A. Additional Theoretical Results

A.1. Implementation of the Exponentiated Gradient Algorithm in § 4.1

We provide the details of the exponentiated gradient algorithm discussed in § 4.1 for finding the predictive disparity minimizing model within the set of good models. Algorithm 3 implements the exponentiated gradient algorithm, except for the best-response functions of the λ-player and the Q_h-player. The best-response function of the λ-player is

$$\text{Best}_\lambda(Q_h) := \begin{cases} 0 \text{ if } \widehat{\text{cost}}(Q_h) - \hat{\epsilon} \leq 0, \\ B_\lambda \text{ otherwise.} \end{cases} \tag{7}$$

The best-response function of the Q_h-player may be constructed through a further reduction to cost-sensitive classification. The Lagrangian for the setting without selective labels can be written as

$$L(h_f, \lambda) = \hat{\mathbb{E}}\left[\mathbb{E}_{Z_\alpha}\left[c_\lambda(\underline{Y}_i^*, A_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]\right] - \lambda\hat{\epsilon}, \tag{8}$$

where

$$c_\lambda(\underline{Y}_i^*, A_i, Z_\alpha) := \frac{\beta_0}{\hat{p}_0} 1\left\{\mathcal{E}_{i,0}\right\} + \frac{\beta_1}{\hat{p}_1} 1\left\{\mathcal{E}_{i,1}\right\} + \lambda c(\underline{Y}_i^*, Z_\alpha)$$

and $\hat{p}_a := \hat{\mathbb{E}}\left[\mathcal{E}_{i,a}\right]$ for $a \in \{0, 1\}$.

The Lagrangian for the setting *with* selective labels can be written as

$$L(h_f, \lambda) = \hat{\mathbb{E}}\left[\mathbb{E}_{Z_\alpha}\left[c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]\right] - \lambda\hat{\epsilon}, \tag{9}$$

where

$$c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha) := \frac{\beta_0}{\hat{p}_0} g(X_i, Y_i)(1 - A_i) + \frac{\beta_1}{\hat{p}_1} g(X_i, Y_i) A_i + \lambda c(\underline{\hat{\mu}}_i, Z_\alpha)$$

and $\hat{p}_a = \hat{\mathbb{E}}[g(X_i, Y_i) 1\left\{A_i = a\right\}]$ for $a \in \{0, 1\}$.

This is solved by calling cost-sensitive classification oracle on an augmented dataset of size $n \times N$ with observations $\{(X_{i,z_\alpha}, C_{i,z_\alpha}\}_{i \in [n], z_\alpha \in \mathcal{Z}_\alpha}$ with $X_{i,z_\alpha} = (X_i, z_\alpha)$ and $C_{i,z_\alpha} = c_\lambda(\underline{Y}_i^*, A_i, z_\alpha)$ for the setting without selective labels and $= c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha)$ for the setting with selective labels. In our empirical implementation, we use the heuristic least-squares reduction described in (Agarwal et al., 2019). The heuristic reduction generally performed well in our empirical work, but its performance will depend on the dataset and the choice of predictive disparity in general.

A.2. Shrinking the Support of the Stochastic Risk Score

As discussed in § 4.1, a key challenge to the practical use of Algorithm 3 is it returns a stochastic prediction function $\hat{Q}_h$ with possibly large support. The number of prediction functions in the support of $\hat{Q}_h$ is equal to the total number of iterations taken by the respective algorithm. As a result, $\hat{Q}_h$ may be complex to describe, time-intensive to evaluate, and memory-intensive to store.

The support of the returned stochastic prediction may be shrunk while maintaining the same guarantees on its performance by solving a simple linear program. To do so, we take the set of prediction functions in the support of $\hat{Q}_h$ and solve the following linear program

$$\min_{p \in \Delta^T} \sum_{t=1}^{T} p_t \widehat{\text{disp}}(h_t) \text{ s.t. } \sum_{t=1}^{T} p_t \widehat{\text{cost}}(h_t) \leq \hat{\epsilon} + 2\nu, \tag{10}$$

where T is the number of iterations of Algorithm 3, Δ^T is the T-dimensional unit simplex and h_t is the t-th prediction function in the support of $\hat{Q}_h$ (i.e., the prediction function constructed at the t-th iteration of Algorithm 3). We then use the randomized prediction function that assigns probability p_t to each prediction function in the support of $\hat{Q}_h$. In practice, we calibrate the constraint in (10) by choosing the smallest $\nu \geq 0$ such that the linear program has a feasible solution, following the practical recommendations in Cotter et al. (2019).

Algorithm 3: Algorithm for finding the predictive disparity minimizing model

Input: Training data $\{(X_i, Y_i, A_i)\}_{i=1}^n$, parameters β_0, β_1, events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$, empirical loss tolerance $\hat{\epsilon}$, bound B_λ, accuracy ν and learning rate η.
Result: ν-approximate saddle point $(\hat{Q}_h, \hat{\lambda})$
Set $\theta_1 = 0 \in \mathbb{R}$;
for $t = 1, 2, \ldots$ **do**
  Set $\lambda_t = B_\lambda \frac{\exp(\theta_t)}{1 + \exp(\theta_t)}$;
  $h_t \leftarrow \text{Best}_h(\lambda_t)$;
  $\hat{Q}_{h,t} \leftarrow \frac{1}{t}\sum_{s=1}^{t} h_s$,
  $\bar{L} \leftarrow L(\hat{Q}_{h,t}, \text{Best}_\lambda(\hat{Q}_{h,t}))$;
  $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{s=1}^{t} \lambda_s$, $\quad \underline{L} \leftarrow L(\text{Best}_h(\hat{\lambda}_t), \hat{\lambda}_t)$;
  $\nu_t \leftarrow$
   $\max\left\{L(\hat{Q}_{h,t}, \hat{\lambda}_t) - \underline{L}, \bar{L} - L(\hat{Q}_{h,t}, \hat{\lambda}_t)\right\}$;
  if $\nu_t \leq \nu$ **then**
    if $\widehat{\text{cost}}(\hat{Q}_{h,t}) \leq \hat{\epsilon} + \frac{|\beta_0| + |\beta_1| + 2\nu}{B_\lambda}$ **then**
      return $(\hat{Q}_{h,t}, \hat{\lambda}_t)$;
    else
      return *null*
    end
  end
  Set $\theta_{t+1} = \theta_t + \eta\left(\widehat{\text{cost}}(h_t) - \hat{\epsilon}\right)$;
end

Lemma 7 of Cotter et al. (2019) shows that the solution to (10) has at most 2 support points and the same performance guarantees as the original solution $\hat{Q}_h$.

