k \quad \forall k \in \mathcal{K}, \end{cases}$$

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

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

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

*Proof of Remark 3.5.* For any sample $i \in [N]$, we have

$$\begin{aligned} w^\top \hat{x}_i + \rho\|w\|_* + b + \varepsilon &= (w,b)^\top(\hat{x}_i, 1) + \rho\|w\|_* + \varepsilon \\ &\le \|(w,b)\|\|(\hat{x}_i,1)\|_* + \rho\|w\|_* + \varepsilon \\ &\le \|(w,b)\|\|(\hat{x}_i,1)\|_* + \rho\|(w,b)\|_* + \varepsilon, \end{aligned}$$

where the first inequality follows from Cauchy–Schwarz inequality [9]. Let $\|\cdot\|$ be the $p$-norm with $p \ge 1$. We know that the dual norm of the $p$-norm is the $q$-norm with $q = \frac{p}{p-1}$ [8]. By Hölder's Inequality [5], when $1 \le p \le 2$, we have

$$\|\cdot\|_* \le \|\cdot\|;$$

and when $p > 2$, we have

$$\|\cdot\|_* \le d^{(\frac{p-1}{p} - \frac{1}{p})} \|\cdot\| \le d\|\cdot\|.$$

Thus, for any $p \ge 1$, we have

$$\|\cdot\|_* \le d\|\cdot\|,$$

and we can further obtain

$$w^\top \hat{x}_i + \rho\|w\|_* + b + \varepsilon \le \|(w,b)\|\|(\hat{x}_i,1)\|_* + \rho\|(w,b)\|_* + \varepsilon \le C + \rho d + \varepsilon \le M.$$

□

*Proof of Remark 3.7.* Maximizing balanced accuracy is equivalent to minimizing the balanced misclassification rate. Therefore, we can easily modify the distributionally robust fair classification problem (8) by

$$\begin{aligned} \min \quad & \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{Q}(Y(w^\top X + b) < \varepsilon | Y = 1)/2 + \mathbb{Q}(Y(w^\top X + b) < \varepsilon | Y = -1)/2 \\ \text{s.t.} \quad & w \in \mathbb{R}^d,\ b \in \mathbb{R},\ \|(w,b)\| \le 1, \\ & \sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{U}_\varepsilon(w,b,\mathbb{Q}) \le \eta. \end{aligned} \tag{A2}$$

where the objective function minimizes the averaged accuracy score computed from the positive and negative classes. □

The following proposition shows that problem (A2) can be reformulated as a mixed binary conic program.

**Proposition A.4** (Balanced misclassification rate reformulation)**.** Let $\mathcal{I}_1 = \{i \in [N] : \hat{y}_i = 1\}$ and $\mathcal{I}_{-1} = \{i \in [N] : \hat{y}_i = -1\}$ be the index sets of positive and negative labels, respectively. The balanced misclassification rate minimization problem (A2) is equivalent to the mixed binary conic optimization problem

$$\begin{array}{lll}
\min & \dfrac{1}{2|\mathcal{I}_1|}\displaystyle\sum_{i\in\mathcal{I}_1} t_i + \dfrac{1}{2|\mathcal{I}_{-1}|}\sum_{i\in\mathcal{I}_{-1}} t_i & \\
\text{s.t.} & w\in\mathbb{R}^d,\ b\in\mathbb{R},\ t\in\{0,1\}^N,\ \lambda^0\in\{0,1\}^N,\ \lambda^1\in\{0,1\}^N,\ \|(w,b)\|\le 1, & \\
& -\hat{y}_i(w^\top\hat{x}_i+b)+\rho\|w\|_*\le Mt_i-\varepsilon & \forall i\in[N],\\
& \dfrac{1}{|\mathcal{I}_{11}|}\displaystyle\sum_{i\in\mathcal{I}_{11}}\lambda_i+\dfrac{1}{|\mathcal{I}_{01}|}\sum_{i\in\mathcal{I}_{01}}\lambda_i-1\le\eta, & \\
& w^\top\hat{x}_i+\rho\|w\|_*+b\le M\lambda_i & \forall i\in\mathcal{I}_{11},\\
& -w^\top\hat{x}_i+\rho\|w\|_*-b\le M\lambda_i & \forall i\in\mathcal{I}_{01}
\end{array} \tag{A3}$$

*Proof of Proposition A2.* The reformulation of the fairness constraint is exactly the same as model (10) in the main text, hence we only show the derivation of the new objective function.

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})}\ \mathbb{Q}(Y(w^\top X+b)<\varepsilon|Y=1)/2+\mathbb{Q}(Y(w^\top X+b)<\varepsilon|Y=-1)/2\\
=&\sup_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})}\ \frac{1}{2}\mathbb{E}_{\mathbb{Q}}[\hat{p}_1^{-1}\mathbb{I}(Y(w^\top X+b)<\varepsilon)\mathbb{I}(Y=1)+\hat{p}_{-1}^{-1}\mathbb{I}(Y(w^\top X+b)<\varepsilon)\mathbb{I}(Y=-1)]\\
=&\frac{1}{N}\left(\hat{p}_1^{-1}\sum_{i\in\mathcal{I}_1}\sup_{x_i:\|x_i-\hat{x}_i\|\le\rho} I(Y(w^\top X+b)<\varepsilon)+\hat{p}_{-1}^{-1}\sum_{i\in\mathcal{I}_{-1}}\sup_{x_i:\|x_i-\hat{x}_i\|\le\rho} I(Y(w^\top X+b)<\varepsilon)\right)\\
=&\left\{\begin{array}{ll}
\min & \dfrac{1}{2|\mathcal{I}_1|}\displaystyle\sum_{i\in\mathcal{I}_1} t_i+\dfrac{1}{2|\mathcal{I}_{-1}|}\sum_{i\in\mathcal{I}_{-1}} t_i\\
\text{s.t.} & t\in\{0,1\}^N,\\
& -\hat{y}_i(w^\top\hat{x}_i+b)+\rho\|w\|_*\le Mt_i-\varepsilon\quad\forall i\in[N],
\end{array}\right.
\end{aligned}$$

where $\hat{p}_1$ and $\hat{p}_{-1}$ denote the probability that a sample belongs to the positive or negative classes, respectively. Combining the outer minimization problem completes the proof. □

The deterministic reformulation (A3) involves $2N$ binary variables, which shares the same complexity as reformulation (10). Thus, for moderate-size imbalanced data, decision makers can consider adopting model (A2) to generate more balanced results.