A.3. Computing the Absolute Predictive Disparity Minimizing Model

In this section, we extend the reductions approach to compute the prediction function that minimizes the absolute predictive disparity over the set of good models (3). We solve

$$\min_{Q \in \Delta(\mathcal{F})} |\text{disp}(Q)| \text{ s.t. } \text{loss}(Q) \leq \epsilon. \tag{11}$$

Through the same discretization argument, this problem may be reduced to a constrained classification problem over the set of threshold classifiers

$$\min_{Q_h \in \Delta(\mathcal{H})} |\text{disp}(Q_h)| \text{ s.t. } \text{cost}(Q_h) \leq \epsilon - c_0. \tag{12}$$

To further deal with the absolute value operator in the objective function, we introduce a slack variable ξ and define the equivalent problem over both $Q_h \in \Delta(\mathcal{H}), \xi \in \mathbb{R}$

$$\begin{aligned} \min_{\xi, Q_h \in \Delta(\mathcal{H})} \; & \xi \\ \text{s.t. } & \text{disp}(Q_h) - \xi \leq 0, \\ & -\text{disp}(Q_h) - \xi \leq 0, \\ & \text{cost}(Q_h) \leq \epsilon - c_0. \end{aligned} \tag{13}$$

We construct solutions to the empirical analogue of (13).

Solving the empirical analogue of (13) is equivalent to finding the saddle point $\min_{Q_h \in \Delta(\mathcal{H}), \xi \in [0, B_\xi]} \max_{\|\lambda\| \leq B_\lambda} L(\xi, Q_h, \lambda)$ with Lagrangian $L(\xi, Q_h, \lambda) = \xi + \lambda_+ \left(\widehat{\text{disp}}(Q_h) - \xi\right) + \lambda_- \left(-\widehat{\text{disp}}(Q_h) - \xi\right) + \lambda_{\text{cost}} \left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}\right)$, $\lambda = (\lambda_+, \lambda_-, \lambda_{\text{cost}})$ and B_ξ is a bound on the slack variable. Since the absolute predictive disparity is bounded by one, we define $B_\xi = 1$ in practice. We search for the saddle point by treating it as the equilibrium of a two-player zero-sum game in which one player chooses (ξ, Q_h) and the other chooses λ.

Algorithm 4 computes a ν-approximate saddle point of $L(\xi, Q_h, \lambda)$. The best-response of the λ-player sets the Lagrange multiplier associated with the maximally violated constraint equal to B_λ. Otherwise, she sets all Lagrange multipliers to zero if all constraints are satisfied. In order to analyze the best-response of the (ξ, Q_h)-player, rewrite the Lagrangian as

$$\begin{aligned} L(\xi, Q_h, \lambda) = & (1 - \lambda_+ - \lambda_-)\xi \\ & + (\lambda_+ - \lambda_-)\widehat{\text{disp}}(Q_h) + \lambda_{\text{cost}}(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}). \end{aligned} \tag{14}$$

For a fixed value of λ, minimizing $L(\xi, Q_h, \lambda)$ over (ξ, Q_h) jointly is equivalent to separately minimizing the first term involving ξ and the remaining terms involving Q_h. To minimize $(1 - \lambda_+ - \lambda_-)\xi$, the best-response is to set $\xi = B_\xi$ if $1 - \lambda_+ - \lambda_- < 0$, and set $\xi = 0$ otherwise. Minimizing

$$(\lambda_+ - \lambda_-)\widehat{\text{disp}}(Q_h) + \lambda_{\text{cost}}(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}) \tag{15}$$

over Q_h can be achieved through a reduction to cost-sensitive classification since minimizing the previous display is equivalent to minimizing

$$\hat{\mathbb{E}}\left[\mathbb{E}_{Z_\alpha}\left[c_\lambda(\underline{Y}_i^*, A_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]\right], \tag{16}$$

where now $c_\lambda(\underline{Y}_i^*, A_i, Z_\alpha) := (\lambda_+ - \lambda_-)\left(\frac{\beta_0}{\hat{p}_0} 1\{\mathcal{E}_{i,0}\} + \frac{\beta_1}{\hat{p}_1} 1\{\mathcal{E}_{i,1}\}\right) + \lambda_{\text{cost}} c(\underline{Y}_i^*, Z_\alpha)$.

We use an analogous linear program reduction (§ A.2) to shrink the support of the solution returned by Algorithm 4.

Algorithm 4: Algorithm for finding the absolute predictive disparity minimizing model among the set of good models

Input: Training data $\{(X_i, Y_i, A_i)\}_{i=1}^n$,
Parameters β_0, β_1, Events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$, and empirical loss tolerance $\hat{\epsilon}$
Bounds B_λ, B_ξ, accuracy ν and learning rate η
Result: ν-approximate saddle point $(\hat{\xi}, \hat{Q}, \hat{\lambda})$
Set $\theta_1 = 0 \in \mathbb{R}^3$;
for $t = 1, 2, \ldots$ **do**
  Set $\lambda_{t,k} = B_\lambda \frac{\exp(\theta_{t,k})}{1 + \sum_{k'} \exp(\theta_{t,k'})}$ for all $k = \{\text{cost}, +, -\}$;
  $h_t \leftarrow \text{Best}_h(\lambda_t)$, $\xi_t \leftarrow \text{Best}_\xi(\lambda_t)$;
  $\hat{Q}_{h,t} \leftarrow \frac{1}{t}\sum_{s=1}^t h_s$, $\hat{\xi}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \xi_t$;
  $\bar{L} \leftarrow L(\hat{\xi}_t, \hat{Q}_t, \text{Best}_\lambda(\hat{\xi}_t.\hat{Q}_t))$;
  $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \lambda_s$,
  $\underline{L} \leftarrow L(\text{Best}_\xi(\lambda_t), \text{Best}_h(\hat{\lambda}_t), \hat{\lambda}_t)$;
  $\nu_t \leftarrow \max\left\{L(\hat{\xi}_t, \hat{Q}_t, \hat{\lambda}_t) - \underline{L}, \bar{L} - L(\hat{\xi}_t, \hat{Q}_t, \hat{\lambda}_t)\right\}$;
  if $\nu_t \leq \nu$ **then**
    if $\widehat{\text{cost}}(\hat{Q}) \leq \hat{\epsilon} + \frac{B_\xi + 2\nu}{B_\lambda}$ **then**
      return $(\hat{\xi}_t, \hat{Q}_t, \hat{\lambda}_t)$;
    else
      return $null$;
    end
  end
  Set $\theta_{t+1} = \theta_t + \eta \begin{pmatrix} \widehat{\text{disp}}(h_t) - \xi_t \\ -\widehat{\text{disp}}(h_t) - \xi_t \\ \widehat{\text{cost}}(h_t) - \hat{\epsilon} \end{pmatrix}$;
end

A.3.1. Error Analysis

We analyze the suboptimality of the solution returned by Algorithm 4.