## A.2. **Auxiliary Results and Proofs of Section 4.**

*Proof of Proposition 4.1.* By definition, we find

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

for every possible value of the classifier parameter $(w,b)$ and any distribution $\mathbb{Q}$. Furthermore, one can verify that for any $(w,b)\in\mathbb{R}^{d+1}$ and $\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})$

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

Thus, by plugging in the optimal solution $(w^\star, b^\star)$, we have

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

which completes the proof. □

*Proof of Remark 4.3.* Similar to Remark 3.7, we can modify the hinge distributionally robust fair classification problem (12) by

$$\begin{array}{ll} \min & \sup\limits_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(w^\top X + b)\}|Y = 1]/2 + \mathbb{E}_{\mathbb{Q}}[\max\{0, 1 - Y(w^\top X + b)\}|Y = -1]/2 \\ \text{s.t.} & w \in \mathbb{R}^d,\ b \in \mathbb{R}, \\ & \sup\limits_{\mathbb{Q}\in\mathbb{B}(\hat{\mathbb{P}})} \mathbb{H}(w, b, \mathbb{Q}) \leq \zeta. \end{array} \tag{A4}$$

The derivation of the reformulation parallels to Remark 3.7; thus, we omit it for brevity. □

A.3. **Auxiliary Results and Proofs of Section 5.** In this section, we will provide the proof of Theorem 5.1. This proof leverages the following duality result.

**Lemma A.5.** (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

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

admits the following dual form

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

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

*Proof of Lemma A.5.* Using the definition of the type-$\infty$ Wasserstein distance, we can re-express the ambiguity set $\mathcal{B}_\gamma(\hat{\mathbb{P}})$ as

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

The worst-case expected loss can now be written as

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

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

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

Moreover, define the following optimal values

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

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 $\pi$, we obtain

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

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

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

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

Equipped with the duality result of Lemma A.5, we now present the proof of Theorem 5.1.

*Proof of Theorem 5.1.* 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 $\phi(X,A,Y)=\mathbb{I}\left(Y(w^\top X+b)<\varepsilon\right)$ is an indicator function. By Lemma A.5, we have

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

Based on Lemma 3.4, the constraint in the above infimum problem is equivalent to

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

Next, we show the derivation for constraints. Employing the result of Lemma A.5 yields

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})}\ \mathbb{Q}(w^\top X+b>-\varepsilon|A=1,Y=1)-\mathbb{Q}(w^\top X+b\ge 0|A=0,Y=1)\\
=&\sup_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})}\ \mathbb{E}_{\mathbb{Q}}[\hat{p}_{11}^{-1}\mathbb{I}(w^\top X+b>-\varepsilon)\mathbb{1}_{(1,1)}(A,Y)-\hat{p}_{01}^{-1}\mathbb{I}(w^\top X+b\ge 0)\mathbb{1}_{(0,1)}(A,Y)]\\
=&\left\{\begin{array}{ll}
\inf & \dfrac{1}{N}\displaystyle\sum_{i\in[N]}\nu'_i+\sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}}\hat{p}_{ay}\mu'_{ay}-\theta'(1-\gamma)\\
\text{s.t.} & \nu'\in\mathbb{R}^N,\ \theta'\in\mathbb{R}_+,\ \mu'\in\mathbb{R}^{2\times 2},\\
& \displaystyle\sup_{\forall x\in\mathcal{X}:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|a-\hat{a}_i|-\kappa_{\mathcal{Y}}|y-\hat{y}_i|}\phi(x,a,y)\le\mu'_{ay}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y)+\nu'_i\\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall i\in[N]\quad\forall(a,y)\in\mathcal{A}\times\mathcal{Y},
\end{array}\right.
\end{aligned}
$$

where the second equation relies on the result of Lemma A.5 by defining

$$
\phi(X,A,Y)=\hat{p}_{11}^{-1}\mathbb{I}(w^\top X+b>-\varepsilon)\mathbb{1}_{(1,1)}(A,Y)-\hat{p}_{01}^{-1}\mathbb{I}(w^\top X+b\ge 0)\mathbb{1}_{(0,1)}(A,Y).
$$

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

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

$$
\begin{aligned}
&\sup_{\forall x\in\mathcal{X}:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|1-\hat{a}_i|-\kappa_{\mathcal{Y}}|1-\hat{y}_i|}\hat{p}_{11}^{-1}\mathbb{I}(w^\top x+b>-\varepsilon)\mathbb{1}_{(1,1)}(1,1)-\hat{p}_{01}^{-1}\mathbb{I}(w^\top x+b\ge 0)\mathbb{1}_{(0,1)}(1,1)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\le\mu'_{a1}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,1)+\nu'_i\\
\iff&\sup_{\forall x\in\mathcal{X}:\|x-\hat{x}_i\|\le\rho-\kappa_{\mathcal{A}}|1-\hat{a}_i|-\kappa_{\mathcal{Y}}|1-\hat{y}_i|}\hat{p}_{11}^{-1}\mathbb{I}(w^\top x+b>-\varepsilon)\le\mu'_{11}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,1)+\nu'_i\\
\iff&\left\{\begin{array}{l}
\lambda_i^1\in\{0,1\},\\
\hat{p}_{11}^{-1}\lambda_i^1\le\mu'_{11}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,1)+\nu'_i,\\
w^\top\hat{x}_i+(\rho-\kappa_{\mathcal{A}}|1-\hat{a}_i|-\kappa_{\mathcal{Y}}|1-\hat{y}_i|)\|w\|_*+b\le M\lambda_i^1-\varepsilon,
\end{array}\right.
\end{aligned}
$$

where the last equation follows from the result of Lemma 3.4.

(2) Case 2: $(a,y)=(0,1)$. There is an active constraint if $\kappa_{\mathcal{A}}|0-\hat{a}_i|+\kappa_{\mathcal{Y}}|1-\hat{y}_i|\leq\rho$, and the constraint is equivalent to

$$\sup_{\forall x\in\mathcal{X}:\|x-\hat{x}_i\|\leq\rho-\kappa_{\mathcal{A}}|0-\hat{a}_i|-\kappa_{\mathcal{Y}}|1-\hat{y}_i|}\hat{p}_{11}^{-1}\mathbb{I}(w^\top x+b>-\varepsilon)\mathbb{1}_{(1,1)}(0,1)-\hat{p}_{01}^{-1}\mathbb{I}(w^\top x+b\geq 0)\mathbb{1}_{(0,1)}(0,1)$$
$$\leq\mu'_{01}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,1)+\nu'_i$$
$$\Longleftrightarrow-\hat{p}_{01}^{-1}\left(1-\sup_{\forall x\in\mathcal{X}:\|x-\hat{x}_i\|\leq\rho-\kappa_{\mathcal{A}}|0-\hat{a}_i|-\kappa_{\mathcal{Y}}|1-\hat{y}_i|}\mathbb{I}(w^\top x+b<0)\right)\leq\mu'_{01}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,1)+\nu'_i$$
$$\Longleftrightarrow\begin{cases}\lambda_i^0\in\{0,1\},\\ \hat{p}_{01}^{-1}(\lambda_i^0-1)\leq\mu'_{01}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,1)+\nu'_i,\\ -w^\top\hat{x}_i+(\rho-\kappa_{\mathcal{A}}|0-\hat{a}_i|-\kappa_{\mathcal{Y}}|1-\hat{y}_i|)\|w\|_*-b\leq M\lambda_i^0.\end{cases}$$

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

$$\sup_{\forall x\in\mathcal{X}:\|x-\hat{x}_i\|\leq\rho-\kappa_{\mathcal{A}}|1-\hat{a}_i|-\kappa_{\mathcal{Y}}|-1-\hat{y}_i|}\hat{p}_{11}^{-1}\mathbb{I}(w^\top x+b>-\varepsilon)\mathbb{1}_{(1,1)}(1,-1)-\hat{p}_{01}^{-1}\mathbb{I}(w^\top x+b\geq 0)\mathbb{1}_{(0,1)}(1,-1)$$
$$\leq\mu'_{1,-1}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,-1)+\nu'_i$$
$$\Longleftrightarrow 0\leq\mu'_{1,-1}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,-1)+\nu'_i.$$

(4) Case 4: $(a,y)=(0,-1)$. There is an active constraint if $\kappa_{\mathcal{A}}|0-\hat{a}_i|+\kappa_{\mathcal{Y}}|-1-\hat{y}_i|\leq\rho$, the constraint is equivalent to

$$\sup_{\forall x\in\mathcal{X}:\|x-\hat{x}_i\|\leq\rho-\kappa_{\mathcal{A}}|0-\hat{a}_i|-\kappa_{\mathcal{Y}}|-1-\hat{y}_i|}\hat{p}_{11}^{-1}\mathbb{I}(w^\top x+b>-\varepsilon)\mathbb{1}_{(1,1)}(0,-1)-\hat{p}_{01}^{-1}\mathbb{I}(w^\top x+b\geq 0)\mathbb{1}_{(0,1)}(0,-1)$$
$$\leq\mu'_{0,-1}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,-1)+\nu'_i$$
$$\Longleftrightarrow 0\leq\mu_{0,-1}-\theta'\mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,-1)+\nu'_i.$$

Notice that at least one of the above four conditions will be satisfied because when $a=\hat{a}_i$ and $y=\hat{y}_i$, we have $\kappa_{\mathcal{A}}|a-\hat{a}_i|+\kappa_{\mathcal{Y}}|y-\hat{y}_i|=0\leq\rho$ 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 $\eta$ for each value of $a\in\mathcal{A}$. This completes the proof. □

The general ground metric (14) defined in Section 5 can also be applied to the HDRFC model (12) derived in Section 4. We consider now the following modification of problem (12) in which the ambiguity set is replaced by $\mathcal{B}_\gamma(\hat{\mathbb{P}})$:

$$\begin{array}{ll}\min & \sup\limits_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})}\ \mathbb{E}_{\mathbb{Q}}[\max\{0,1-Y(w^\top X+b)\}]\\ \text{s.t.} & w\in\mathbb{R}^d,\ b\in\mathbb{R},\\ & \sup\limits_{\mathbb{Q}\in\mathcal{B}_\gamma(\hat{\mathbb{P}})}\ \mathbb{H}(w,b,\mathbb{Q})\leq\zeta.\end{array}\tag{A6}$$

We now present the main result of this section, which provides the reformulation for problem (A6).

**Theorem A.6.** (HDRFC reformulation) Suppose that the ground metric is prescribed using (14), for any $\gamma \in (0,1)$, problem (A6) is equivalent to the following conic program

$$
\begin{array}{ll}
\inf & \dfrac{1}{N}\displaystyle\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,\ \theta \in \mathbb{R}_+,\ \mu \in \mathbb{R}^{2\times 2},\ \nu' \in \mathbb{R}^N,\ \theta' \in \mathbb{R}_+,\ \mu' \in \mathbb{R}^{2\times 2}, \\
& \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 - yw^\top \hat{x}_i + (\rho - \kappa_{\mathcal{A}}|a-\hat{a}_i| - \kappa_{\mathcal{Y}}|y-\hat{y}_i|)\|w\|_* \\
\qquad\qquad \le \mu_{ay} - \theta \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(a,y) + \nu_i
\end{array}\right\} \quad \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{0} \le \hat{p}_{11}(\mu'_{1,1} - \theta' \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(1,1) + \nu'_i), \\
\quad 1 + w^\top \hat{x}_i + (\rho - \kappa_{\mathcal{A}}|1-\hat{a}_i| - \kappa_{\mathcal{Y}}|1-\hat{y}_i|)\|w\|_* + b \\
\qquad\qquad \le \hat{p}_{11}(\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 \hat{0} \le \hat{p}_{01}(\mu'_{01} - \theta' \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,1) + \nu'_i), \\
\quad 1 - w^\top \hat{x}_i + (\rho - \kappa_{\mathcal{A}}|0-\hat{a}_i| - \kappa_{\mathcal{Y}}|1-\hat{y}_i|)\|w\|_* - b \\
\qquad\qquad \le \hat{p}_{01}(\mu'_{01} - \theta' \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,1) + \nu'_i) \\
\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^a \mathbb{1}_{(\hat{a}_i,\hat{y}_i)}(0,-1) + \nu'_i
\end{array}\right\} \quad \forall i \in [N], \\
& \dfrac{1}{N}\displaystyle\sum_{i\in[N]} \nu'_i + \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} \hat{p}_{ay}\mu'_{ay} - \theta'(1-\gamma) - 1 \le \zeta.
\end{array}
$$

*Proof of Theorem A.6.* The proof parallels that of Theorem 5.1; thus, we omit it for brevity. □

## Appendix B. Marginal Constraints and Finite-sample Guarantees

In this section, we illustrate how to handle ambiguity in the marginal distributions and obtain a generalized model with finite sample guarantees. To relax the marginal constraints in the ambiguity set (4), we first construct four ambiguity sets around the empirical conditional distributions as

$$
\mathbb{B}_{ay}(\hat{\mathbb{P}}_{ay}) = \left\{ \mathbb{Q}_{ay} \in \mathcal{M}(\mathcal{X}) :\ \mathbb{W}_\infty(\mathbb{Q}_{ay}, \hat{\mathbb{P}}_{ay}) \le \rho_{ay} \right\}, \tag{A7}
$$

where $\hat{\mathbb{P}}_{ay} \triangleq \frac{1}{|\mathcal{I}_{ay}|}\sum_{i\in\mathcal{I}_{ay}} \delta_{\hat{x}_i}$ is the empirical conditional distribution. Notice that the notation $\mathbb{W}_\infty$ in the above definition of the ambiguity set is used with a slight abuse of notation: $\mathbb{W}_\infty$ in this case is a distance on $\mathcal{M}(\mathcal{X})$, and it is no longer a distance on the *joint* space $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$ as is used in the main paper. Next, we construct an ambiguity set for the marginal distribution $p \in \mathbb{R}^4$ based on the $\chi^2$-divergence by

$$
\Delta = \left\{ p \in \mathbb{R}^4_{++} :\ 1^\top p = 1,\ \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} (p_{ay} - \hat{p}_{ay})^2 / p_{ay} \le \delta_p \right\} \tag{A8}
$$

Combining the two ambiguity sets, we define the following generalized ambiguity set:

$$
\mathbb{B}_g(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X}\times\mathcal{A}\times\mathcal{Y}) :\ \begin{array}{c} \exists \mathbb{Q}_{ay} \in \mathcal{M}(\mathbb{X}) \quad \forall (a,y) \in \mathcal{A}\times\mathcal{Y},\ p \in \mathbb{R}^4 \text{ such that :} \\ \mathbb{Q}(\mathbb{X}\times\{a\}\times\{y\}) = p_{ay}\mathbb{Q}_{ay}(\mathbb{X}) \quad \text{for all } \mathbb{X} \text{ measurable, } \forall (a,y) \in \mathcal{A}\times\mathcal{Y}, \\ p \in \Delta, \mathbb{Q}_{ay} \in \mathbb{B}_{ay}(\hat{\mathbb{P}}_{ay}) \quad \forall (a,y) \in \mathcal{A}\times\mathcal{Y} \end{array} \right\}.
$$