Theorem 3. *Suppose Assumption 1 holds for* $C' \geq 2C + 2 + \sqrt{2\ln(8N/\delta)}$ *and* $C'' \geq \sqrt{\frac{-\log(\delta/8)}{2}}$.

Then, Algorithm 4 with $\nu \propto n^{-\phi}, B_\lambda \propto n^{\phi}, N \propto n^{\phi}$ *terminates in at most* $O(n^{4\phi})$ *iterations. It returns* $\hat{Q}_h$*, which when viewed as a distribution over* $\mathcal{F}$*, satisfies with probability at least* $1-\delta$ *either one of the following: 1)* $\hat{Q}_h \neq null$, $\text{loss}(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$ *and* $\left|\text{disp}(\hat{Q}_h)\right| \leq \left|\text{disp}(\tilde{Q})\right| + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi})$ *for any* $\tilde{Q}$ *that is feasible in (11); or 2)* $\hat{Q}_h = null$ *and (11) is infeasible.*

We next provide an oracle result for the absolute disparity minimizing algorithm under selective labels.

Theorem 4 (Selective Labels for Algorithm 4)**.** *Suppose Assumption 2 holds and Algorithm 4 is given as input the modified training data* $\{(X_i, A_i, \mu(X_i)\}_{i=1}^n$.

Under the same conditions as Theorem 3, Algorithm 4 terminates in at most $O(n^{4\phi})$ *iterations. It returns* $\hat{Q}_h$*, which when viewed as a distribution over* $\mathcal{F}$*, satisfies with probability at least* $1-\delta$ *either one of the following: 1)* $\hat{Q}_h \neq null$, $\text{loss}_\mu(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$ *and* $\left|\text{disp}(\hat{Q}_h)\right| \leq \left|\text{disp}(\tilde{Q})\right| + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi})$ *for any* $\tilde{Q}$ *that is feasible in (11); or 2)* $\hat{Q}_h = null$ *and (11) is infeasible.*

We omit the proof of Theorem 4 since the analogous steps are given in proofs of Theorems 2-3 below.

A.4. Bounded Group Loss Disparity

Bounded group loss is a common notion of predictive fairness that examines the variation in average loss across values of the protected or sensitive attribute. It is commonly used to ensure that the prediction function achieves some minimal threshold of predictive performance across all values of the attribute (Agarwal et al., 2019). We define a bounded group loss disparity to be the difference in average loss across values of the attribute, $\text{disp}(f) = \mathbb{E}\left[l(Y_i^*, f(X_i)) \mid A_i = 1\right] - \mathbb{E}\left[l(Y_i^*, f(X_i)) \mid A_i = 0\right]$. This choice of predictive disparity measure is convenient as it allows us to drastically simplify our algorithm by skipping the discretization step entirely and reducing the problem to an instance of weighted loss minimization. Agarwal et al. (2019) apply the same idea in their analysis of fair regression under bounded group loss.

Take, for example, the problem of finding the range of bounded group loss disparities that are possible over the set of good models. Letting $\text{loss}(f \mid A_i = a) := \mathbb{E}\left[l(Y_i^*, f(X_i) \mid A_i = a\right]$ and $\text{loss}(Q \mid A_i = a) := \sum_{f\in\mathcal{F}} Q(f)\,\text{loss}(f \mid A_i = a)$, we solve

$$\min_{Q\in\Delta(\mathcal{F})} \text{loss}(Q \mid A_i = 1) - \text{loss}(f \mid A_i = 0)$$
$$\text{s.t. } \text{loss}(Q) \leq \epsilon.$$

The sample version of this problem is to minimize $\widehat{\text{loss}}(Q \mid A_i = 1) - \widehat{\text{loss}}(f \mid A_i = 0)$ subject to $\widehat{\text{loss}}(Q) \leq \epsilon$. We solve the sample problem by finding a saddle point of the associated Lagrangian $L(Q, \lambda) = \widehat{\text{loss}}(Q \mid A_i = 1) - \widehat{\text{loss}}(f \mid A_i = 0) + \lambda(\widehat{\text{loss}}(Q) - \epsilon)$. We compute a ν-approximate saddle point by treating it as a zero-sum game between a Q-player and a λ-player. The best response of the λ-player is the same as before: if the constraint $\hat{\text{loss}}(Q) - \epsilon$ is violated, she sets $\lambda = B_\lambda$, and otherwise she sets $\lambda = 0$. The best-response of the Q-player may reduced to an instance of weighted loss minimization since

$$\widehat{\text{loss}}(f|\mathcal{E}_{i,0}) - \widehat{\text{loss}}(f|\mathcal{E}_{i,1}) + \lambda(\widehat{\text{loss}}(f) - \epsilon)$$
$$= \hat{\mathbb{E}}\left[\left(\frac{1}{\hat{p}_0}1\{\mathcal{E}_0\} - \frac{1}{\hat{p}_1}1\{\mathcal{E}_1\} + \lambda\right) l(Y_i, f(X_i))\right]$$

Therefore, defining the weights $W_i = \frac{1}{\hat{p}_0}1\{\mathcal{E}_{i,0}\} - \frac{1}{\hat{p}_1}1\{\mathcal{E}_{i,1}\} + \lambda$, we see that minimizing $L(h, \lambda)$ is equivalent to solving an instance of weighted loss minimization. Algorithm 5 formally states the procedure for finding the range of bounded group loss disparities. We may analogously extend Algorithm 4 to find the absolute bounded group loss-minimizing model among the set of good models.