It can be verified that when $\Delta$ contains only the empirical marginal distribution $\hat{p}$, the generalized ambiguity set reduces to the ambiguity set $\mathbb{B}(\hat{\mathbb{P}})$ defined in (4). As the conditional probability measures are supported on $\mathcal{M}(\mathbb{X})$, with a slight abuse of notation, we use the following ground metric:

$$c\big(x', x\big) = \|x - x'\|. \tag{A9}$$

Now, consider the $\varepsilon$-DRFC problem with the generalized ambiguity set

$$\begin{array}{lll} \min & \sup\limits_{\mathbb{Q}\in\mathbb{B}_g(\hat{\mathbb{P}})} & \mathbb{Q}(Y(w^\top X + b) < \varepsilon) \\ \text{s.t.} & w \in \mathbb{R}^d,\ b \in \mathbb{R},\ \|(w,b)\| \le 1 & \\ & \sup\limits_{\mathbb{Q}\in\mathbb{B}_g(\hat{\mathbb{P}})} & \mathbb{U}_\varepsilon(w, b, \mathbb{Q}) \le \eta. \end{array} \tag{A10}$$

Notice that the above optimization is similar to problem (8): the only two differences are the ambiguity set $\mathbb{B}_g(\hat{\mathbb{P}})$ and the unfairness measure $\mathbb{U}_\varepsilon$. The next theorem asserts that the generalized model is equivalent to a mixed binary second-order cone program.

**Theorem B.1** (Generalized $\varepsilon$-DRFC reformulation)**.** Suppose the ground metric is prescribed using (A9), then the generalized $\varepsilon$-DRFC problem (A10) is equivalent to the mixed binary second-order cone program

$$\begin{array}{lll} \min & \delta_p\zeta - \theta - 2\hat{p}^\top r + 2\zeta 1^\top \hat{p} & \\ \text{s.t.} & w \in \mathbb{R}^d,\ b \in \mathbb{R},\ \zeta \in \mathbb{R}_+,\ \theta \in \mathbb{R},\ r \in \mathbb{R}^4,\ s \in \mathbb{R}^4,\ t \in \{0,1\}^N,\ \lambda^0 \in \{0,1\}^N, & \\ & \|(w,b)\| \le 1, & \\ & s_{ay} + \theta \le \zeta,\ \sqrt{4r_{ay}^2 + (s_{ay}+\theta)^2} \le 2\zeta - s_{ay} - \theta & \forall (a,y) \in \mathcal{A}\times\mathcal{Y}, \\ & \dfrac{1}{|\mathcal{I}_{ay}|}\displaystyle\sum_{i\in\mathcal{I}_{ay}} t_i \le s_{ay} & \forall (a,y) \in \mathcal{A}\times\mathcal{Y}, \\ & -\hat{y}_i(w^\top \hat{x}_i + b) + \rho_{\hat{a}_i\hat{y}_i}\|w\|_* \le Mt_i - \varepsilon & \forall i \in [N], \\ & \dfrac{1}{|\mathcal{I}_{11}|}\displaystyle\sum_{i\in\mathcal{I}_{11}} \lambda_i + \dfrac{1}{|\mathcal{I}_{01}|}\sum_{i\in\mathcal{I}_{01}} \lambda_i - 1 \le \eta, & \\ & w^\top \hat{x}_i + \rho_{11}\|w\|_* + b + \varepsilon \le M\lambda_i \quad \forall i \in \mathcal{I}_{11}, & \\ & -w^\top \hat{x}_i + \rho_{01}\|w\|_* - b \le M\lambda_i \quad \forall i \in \mathcal{I}_{01}, & \end{array} \tag{A11}$$

where $M$ is the big-M parameter.

For the remainder of this section, we will provide the proof for Theorem B.1. This proof relies on the following Lemma.

**Lemma B.2** ($\chi^2$-divergence reformulation, Theorem 4.1 in [2])**.** Define $\Delta$ as in (A8). For any $m \in \mathbb{N}_+$ and $\varphi \in \mathbb{R}^m$, both the optimal value and a maximizer of the worst-case expectation problem $\sup_{p\in\Delta} \varphi^\top p$ can be obtained by solving the second-order cone program

$$\begin{array}{ll} \sup & \varphi^\top p \\ \text{s.t.} & p \in \mathbb{R}^m_+,\ q \in \mathbb{R}^m_+,\ 1^\top p = 1,\ 1^\top q \le \delta_p, \\ & \sqrt{(p_j - \hat{p}_j)^2 + \frac{1}{4}p_j^2 + q_j^2} \le \frac{1}{2}p_j + q_j \qquad \forall j \in [m]. \end{array}$$

The optimal value can also be computed by solving the dual problem:

$$\begin{array}{ll}\inf & \delta_p\zeta-\theta-2\hat{p}^\top r+2\zeta 1^\top\hat{p}\\ \text{s.t.} & \zeta\in\mathbb{R}_+,\ \theta\in\mathbb{R},\ r\in\mathbb{R}^m,\ s\in\mathbb{R}^m,\\ & \varphi_j\le s_j,\ s_j+\theta\le\zeta,\ \sqrt{4r_j^2+(s_j+\theta)^2}\le 2\zeta-s_j-\theta \qquad \forall j\in[m].\end{array}$$

We are now ready to present the proof.

*Proof of Theorem B.1.* Observe that problem (A10) can be equivalently written as

$$\begin{array}{ll}\min & \sup\limits_{p\in\Delta}\sum\limits_{a\in\mathcal{A},y\in\mathcal{Y}}\sup\limits_{\mathbb{Q}_{ay}\in\mathbb{B}_{ay}(\hat{\mathbb{P}}_{ay})} p_{ay}\mathbb{Q}_{ay}(Y(w^\top X+b)<\varepsilon)\\ \text{s.t.} & w\in\mathbb{R}^d,\ b\in\mathbb{R},\ \|(w,b)\|\le 1\\ & \sup\limits_{\mathbb{Q}\in\mathbb{B}_g(\hat{\mathbb{P}})}\mathbb{U}_\varepsilon(w,b,\mathbb{Q})\le\eta.\end{array}$$