B. Proofs of Main Results

Proof of Lemma 1

Fix $f \in \mathcal{F}$. For $x \in \mathcal{X}$ and $z_\alpha \in \mathcal{Z}_\alpha$

$$h_f(x, z_\alpha) = 1\{f(x) \geq z_\alpha\} = 1\{\underline{f}(x) \geq z_\alpha\},$$

Therefore,

$$\mathbb{E}_{Z_\alpha}\left[h_f(x, Z_\alpha)\right] = \mathbb{E}_{Z_\alpha}\left[1\{\underline{f}(x) \geq Z_\alpha\}\right] = \underline{f}(x),$$

and for any $a \in \{0, 1\}$,

$$\begin{aligned}
&\left|\mathbb{E}\left[h_f(X, Z_\alpha)|\mathcal{E}_{i,a}\right] - \mathbb{E}\left[f(X)|\mathcal{E}_{i,a}\right]\right| \\
&= \left|\mathbb{E}\left[\mathbb{E}_{Z_\alpha}\left[h_f(X, Z_\alpha)\right] - f(X)|\mathcal{E}_{i,a}\right]\right| \\
&= \left|\mathbb{E}\left[\underline{f}(X) - f(X)|\mathcal{E}_{i,a}\right]\right| \leq \alpha
\end{aligned}$$

where the first equality uses iterated expectations plus the fact that Z_α is independent of (X, A, Y^*) and the final equality follows by the definition of $\underline{f}(X)$. The claim is immediate after noticing $\text{disp}(h_f) - \text{disp}(f)$ equals $\beta_0\left(\mathbb{E}\left[h_f(X, Z_\alpha) - f(X)|\mathcal{E}_{i,0}\right]\right) + \beta_1\left(\mathbb{E}\left[h_f(X, Z_\alpha) - f(X)|\mathcal{E}_{i,1}\right]\right)$ and applying the triangle inequality. □

Algorithm 5: Algorithm for finding the bounded group loss disparity minimizing model over the set of good models

Input: Training data $\{(X_i, Y_i, A_i)\}_{i=1}^n$, Parameters β_0, β_1, Events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$, and loss tolerance $\hat{\epsilon}$
Bound B_λ, accuracy ν and learning rate η
Result: ν-approximate saddle point $(\hat{Q}_h, \hat{\lambda})$
Set $\theta_1 = 0 \in \mathbb{R}$;
for $t = 1, 2, \ldots$ **do**
  Set $\lambda_t = B_\lambda \frac{\exp(\theta_t)}{1+\exp(\theta_t)}$;
  $f_t \leftarrow \text{Best}_f(\lambda_t)$;
  $\hat{Q}_t \leftarrow \frac{1}{t}\sum_{s=1}^t f_s, \quad \bar{L} \leftarrow L(\hat{Q}_t, \text{Best}_\lambda(\hat{Q}_t))$;
  $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \lambda_s, \quad \underline{L} \leftarrow L(\text{Best}_f(\hat{\lambda}_t), \hat{\lambda}_t)$;
  $\nu_t \leftarrow \max\left\{L(\hat{Q}_t, \hat{\lambda}_t) - \underline{L}, \bar{L} - L(\hat{Q}_t, \hat{\lambda}_t)\right\}$;
  if $\nu_t \leq \nu$ **then**
    if $\widehat{\text{loss}}(\hat{Q}_t) \leq \hat{\epsilon} + \frac{|\beta_0|+|\beta_1|+2\nu}{B_\lambda}$ **then**
      return $(\hat{Q}_t, \hat{\lambda}_t)$;
    else
      return *null*
    end
  end
  Set $\theta_{t+1} = \theta_t + \eta\left(\widehat{\text{loss}}(f_t) - \hat{\epsilon}\right)$;
end

Proof of Theorem 1

The claim about the iteration complexity of Algorithm 3 follows immediately from Lemma 2, substituting in the stated choices of ν and B.

The proof strategy for the remaining claims follows the proof of Theorems 2-3 in (Agarwal et al., 2019). We consider two cases.

Case 1: There is a feasible solution Q^* to the population problem (4) Using Lemmas 4-5, the ν-approximate saddle point $\hat{Q}_h$ satisfies

$$\widehat{\text{disp}}(\hat{Q}_h) \leq \widehat{\text{disp}}(Q_h) + 2\nu \tag{17}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} + \frac{|\beta_0| + |\beta_1| + 2\nu}{B} \tag{18}$$

for any distribution Q_h that is feasible in the empirical problem. This implies that Algorithm 3 returns $\hat{Q} \neq null$. We now show that the returned $\hat{Q}_h$ provides an approximate solution to the discretized population problem.

First, define $\widehat{\text{cost}}_z(h) := \hat{\mathbb{E}}\left[c(\underline{Y}_i^*, z)h(X_i, z)\right]$ and $\text{cost}_z(h) := \mathbb{E}\left[c(\underline{Y}_i^*, z)h(X_i, z)\right]$. Since $c(\underline{Y}_i^*, z) \in [-1, 1]$, we invoke Lemma 7 with $S_i = c(\underline{Y}_i^*, z_i)$, $U_i = (X_i, z)$, $\mathcal{G} = \mathcal{H}$ and $\psi(s,t) = st$ to obtain that with probability at least $1 - \frac{\delta}{4}$ for all $z \in \mathcal{Z}_\alpha$ and $h \in \mathcal{H}$

$$\left|\widehat{\text{cost}}_z(h) - \text{cost}_z(h)\right| \leq$$

$$2R_n(\mathcal{H}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(8N/\delta)}{n}} = \tilde{O}(n^{-\phi}),$$

where the last equality follows by the bound on $R_n(\mathcal{H})$ in Assumption 1 and setting $N \propto n^\phi$. Averaging over $z \in \mathcal{Z}_\alpha$ and taking a convex combination of according to $Q_h \in \Delta(\mathcal{H})$ then delivers via Jensen's Inequality that with probability at least $1 - \delta/4$ for all $Q \in \Delta(\mathcal{H})$

$$\left|\widehat{\text{cost}}(Q_h) - \text{cost}(Q_h)\right| \leq \tilde{O}(n^{-\phi}). \tag{19}$$