We first derive the reformulation of the objective function. By the definition of $\Delta$ and the result of Lemma B.2, we have

$$\begin{aligned}&\sup_{p\in\Delta}\sum_{a\in\mathcal{A},y\in\mathcal{Y}}\sup_{\mathbb{Q}\in\mathbb{B}_{ay}(\hat{\mathbb{P}}_{ay})} p_{ay}\mathbb{Q}_{ay}(Y(w^\top X+b)<\varepsilon)\\ =&\min\left\{\delta_p\zeta-\theta-2\hat{p}^\top r+2\zeta 1^\top\hat{p}:\ \begin{array}{ll}\zeta\in\mathbb{R}_+,\ \theta\in\mathbb{R},\ r\in\mathbb{R}^4,\ s\in\mathbb{R}^4,\\ \sup\limits_{\mathbb{Q}_{ay}\in\mathbb{B}_{ay}(\hat{\mathbb{P}}_{ay})}\mathbb{Q}_{ay}(Y(w^\top X+b)<\varepsilon)\le s_{ay} & \forall(a,y)\in\mathcal{A}\times\mathcal{Y},\\ s_{ay}+\theta\le\zeta,\ \sqrt{4r_{ay}^2+(s_{ay}+\theta)^2}\le 2\zeta-s_{ay}-\theta & \forall(a,y)\in\mathcal{A}\times\mathcal{Y}\end{array}\right\}\\ =&\min\left\{\delta_p\zeta-\theta-2\hat{p}^\top r+2\zeta 1^\top\hat{p}:\ \begin{array}{ll}\zeta\in\mathbb{R}_+,\ \theta\in\mathbb{R},\ r\in\mathbb{R}^4,\ s\in\mathbb{R}^4,\ t\in\{0,1\}^N,\\ s_{ay}+\theta\le\zeta,\ \sqrt{4r_{ay}^2+(s_{ay}+\theta)^2}\le 2\zeta-s_{ay}-\theta & \forall(a,y)\in\mathcal{A}\times\mathcal{Y},\\ \dfrac{1}{|\mathcal{I}_{ay}|}\sum\limits_{i\in\mathcal{I}_{ay}}t_i\le s_{ay} & \forall(a,y)\in\mathcal{A}\times\mathcal{Y},\\ -\hat{y}_i(w^\top\hat{x}_i+b)+\rho_{\hat{a}_i\hat{y}_i}\|w\|_*\le Mt_i-\varepsilon & \forall i\in[N]\end{array}\right\}.\end{aligned}$$

For the constraint, we have

$$\begin{aligned}&\sup_{\mathbb{Q}\in\mathbb{B}_g(\hat{\mathbb{P}})}\mathbb{Q}_{11}(w^\top X+b>-\varepsilon)-\mathbb{Q}_{01}(w^\top X+b\ge 0)\\ =&\sup_{\mathbb{Q}_{11}\in\mathbb{B}_{11}(\hat{\mathbb{P}}_{11}),\mathbb{Q}_{01}\in\mathbb{B}_{01}(\hat{\mathbb{P}}_{01})}\mathbb{Q}_{11}(w^\top X+b>-\varepsilon)-\mathbb{Q}_{01}(w^\top X+b\ge 0)\\ =&\sup_{\mathbb{Q}_{11}\in\mathbb{B}_{11}(\hat{\mathbb{P}}_{11})}\mathbb{Q}_{11}(w^\top X+b>-\varepsilon)-\inf_{\mathbb{Q}_{01}\in\mathbb{B}_{01}(\hat{\mathbb{P}}_{01})}\mathbb{Q}_{01}(w^\top X+b\ge 0)\\ =&\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\sup_{x_i:\|x_i-\hat{x}_i\|\le\rho_{11}}\mathbb{I}(w^\top x_i+b>-\varepsilon)-\frac{1}{|\mathcal{I}_{01}|}\left(|\mathcal{I}_{01}|-\sum_{i\in\mathcal{I}_{01}}\sup_{x_i:\|x_i-\hat{x}_i\|\le\rho_{01}}\mathbb{I}(w^\top x_i+b<0)\right)\\ =&\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\sup_{x_i:\|x_i-\hat{x}_i\|\le\rho_{11}}\mathbb{I}(w^\top x_i+b>-\varepsilon)+\frac{1}{|\mathcal{I}_{01}|}\sum_{i\in\mathcal{I}_{01}}\sup_{x_i:\|x_i-\hat{x}_i\|\le\rho_{01}}\mathbb{I}(w^\top x_i+b<0)-1\\ =&\left\{\begin{array}{ll}\min & \dfrac{1}{|\mathcal{I}_{11}|}\sum\limits_{i\in\mathcal{I}_{11}}\lambda_i+\dfrac{1}{|\mathcal{I}_{01}|}\sum\limits_{i\in\mathcal{I}_{01}}\lambda_i-1\\ \text{s.t.} & \lambda\in\{0,1\}^N\\ & w^\top\hat{x}_i+\rho_{11}\|w\|_*+b+\varepsilon\le M\lambda_i \quad \forall i\in\mathcal{I}_{11}\\ & -w^\top\hat{x}_i+\rho_{01}\|w\|_*-b\le M\lambda_i \quad \forall i\in\mathcal{I}_{01},\end{array}\right.\end{aligned}$$

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

We now investigate the finite-sample guarantee of this generalized problem.

**Theorem B.3** (Finite-sample guarantee)**.** Let $\mathbb{P}^\star$ denote the true joint distribution of $(X, A, Y)$. Assume that for all $(a, y) \in \mathcal{A} \times \mathcal{Y}$, the conditional distribution $\mathbb{P}^\star_{ay}$ of $X \in \mathbb{R}^d$, with $d \geq 2$, has a density function $\tau_{ay} : \bar{\mathbb{X}} \to [0, \infty)$, where $\bar{\mathbb{X}} \subseteq \mathbb{X} \subseteq \mathbb{R}^d$ is an open, connected, and bounded set with a Lipschitz boundary, and there exists a constant $\lambda \geq 1$ such that $1/\lambda \leq \tau_{ay}(x) \leq \lambda$ for all $x \in \bar{\mathbb{X}}$ and $(a, y) \in \mathcal{A} \times \mathcal{Y}$. Let $v^\star$ be the optimal value of (A10), and $(w^\star, b^\star) \in \mathbb{R}^{d+1}$ be the corresponding optimal solution. Then for any $\alpha > 2$, setting $\rho_{ay} = C_1 \frac{\log(|\mathcal{I}_{ay}|)^{1/d}}{|\mathcal{I}_{ay}|^{1/d}}$ and $\delta_p > \frac{k}{N}$ implies

$$\text{Prob}\left(\mathbb{P}^\star(Y(w^\top X + b) \leq 0) \leq v^\star\right) \geq 1 - C_2 e^{-\frac{1}{2}\left(N\delta_p - \sqrt{2kN\delta_p - k^2}\right)} - C_3 \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} |\mathcal{I}_{ay}|^{-\frac{\alpha}{2}},$$

and

$$\text{Prob}\left(\mathbb{U}(w, b, \mathbb{P}^\star) \leq \eta\right) \geq 1 - C_2 e^{-\frac{1}{2}\left(N\delta_p - \sqrt{2kN\delta_p - k^2}\right)} - C_3 \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} |\mathcal{I}_{ay}|^{-\frac{\alpha}{2}},$$

where $C_1$ is a constant which depends on the true distribution $\mathbb{P}^\star$ and $\alpha$, $C_2$ is a constant which depends on the conditional distribution $\mathbb{P}^\star$, and $C_3$ is an universal constant.

The proof of Theorem B.3 relies on the following theorem, which provides the concentration inequality for the type $\infty$-Wasserstein distance.

**Theorem B.4** ($\infty$-Wasserstein concentration, Theorem 1.1 in [11])**.** Assume that the probability distribution of $\xi \in \mathbb{R}^d$ has a density function $\tau : \bar{\Xi} \to [0, \infty)$, where $\bar{\Xi} \subseteq \Xi \subseteq \mathbb{R}^d$ is an open, connected, and bounded set with a Lipschitz boundary, and there exists a constant $\lambda \geq 1$ such that $1/\lambda \leq \tau(\xi) \leq \lambda$ for all $\xi \in \bar{\Xi}$. Then, for any fixed $\alpha > 2$,

$$\text{Prob}\left(\mathbb{W}_\infty(\mathbb{P}^\star, \hat{\mathbb{P}}) > C \left\{ \begin{array}{ll} \frac{\log(N)^{3/4}}{N^{1/2}}, & \text{if } d = 1 \\ \frac{\log(N)^{1/d}}{N^{1/d}}, & \text{if } d \geq 2 \end{array} \right\}\right) = \mathcal{O}(N^{-\frac{\alpha}{2}}),$$

where $C$ is a constant which depends only on $\alpha$, $\bar{\Xi}$, and $\lambda$.

Equipped with Theorem B.4, we are now ready to show the proof of Theorem B.3.

*Proof of Theorem B.3.* For any $(a, y) \in \mathcal{A} \times \mathcal{Y}$, $\hat{p}_{ay}$ is an estimator given by

$$\hat{p}_{ay} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(a,y)}(\hat{a}_i, \hat{y}_i).$$

Let $p_{ay}$ denote the mass of the true marginal distribution on $(a, y)$. It can be verified that $N \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} (p_{ay} - \hat{p}_{ay})^2 / p_{ay}$ asymptotically converges to $\chi^2$ with degree $k = 3$ [3]. To obtain an explicit and concise result, we employ the $\chi^2$ test-statistic upper tail bound [6, Lemma 1], which can be written as

$$\text{Prob}\left(Z \geq k + 2\sqrt{kt} + 2t\right) \leq e^{-t},$$

where $Z \sim \chi^2_k$ and $t \geq 0$. By letting $N\delta_p = k + \sqrt{kt} + 2t$ and solving for $t$, we further obtain

$$\text{Prob}\left(N \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} (p_{ay} - \hat{p}_{ay})^2 / p_{ay} > N\delta_p\right) \leq C_2 e^{-\frac{1}{2}\left(N\delta_p - \sqrt{2kN\delta_p - k^2}\right)},$$

where $C_2$ is a constant that depends on the underlying distribution. We assume the conditions of Theorem B.4 hold and $d \geq 2$. Then there exists $C_3 > 0$ such that setting $\rho_{ay} = C_1 \frac{\log(|\mathcal{I}_{ay}|)^{1/d}}{|\mathcal{I}_{ay}|^{1/d}}$ implies

$$\text{Prob}\left(\mathbb{W}_\infty(\mathbb{P}^\star_{ay}, \hat{\mathbb{P}}_{ay}) > \rho\right) \leq C_3 |\mathcal{I}_{ay}|^{-\frac{\alpha}{2}} \qquad \forall (a, y) \in \mathcal{A} \times \mathcal{Y}.$$

By union bound, we further obtain

$$\text{Prob}\left(\mathbb{P}^\star_{ay} \in \mathbb{B}_{ay}(\hat{\mathbb{P}}_{ay})\ \forall (a,y) \in \mathcal{A}\times\mathcal{Y} \text{ and } p \in \Delta\right) \geq 1 - C_2 e^{-\frac{1}{2}\left(N\delta_p - \sqrt{2kN\delta_p - k^2}\right)} - \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} C_3 |\mathcal{I}_{ay}|^{-\frac{\alpha}{2}}.$$

This result immediately leads to the statements in the theorem. □

Theorem B.3 requires choosing the radius $\rho$ based on the constants $C_1$ and $C_2$, and the probabilistic guarantee also depends on the universal constant $C_3$. However, these parameters depend on the properties of the underlying distribution $\mathbb{P}^\star$, which are typically unknown to decision makers. Moreover, the generalized model introduces one more tuning parameter $\delta_p$, making the cross-validation procedure more demanding. Therefore, in practice, we adopt the simpler ambiguity set (4). Nonetheless, Theorem 1 describes an explicit rate for decreasing the Wasserstein radius $\rho$, which provides useful insight to determine the parameter value in practice.

## Appendix C. Hinge Loss and CVaR

In this section, we show that the expected hinge loss minimization problem is exactly the Conditional Value at Risk (CVaR) approximation of the misclassification probability minimization problem (1). For any distribution $\mathbb{Q} \in \mathcal{M}$, the hinge loss minimization problem is defined as

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \ \mathbb{E}_{\mathbb{Q}}\left[\max\left\{0, 1 - Y(w^\top X + b)\right\}\right]. \tag{A12}$$

Meanwhile, we observe that the misclassification probability minimization problem under distribution $\mathbb{Q}$ can be equivalently written as

$$\begin{array}{ll} \inf & t \\ \text{s.t.} & w \in \mathbb{R}^d,\ b \in \mathbb{R},\ t \in \mathbb{R}, \\ & \mathbb{Q}\left(Y(w^\top X + b) > 0\right) \geq 1 - t. \end{array} \tag{A13}$$

The optimization problem above can be regarded as an extension of chance constrained programs, with the quantile $t$ being a decision variable. However, the feasible set of such a chance constraint is non-convex, which makes optimization problematic in the face of large instances. A natural way to overcome this difficulty is to replace the chance constraint with a tractable approximation. To obtain an efficient model for large sample sizes, we require the approximation to be convex. Moreover, we also like the approximation to be conservative, i.e., if a solution is feasible in the approximated problem then it is also feasible to the original problem. If these two conditions hold, we refer to the approximation as a convex conservative approximation.

To this end, the well-known CVaR can be used to derive such an approximation. The core idea lies in the fact that the chance constraint in (A13) can be written as a Value at Risk (VaR) constraint. This yields the equivalent reformulation

$$\begin{array}{ll} \inf & t \\ \text{s.t.} & w \in \mathbb{R}^d,\ b \in \mathbb{R},\ t \in \mathbb{R}, \\ & \text{VaR}_{(t,\mathbb{Q})}(-Y(w^\top X + b)) < 0, \end{array}$$

where $\text{VaR}_{(t,\mathbb{Q})}(Z)) \triangleq \inf\{\tau \in \mathbb{R} : \mathbb{Q}(Z \leq \tau) \geq 1 - t\}$ can be interpreted as the $1-t$ quantile of $Z$. Recall that the CVaR of a random variable $Z$ is defined as

$$\text{CVaR}_{(t,\mathbb{Q})}(Z) \triangleq \inf\left\{\tau \in \mathbb{R} : \tau + \frac{1}{t}\mathbb{E}_{\hat{\mathbb{P}}}\left[\max\{0, Z - \tau\}\right]\right\}.$$

It can be verified that $\mathrm{VaR}_{(t,\mathbb{Q})}(Z)$ is a minimizer of the right hand side problem [7]. Thus, the relation $\mathrm{VaR}_{(t,\mathbb{Q})}(Z)) \leq \mathrm{CVaR}_{(t,\mathbb{Q})}(Z)$ holds for any distribution, and we can replace the VaR term using CVaR, and obtain the following convex conservative approximation to the chance constrained program (A13):

$$\begin{array}{ll} \inf & t \\ \text{s.t.} & w \in \mathbb{R}^d,\ b \in \mathbb{R},\ t \in \mathbb{R}, \\ & \mathrm{CVaR}_{(t,\mathbb{Q})}(-Y(w^\top X + b)) < 0. \end{array} \tag{A14}$$

Interestingly, we find that the CVaR approximation problem (A14) is exactly equivalent to the hinge loss minimization model (A12). That is, the optimal value of (A14) coincides with the optimal value of (A12), and the two problems yield the same optimal classifiers. We remark that this result also connects the well-known SVM model, which minimizes the empirical expected hinge loss, with the misclassification rate minimization problem from the perspective of conservative approximation.