Next, define $\widehat{\text{disp}}_z(h) := \beta_0\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,0}\right] + \beta_1\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,1}\right]$ and $\text{disp}_z(h) := \beta_0\mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,0}\right] + \beta_1\mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,1}\right]$, where the difference can be expressed as

$$\widehat{\text{disp}}_z(h) - \text{disp}_z(h) =$$
$$\beta_0\left(\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,0}\right]\right) +$$
$$\beta_1\left(\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,1}\right]\right).$$

Therefore, by the triangle inequality,

$$\left|\widehat{\text{disp}}_z(h) - \text{disp}_z(h)\right| \leq$$
$$|\beta_0|\left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,0}\right]\right| +$$
$$|\beta_1|\left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,1}\right]\right|.$$

For each term on the right-hand side of the previous display, we invoke Lemma 7 applied to the data distribution conditional on $\mathcal{E}_0$ and $\mathcal{E}_1$. We set $S = 1$, $U = (X_i, z)$, $\mathcal{G} = \mathcal{H}$ and $\psi(s,t) = st$. With probability at least $1 - \frac{\delta}{4}$ for all $z \in \mathcal{Z}_\alpha$,

$$\left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,0}\right]\right| \leq$$
$$R_{n_0}(\mathcal{H}) + \frac{2}{\sqrt{n_0}} + \sqrt{\frac{2\ln(8N/\delta)}{n_0}},$$
$$\left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_{i,1}\right]\right| \leq$$
$$R_{n_1}(\mathcal{H}) + \frac{2}{\sqrt{n_1}} + \sqrt{\frac{2\ln(8N/\delta)}{n_1}}.$$

Then, averaging over $z \in \mathcal{Z}_\alpha$ and taking a convex combination according to $Q_h \in \Delta(\mathcal{H})$ delivers via Jensen's Inequality that with probability at least $1 - \delta/4$ for all $Q \in \Delta(\mathcal{H})$

$$\left|\hat{\mathbb{E}}\left[Q_h|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[Q_h|\mathcal{E}_{i,0}\right]\right| \leq R_{n_0}(\mathcal{H}) + \frac{2}{\sqrt{n_0}} + \sqrt{\frac{2\ln(8N/\delta)}{n_0}} \tag{20}$$

$$\left|\hat{\mathbb{E}}\left[Q_h|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[Q_h|\mathcal{E}_{i,1}\right]\right| \leq R_{n_1}(\mathcal{H}) + \frac{2}{\sqrt{n_1}} + \sqrt{\frac{2\ln(8N/\delta)}{n_1}} \tag{21}$$

By the union bound, both inequalities hold with probability at least $1-\delta/2$.

Finally, Hoeffding's Inequality implies that with probability at least $1-\delta/4$,

$$|\hat{c}_0 - c_0| \leq \sqrt{\frac{-\log(\delta/8)}{2n}}. \tag{22}$$

From Lemma 6, we have that Algorithm 3 terminates and delivers a distribution $\hat{Q}_h$ that compares favorably against any feasible Q in the discretized sample problem. That is, for any such Q_h,

$$\widehat{\text{disp}}(\hat{Q}_h) \leq \widehat{\text{disp}}(Q_h) + O(n^{-\phi}) \tag{23}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} + O(n^{-\phi}) \tag{24}$$

where we used the fact that $\nu \propto n^{-\phi}$ and $B \propto n^{\phi}$ by assumption. First, (19), (22), (24) imply

$$\text{cost}(\hat{Q}_h) \leq \hat{\epsilon} + \tilde{O}(n^{-\phi}) \leq \epsilon - c_0 + \tilde{O}(n^{-\phi}), \tag{25}$$

where we used that $\hat{\epsilon} = \epsilon - \hat{\mathbb{E}}[l(\underline{Y}_i^*, \frac{\alpha}{2})] + C'n^{-\phi} - C''n^{-1/2}$. by assumption. Second, the bounds in (20), (21) imply

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(Q_h) + \tilde{O}(n_0^{-\beta}) + \tilde{O}(n_1^{-\phi}). \tag{26}$$

We assumed that Q_h was a feasible point in the discretized sample problem. Assuming that (19) holds implies that any feasible solution of the population problem is also feasible in the empirical problem due to how we have set C' and C''. Therefore, we have just shown in (25), (26) that $\hat{Q}_h$ is approximately feasible and approximately optimal in the discretized population problem (5). Our last step is to relate $\hat{Q}_h$ to the original problem over $f \in \mathcal{F}$ (2).

From Lemma 1 in Agarwal et al. (2019) and (25), we observe that

$$\text{loss}_\alpha(\hat{Q}_h) \overset{(1)}{\leq} \epsilon + \tilde{O}(n^{-\phi}),$$

$$\text{loss}(\hat{Q}_h) \overset{(2)}{\leq} \epsilon + \tilde{O}(n^{-\phi}),$$

where (1) used Lemma 1 in Agarwal et al. (2019) and we now view $\hat{Q}_h$ as a distribution of risk scores $f \in \mathcal{F}$, (2) used that $\text{loss}(Q) \leq \text{loss}_\alpha(Q) + \alpha$. Next, from Lemma 1 and (26), we observe that

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + (|\beta_0| + |\beta_1|)\,\alpha + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}).$$

where $\hat{Q}_h$ is viewed as a distribution over risk scores $f \in \mathcal{F}$ and $\tilde{Q}$ is now any distribution over risk scores $f \in \mathcal{F}$ that is feasible in the fairness frontier problem. This proves the result for Case I.

Case II: There is no feasible solution to the population problem (4) This follows the proof of Case II in Theorem 3 of Agarwal et al. (2019). If the algorithm returns a ν-approximate saddle point $\hat{Q}_h$, then the theorem holds vacuously since there is no feasible $\tilde{Q}$. Similarly, if the algorithm returns $null$, then the theorem also holds. □

Proof of Theorem 2

Under oracle access to $\mu(x)$, the iteration complexity and bound on cost hold immediately from Theorem 1. The bound on disparity holds immediately for choices $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ that depend on only A. For choices of $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ that depends on Y_i, such as the qualified affirmative action fairness-enhancing intervention, we rely on Lemma 8. We first observe that under oracle access to $\mu(x)$, we can identify any disparity as

$$\frac{\beta_1 \mathbb{E}[f(X)g(\mu(X)) \mid A=1]}{\mathbb{E}[g(\mu(X)) \mid A=1]} - \frac{\beta_0 \mathbb{E}[f(X)g(\mu(X)) \mid A=0]}{\mathbb{E}[g(\mu(X)) \mid A=0]}, \tag{27}$$

where $g(x) = x$ for the balance for the positive class and qualified affirmative action criteria; $g(x) = (1-x)$ for balance for the negative class; and $g(x) = 1$ for the statistical parity and the affirmative action criteria (see proof of Lemma 8 below proof for an example). We define the shorthand

$$\begin{aligned}
\omega_1 &:= \mathbb{E}[f(X)g(\mu(X)) \mid A=1] \\
\bar{\omega}_1 &:= \mathbb{E}[g(\mu(X)) \mid A=1] \\
\omega_0 &:= \mathbb{E}[f(X)g(\mu(X)) \mid A=0] \\
\bar{\omega}_0 &:= \mathbb{E}[g(\mu(X)) \mid A=0]
\end{aligned}$$

and we use $\hat{\omega}_1, \hat{\bar{\omega}}_1, \hat{\omega}_0$, and $\hat{\bar{\omega}}_0$ to denote their empirical estimates. Lemma 8 gives the following bound on the empirical estimate of the disparity:

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \\
&\leq 4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \\
&+ 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]
\end{aligned}$$

where $\omega_\vee = \max(\omega_1, \omega_0)$, $\bar{\omega}_\wedge = \min(\bar{\omega}_1, \bar{\omega}_0)$ and $\beta = \max(|\beta_1|, |\beta_0|)$.

We now proceed to relax and simplify the bound. For $\epsilon \leq 4\frac{\beta\omega_\vee}{\bar{\omega}_\wedge}$, we have

$$2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \geq 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]$$

Case 1: We first consider the likely case that $\bar{\omega}_\wedge \geq \omega_\vee$.

Then we have

$$2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \leq 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^2}{64\beta^2}\right]$$

1a) If

$$\frac{\epsilon\bar{\omega}_\wedge}{8\beta} \geq 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} \tag{28}$$

then

$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^2}{64\beta^2}\right] \leq \exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right]$$

Then we have

$$\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \tag{29}$$

$$\leq 8\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \tag{30}$$

Inverting this bound yields the following: with probability at least $1-\delta$,

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq$$

$$\frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2}{n}\log\left(\frac{8}{\delta}\right)}\right)$$

1b)

$$\frac{\epsilon\bar{\omega}_\wedge}{8\beta} < 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} \tag{31}$$

implies that

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq$$

$$\frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}}\right).$$

Case 2: We now consider the unlikely but plausible case that $\bar{\omega}_\wedge < \omega_\vee$. Then we have

$$\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \leq$$

$$\exp\left[-\frac{n}{2}\left(\frac{\epsilon\omega_\vee}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right]$$

and

$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \leq \exp\left[\frac{-n\epsilon^2\omega_\vee^2}{64\beta^2}\right]$$

We proceed with the same steps as in Case 1 to conclude that with probability at least $1-\delta$,

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq$$

$$\frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2}{n}\log\left(\frac{8}{\delta}\right)}\right)$$

Applying our assumption that

$$R_n(\mathcal{H}) \leq Cn^{-\phi} \text{ and } \hat{\epsilon} = \epsilon - \hat{c}_0 + C'n^{-\phi} - C''n^{-1/2}.$$

for $\phi \leq 1/2$ and $C' \geq 2C + 2 + \sqrt{2\ln(8N/\delta)}$ and $C'' \geq \sqrt{\frac{-\log(\delta/8)}{2}}$, then

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + \tilde{O}(n^{-\phi}), \tag{32}$$

which implies

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}). \tag{33}$$

□

Proof of Theorem 3

The claim about the iteration complexity of Algorithm 4 follows from Lemma 9 after substituting in the stated choices of ν, B_λ. We consider two cases.

Case 1: There is a feasible solution $\tilde{Q}$ to the population problem (11) Using Lemmas 11-13, the ν-approximate saddle point $(\hat{\xi}, \hat{Q}_h)$ satisfies

$$\widehat{\text{disp}}(\hat{Q}_h) - \hat{\xi} \leq \frac{B_\xi + 2\nu}{B_\lambda}, \tag{34}$$

$$-\widehat{\text{disp}}(\hat{Q}_h) - \hat{\xi} \leq \frac{B_\xi + 2\nu}{B_\lambda} \tag{35}$$

$$\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}_{\text{cost}} \leq \frac{B_\xi + 2\nu}{B_\lambda} \tag{36}$$

for any (ξ, Q) that is feasible in the empirical problem. This implies that Algorithm 4 returns $\hat{Q} \neq null$. We will now show that the $(\hat{\xi}, \hat{Q})$ provides an approximate solution to the discretized population problem.

First, through the same argument as in the proof of Theorem 1, we obtain that with probability at least $1 - \delta/4$ for all $Q_h \in \Delta(\mathcal{H})$

$$\left|\widehat{\text{cost}}(Q_h) - \text{cost}(Q_h)\right| \leq \tilde{O}(n^{-\phi}). \tag{37}$$

Second, with probability at least $1 - \delta/2$ for all $Q \in \Delta(\mathcal{H})$,

$$\left|\hat{\mathbb{E}}[Q_h|\mathcal{E}_{i,0}] - \mathbb{E}[Q_h|\mathcal{E}_{i,0}]\right| \leq \tilde{O}(n_0^{-\phi}) \tag{38}$$

$$\left|\hat{\mathbb{E}}[Q_h|\mathcal{E}_{i,1}] - \mathbb{E}[Q_h|\mathcal{E}_{i,1}]\right| \leq \tilde{O}(n_1^{-\phi}). \tag{39}$$

Finally, Hoeffding's Inequality implies that with probability at least $1-\delta/4$,

$$|\hat{c}_0 - c_0| \leq \sqrt{\frac{-\log(\delta/8)}{2n}}. \tag{40}$$

From Lemma 14, we have that Algorithm 4 terminates and delivers $(\hat{\xi}, \hat{Q}_h)$ that compares favorable with any feasible (ξ, Q_h) in the discretized sample problem. That is, for any such (ξ, Q_h),

$$\hat{\xi} \leq \xi + O(n^{-\phi}), \tag{41}$$

$$\widehat{\text{disp}}(\hat{Q}_h) \leq \hat{\xi} + O(n^{-\phi}), \tag{42}$$

$$-\widehat{\text{disp}}(\hat{Q}_h) \leq \hat{\xi} + O(n^{-\phi}) \tag{43}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon}_{\text{cost}} + O(n^{-\phi}) \tag{44}$$