**Theorem C.1** (CVaR equivalence)**.** The CVaR approximation (A14) is equivalent to the SVM model (A12).

*Proof of Theorem C.1.* Let $t^\star$ be the optimal value of (A14). When $t^\star = 0$, one can verify that this condition implies the dataset is linearly separable, and the optimal value of the SVM model (A12) will also be zero. Now, without loss of generality, we assume $t^\star > 0$. By the definition of CVaR, we have

$$\begin{aligned} t^\star &= \inf \left\{ t : \begin{array}{l} w \in \mathbb{R}^d,\ b \in \mathbb{R},\ t \in \mathbb{R}_{++}, \\ \inf_{\beta \in \mathbb{R}} -\beta + \frac{1}{t}\mathbb{E}_{\mathbb{Q}}\left[\max\{\beta - Y(w^\top X + b), 0\}\right] < 0 \end{array} \right\} \\ &= \inf \left\{ t : \begin{array}{l} w \in \mathbb{R}^d,\ b \in \mathbb{R},\ \beta \in \mathbb{R},\ t \in \mathbb{R}_{++}, \\ -\beta + \frac{1}{t}\mathbb{E}_{\mathbb{Q}}\left[\max\{\beta - Y(w^\top X + b), 0\}\right] < 0. \end{array} \right\}. \end{aligned}$$

Notice that when $\beta \leq 0$, the constraint is always infeasible; hence, we can restrict $\beta$ to be positive without changing the feasible region, which yields

$$\begin{aligned} t^\star &= \inf \left\{ t : \begin{array}{l} w \in \mathbb{R}^d,\ b \in \mathbb{R},\ \beta \in \mathbb{R}_{++},\ t \in \mathbb{R}_{++}, \\ -\beta + \frac{1}{t}\mathbb{E}_{\mathbb{Q}}\left[\max\{\beta - Y(w^\top X + b), 0\}\right] < 0 \end{array} \right\} \\ &= \inf \left\{ t : \begin{array}{l} w \in \mathbb{R}^d,\ b \in \mathbb{R},\ \beta \in \mathbb{R}_{++},\ t \in \mathbb{R}_{++}, \\ \mathbb{E}_{\mathbb{Q}}\left[\max\{\beta - Y(w^\top X + b), 0\}\right] < \beta t \end{array} \right\} \\ &= \inf \left\{ t : \begin{array}{l} w \in \mathbb{R}^d,\ b \in \mathbb{R},\ \beta \in \mathbb{R}_{++},\ t \in \mathbb{R}_{++}, \\ \mathbb{E}_{\mathbb{Q}}\left[\max\left\{1 - Y\left(\left(\frac{w}{\beta}\right)^\top X + \left(\frac{b}{\beta}\right)\right), 0\right\}\right] < t \end{array} \right\} \\ &= \left\{ \begin{array}{ll} \min & \mathbb{E}_{\mathbb{Q}}\left[\max\left\{1 - Y\left(w'^\top X + b'\right), 0\right\}\right] \\ \text{s.t.} & w' \in \mathbb{R}^d,\ b' \in \mathbb{R}, \end{array} \right. \end{aligned}$$

where the penultimate equality holds by setting $w' = w/\beta$ and $\beta' = b/\beta$. Thus, the optimal value of these two problem coincides. Furthermore, by noticing that $w' = w/\beta$ and $b' = b/\beta$, the corresponding optimal hyperplanes $w^\top X + b = 0$ and $w'^\top X + b' = 0$ are also the same. This completes the proof. □

## Appendix D. Details of the Numerical Experiments

In this section, we provide further details about the experiments in Section 6.

**Synthetic Experiments.** In the first part of synthetic experiments, we use 2-d binary classification samples to depict the classification boundary for different models. The setup for this experiment follows from [10]. We generate 5000 samples for the majority group ($A = q$) and 2000 samples for the minority

group $(A = 0)$ with $\mathbb{P}^\star(Y = 1) = \mathbb{P}^\star(Y = -1) = 0.5$ in both of them. The conditional probabilities are set as the following Gaussian distributions

$$X|A = 1, Y = 1 \sim \mathcal{N}([6, 0], [3.5, 0; 0, 3.5])$$
$$X|A = 1, Y = 0 \sim \mathcal{N}([2, 0], [3.5, 0; 0, 3.5])$$
$$X|A = 0, Y = 1 \sim \mathcal{N}([-2, 0], [5, 0; 0, 5])$$
$$X|A = 0, Y = 0 \sim \mathcal{N}([-4, 0], [5, 0; 0, 5])$$

We then employ the stratified sampling method to obtain $N = 25$ samples as the training dataset. The rest of the dataset is merged to be the testing dataset, and the accuracy and unfairness scores in Table 2 are evaluated using this testing set. In the fairness-aware classifiers, we set the unfairness threshold $\eta = 0.2$ for the $\varepsilon$-FC and $\varepsilon$-DRFC model, and $\zeta = 1.35$ for the HFC and HDRFC model. The Wasserstein radii are set to 0.1 and 0.2 for the $\varepsilon$-DRFC and HDRFC models, respectively.