Notice that (37), (40) and (44) imply that

$$\text{cost}(\hat{Q}_h) \leq \epsilon - c_0 + \tilde{O}(n^{-\phi}), \tag{45}$$

where we used that $\hat{\epsilon} = \epsilon - \hat{c}_0 + C'n^{-\phi} - C''n^{-\phi}$. For any feasible (ξ, Q_h), then $(|\text{disp}(Q_h)|, Q_h)$ is also feasible. Then, combining (41)-(43) yields

$$\left|\text{disp}(\hat{Q}_h)\right| \leq |\text{disp}(Q_h)| + \tilde{O}(n^{-\phi}) \tag{46}$$

Second, notice that this implies that

$$\left|\text{disp}(\hat{Q}_h)\right| \leq |\text{disp}(Q_h)| + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}) \tag{47}$$

We assumed that (ξ, Q_h) were feasible in the discretized sample problem. Assuming that (37) holds implies that any feasible solution of the population problem is also feasible in the empirical problem due to how we set C', C''. Therefore, we have just shown that $(\hat{\xi}, \hat{Q}_h)$ are approximately optimal in the discretized population problem.

Then, following the proof of Theorem 1, we observe that $\text{loss}(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$, where we now interpret $\hat{Q}_h$ as a distribution over risk scores $f \in \mathcal{F}$. This proves the result for Case I.

Case II: There is no feasible solution to the population problem (11) This follows the proof of Case II of Theorem 3 in Agarwal et al. (2019). If the algorithm returns a ν-approximate saddle point $\hat{Q}_h$, then the theorem holds vacuously since there is no feasible $\tilde{Q}$. Similarly, if the algorithm returns $null$, then the theorem also holds. □

C. Auxiliary Lemmas for Main Results

In this section, we state and prove a series of auxiliary lemmas that are used in the proofs of our main results in the text.

C.1. Auxiliary Lemmas for the Proof of Theorem 1

C.1.1. Iteration Complexity of Algorithm 3

Lemma 2. *Letting* $\rho := \max_{h\in\mathcal{H}} |\widehat{\text{cost}}(h) - \hat{\epsilon}|$, *Algorithm 3 satisfies the inequality*

$$\nu_t \leq \frac{B\log(2)}{\eta t} + \eta\rho^2 B.$$

For $\eta = \frac{\nu}{2\rho^2 B}$, *Algorithm 3 will return a* ν*-approximate saddle point of* L *in at most* $\frac{4\rho^2 B^2 \log(2)}{\nu^2}$. *Since in our setting,* $\rho \leq 1$, *the iteration complexity of Algorithm 3 is* $4B^2\log(2)/\nu^2$.

Proof. Follows immediately from the proof of iteration complexity in Theorem 3 of Agarwal et al. (2019). Since the cost is bounded on $[-1, 1]$ and $\widehat{\text{cost}}(h) - \hat{\epsilon} \leq \widehat{\text{cost}}(h) \leq 1$ for any $h \in \mathcal{H}$, we see that $\rho \leq 1$. □

C.1.2. Solution Quality for Algorithm 3

Let $\Lambda := \{\lambda \in \mathbb{R}_+ : \lambda \leq B\}$ denote the domain of λ. Throughout this section, we assume we are given a pair $(\hat{Q}_h, \hat{\lambda})$ that is a ν-approximate saddle point of the Lagrangian

$$L(\hat{Q}_h, \hat{\lambda}) \leq L(Q_h, \hat{\lambda}) + \nu \text{ for all } Q_h \in \Delta(\mathcal{H}),$$
$$L(\hat{Q}_h, \hat{\lambda}) \geq L(\hat{Q}_h, \lambda) - \nu \text{ for all } 0 \leq \lambda \leq B.$$

We extend Lemma 1, Lemma 2 and Lemma 3 of Agarwal et al. (2018) to our setting.

Lemma 3. *The pair* $(\hat{Q}_h, \hat{\lambda})$ *satisfies*

$$\hat{\lambda}\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right) \geq B\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right)_+ - \nu,$$

where $(x)_+ = \max\{x, 0\}$.

Proof. We consider a dual variable λ that is defined as

$$\lambda = \begin{cases} 0 \text{ if } \widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} \\ B \text{ otherwise.} \end{cases}$$

From the ν-approximate optimality conditions,

$$\begin{aligned}\widehat{\text{disp}}(\hat{Q}) + \hat{\lambda}\left(\widehat{\text{cost}}(\hat{Q}) - \hat{\epsilon}\right) &= L(\hat{Q}, \hat{\lambda}) \\ &\geq L(\hat{Q}, \lambda) - \nu \\ &= \widehat{\text{disp}}(\hat{Q}) + \lambda\left(\widehat{\text{cost}}(Q) - \hat{\epsilon}\right),\end{aligned}$$

and the claim follows by our choice of λ. □

Lemma 4. *The distribution* $\hat{Q}_h$ *satisfies*

$$\widehat{\text{disp}}(\hat{Q}_h) \leq \widehat{\text{disp}}(Q_h) + 2\nu$$

for any Q_h *satisfying the empirical constraint (i.e., any* Q_h *such that* $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$*).*

Proof. Assume Q_h satisfies $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$. Since $\hat{\lambda} \geq 0$, we have that

$$L(Q_h, \hat{\lambda}) = \widehat{\text{disp}}(Q_h) + \hat{\lambda}\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}\right) \leq \widehat{\text{disp}}(Q_h).$$

Moreover, the ν-approximate optimality conditions imply that $L(\hat{Q}_h, \hat{\lambda}) \leq L(Q_h, \hat{\lambda}) + \nu$. Together, these inequalities imply that

$$L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disp}}(Q_h) + \nu.$$

Next, we use Lemma 3 to construct a lower bound for $L(\hat{Q}_h, \hat{\lambda})$. We have that

$$\begin{aligned} L(\hat{Q}_h, \hat{\lambda}) &= \widehat{\text{disp}}(\hat{Q}_h) + \hat{\lambda}\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}'\right) \\ &\geq \widehat{\text{disp}}(\hat{Q}_h) + B\left(\widehat{\text{cost}}(\hat{Q}) - \hat{\epsilon}'\right)_+ - \nu \\ &\geq \widehat{\text{disp}}(\hat{Q}_h) - \nu. \end{aligned}$$