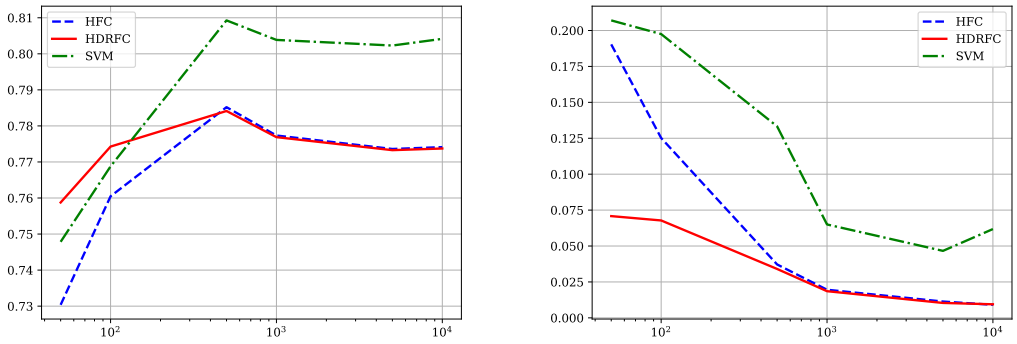
**Statistics on the Training Datasets.** Table A1 reports the statistics of different methods on the training dataset. We first illustrate the effect of the distributionally robust scheme by comparing the DOB+ method with other DRO models. One can observe that the non-robust model performs well within the training dataset, achieving high accuracy while yielding low unfairness scores. However, it cannot preserve these nice in-sample performances in the out-of-sample data because the model does not consider possible noises from limited historical observations. Conversely, the robust models, which deviate from empirical objectives through carefully crafted regularization terms, do not show significant advantages compared with the empirical-based model. Yet, they demonstrate a superior ability to maintain the performance when applied to unseen datasets, as illustrated in Table 4. Moreover, we also notice that the $\varepsilon$-DRFC model yields higher accuracy and lower unfairness scores compared with other fair classifiers. This nice performance is due to the tight conservative approximation of the objective and unfairness measures.

| Dataset | Metric | HC(SVM) | DOB+ | DRFLR | HDRFC | $\varepsilon$-DRFC |
|---|---|---|---|---|---|---|
| German | Accuracy | $\mathbf{0.87 \pm 0.03}$ | $0.84 \pm 0.02$ | $0.82 \pm 0.01$ | $0.82 \pm 0.01$ | $0.85 \pm 0.01$ |
| | $F_1$-score | $\mathbf{0.91 \pm 0.03}$ | $0.83 \pm 0.03$ | $0.83 \pm 0.02$ | $0.82 \pm 0.02$ | $0.89 \pm 0.01$ |
| | Unfairness ($\lvert\mathbb{U}\rvert$) | $0.07 \pm 0.06$ | $0.02 \pm 0.01$ | $0.02 \pm 0.01$ | $\mathbf{0.01 \pm 0.00}$ | $0.02 \pm 0.01$ |
| COMPAS | Accuracy | $\mathbf{0.68 \pm 0.03}$ | $0.64 \pm 0.04$ | $0.61 \pm 0.03$ | $0.62 \pm 0.03$ | $0.67 \pm 0.03$ |
| | $F_1$-score | $\mathbf{0.63 \pm 0.03}$ | $0.56 \pm 0.03$ | $0.50 \pm 0.04$ | $0.52 \pm 0.04$ | $\mathbf{0.63 \pm 0.03}$ |
| | Unfairness ($\lvert\mathbb{U}\rvert$) | $0.22 \pm 0.06$ | $\mathbf{0.05 \pm 0.04}$ | $\mathbf{0.05 \pm 0.04}$ | $\mathbf{0.05 \pm 0.04}$ | $0.06 \pm 0.03$ |
| Arrhythmia | Accuracy | $0.78 \pm 0.04$ | $0.73 \pm 0.03$ | $0.72 \pm 0.03$ | $0.72 \pm 0.04$ | $\mathbf{0.82 \pm 0.03}$ |
| | $F_1$-score | $0.80 \pm 0.04$ | $0.79 \pm 0.03$ | $0.78 \pm 0.03$ | $0.79 \pm 0.03$ | $\mathbf{0.84 \pm 0.03}$ |
| | Unfairness ($\lvert\mathbb{U}\rvert$) | $0.09 \pm 0.07$ | $\mathbf{0.04 \pm 0.03}$ | $\mathbf{0.04 \pm 0.02}$ | $\mathbf{0.04 \pm 0.03}$ | $0.05 \pm 0.03$ |
| Adult | Accuracy | $\mathbf{0.84 \pm 0.04}$ | $\mathbf{0.84 \pm 0.03}$ | $0.82 \pm 0.02$ | $0.77 \pm 0.02$ | $0.82 \pm 0.03$ |
| | $F_1$-score | $0.57 \pm 0.04$ | $0.49 \pm 0.03$ | $0.51 \pm 0.03$ | $0.62 \pm 0.03$ | $\mathbf{0.73 \pm 0.03}$ |
| | Unfairness ($\lvert\mathbb{U}\rvert$) | $0.26 \pm 0.16$ | $0.11 \pm 0.08$ | $0.10 \pm 0.07$ | $0.09 \pm 0.07$ | $\mathbf{0.05 \pm 0.04}$ |
| Drug | Accuracy | $\mathbf{0.89 \pm 0.03}$ | $0.83 \pm 0.02$ | $0.80 \pm 0.02$ | $0.78 \pm 0.02$ | $0.80 \pm 0.03$ |
| | $F_1$-score | $0.69 \pm 0.03$ | $0.54 \pm 0.04$ | $0.50 \pm 0.04$ | $0.60 \pm 0.03$ | $\mathbf{0.75 \pm 0.03}$ |
| | Unfairness ($\lvert\mathbb{U}\rvert$) | $0.16 \pm 0.12$ | $0.09 \pm 0.07$ | $0.08 \pm 0.04$ | $0.09 \pm 0.05$ | $\mathbf{0.06 \pm 0.03}$ |

TABLE A1. Train accuracy, $F_1$-score and unfairness (average $\pm$ standard deviation) for $N = 100$. The best results for each dataset are highlighted in bold.

**Experiment with different sample sizes.** We also conducted experiments with varying sample sizes on the Adult dataset to examine the performance of the aforementioned convex models. The results

averaged from 10 trials are presented in Figure A1. Initially, it can be observed that the plain-vanilla classifier yields highly unfair results when dealing with small sample sizes. This observation underscores the fact that insufficient training data can further exacerbate the fairness score in machine learning problems. In such a scenario, introducing a fairness constraint could be a viable solution to promote fairness. Unfortunately, due to the limited observations and the consequent inaccurate estimation of the underlying distribution, the performance of the empirical fair classifier is unsatisfactory. As shown in Figure A1, the empirical model HFC still yields a high unfairness score for small sample sizes. Distributionally robust optimization is an excellent approach to address this fundamental shortcoming for small datasets. Compared with the empirical fair classifier, the HDRFC model achieves nice performance regarding both accuracy and fairness for small datasets. When the sample size increases, the empirical distribution gradually constitutes a good estimation of the true distribution. In this case, the distributionally robust model converges to the empirical model, and the improvement brought by the distributionally robust model becomes modest.




(A) Accuracy (B) Unfairness Score

FIGURE A1. Accuracy and Unfairness score on the Adult dataset.

## REFERENCES

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

[2] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[3] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Summer School on Machine Learning*, pages 208–240. Springer, 2003.

[4] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2020.

[5] O. Hölder and H. Schwartz. *Über einen mittelwerthssatz*. 1889.

[6] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

[7] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

[8] L. J. Rogers. An extension of a certain theorem in inequalities. *Messenger of Math.*, 17:145–150, 1888.

[9] H. A. Schwarz. Über ein die flächen kleinsten flächeninhalts betreffendes problem der variationsrechnung. In *Gesammelte Mathematische Abhandlungen*, pages 223–269. Springer, 1890.

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

[11] N. G. Trillos and D. Slepčev. On the rate of convergence of empirical measures in $\infty$-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.