By combining the inequalities $L(\hat{Q}_h, \hat{\lambda}) \geq \widehat{\text{disp}}(\hat{Q}_h) - \nu$ and $L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disp}}(Q_h) + \nu$, we arrive at the claim. □

Lemma 5. *Assume the empirical constraint $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$ is feasible. Then, the distribution $\hat{Q}_h$ approximately satisfies the empirical cost constraint with*

$$\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon} \leq \frac{|\beta_0| + |\beta_1| + 2\nu}{B}.$$

Proof. Let Q_h satisfy $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$. Recall from the proof of Lemma 4, we showed that

$$\widehat{\text{disp}}(\hat{Q}_h) + B\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right)_+ - \nu \leq L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disp}}(Q_h) + \nu.$$

Therefore, we observe that

$$B\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}\right) \leq \left(\widehat{\text{disp}}(Q_h) - \widehat{\text{disp}}(\hat{Q}_h)\right) + 2\nu.$$

Since we can bound $\widehat{\text{disp}}(Q_h) - \widehat{\text{disp}}(\hat{Q}_h)$ by $|\beta_0| + |\beta_1|$, the result follows. □

Lemma 6. *Suppose that Q_h is any feasible solution to discretized sample problem. Then, the solution $\hat{Q}_h$ returned by Algorithm 3 satisfies*

$$\begin{aligned} \widehat{\text{disp}}(\hat{Q}_h) &\leq \widehat{\text{disp}}(Q_h) + 2\nu \\ \widehat{\text{cost}}(\hat{Q}_h) &\leq \hat{\epsilon} + \frac{|\beta_0| + |\beta_1| + 2\nu}{B}. \end{aligned}$$

Proof. This is an immediate consequence of Lemma 2, Lemma 4 and Lemma 5. If the algorithm returns *null*, then these inequalities are vacuously satisfied. □

C.1.3. Concentration Inequality

We restate Lemma 2 in Agarwal et al. (2019), which provides a uniform concentration inequality on the convergence of a sample moment over a function class.

Let $\mathcal{G}$ be a class of functions $g\colon \mathcal{U} \to \mathbb{R}$ over some space $\mathcal{U}$. The *Rademacher complexity* of the function class $\mathcal{G}$ is defined as

$$R_n(\mathcal{G}) := \sup_{u_1,\ldots,u_n \in \mathcal{U}} \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(u_i)\right|\right],$$

where the expectation is defined over the i.i.d. random variables $\sigma_1, \ldots, \sigma_n$ with $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$.

Lemma 7 (Lemma 2 in Agarwal et al. (2019))**.** *Let D be a distribution over a pair of random variables (S, U) taking values in $\mathcal{S} \times \mathcal{U}$. Let $\mathcal{G}$ be a class of functions $g\colon \mathcal{U} \to [0,1]$, and let $\psi\colon \mathcal{S} \times [0,1] \to [-1,1]$ be a contraction in its second argument (i.e., for all $s \in \mathcal{S}$ and $t, t' \in [0,1]$, $|\psi(s,t) - \psi(s,t')| \leq |t - t'|$). Then, with probability $1-\delta$, for all $g \in \mathcal{G}$,*

$$\left|\hat{\mathbb{E}}\left[\psi(S, g(U))\right] - \mathbb{E}\left[\psi(S, g(U))\right]\right| \leq 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}},$$

where the expectation is with respect to D and the empirical expectation is based on n i.i.d. draws from D. If ψ is linear in its second argument, then a tighter bound holds with $4R_n(\mathcal{G})$ replaced by $2R_n(\mathcal{G})$.

C.2. Auxiliary Lemmas for the Proof of Theorem 2

C.2.1. Concentration result for disparity under selective labels

Lemma 8.

$$\begin{aligned} &\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \\ &\leq 4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \\ &+ 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right] \end{aligned}$$

where $\omega_\vee = \max(\omega_1, \omega_0)$, $\bar{\omega}_\wedge = \min(\bar{\omega}_1, \bar{\omega}_0)$ and $\beta = \max(|\beta_1|, |\beta_0|)$

Proof. For exposition, we first show the steps for qualified affirmative action and then extend the result to the general disparity. We can rewrite the qualified affirmative action criterion as

$$\mathbb{E}[f(X)|Y=1, A=1] = \frac{\mathbb{E}[f(X)\mu(X)|A=1]}{\mathbb{E}[\mu(X)|A=1]} \quad (48)$$

where $\mu(x) := \mathbb{E}[Y \mid X = x]$.

$$\mathbb{E}[f(X)|Y=1, A=1]$$

$$= \frac{\mathbb{E}[f(X)1\{Y=1\}|A=1]}{P(Y=1|A=1)} \tag{49}$$

$$= \frac{\mathbb{E}[f(X)\mathbb{E}[1\{Y=1\}|X,A=1]|A=1]}{E[P(Y=1|X,A=1)|A=1]} \tag{50}$$

$$= \frac{\mathbb{E}[f(X)P(Y=1|X,A=1)|A=1]}{E[\mu(X)|A=1]} \tag{51}$$

$$= \frac{\mathbb{E}[f(X)\mu(X)|A=1]}{E[\mu(X)|A=1]} \tag{52}$$

Assuming access to the oracle μ function, we can estimate this on the full training data as

$$\frac{\hat{\mathbb{E}}[f(X)\mu(X, A=1)|A=1]}{\hat{\mathbb{E}}[\mu(X, A=1)|A=1]} \tag{53}$$

Next we will make use of Lemma 2 of Agarwal et al. (2019), which we restate here again for convenience. Under certain conditions on ϕ and g, with probability at least $1-\delta$

$$\left|\hat{\mathbb{E}}\left[\phi(S, g(U))\right] - \mathbb{E}\left[\phi(S, g(U))\right]\right| \leq$$

$$4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}}.$$

We invert the bound by setting $\epsilon = 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}}$ and solving for δ to get

$$\delta = 2\exp\left[-\frac{n}{2}\left(\epsilon - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \tag{54}$$

Now we can restate Lemma 2 of Agarwal et al. (2019) as

$$\mathbb{P}\Big[\left|\hat{\mathbb{E}}\left[\phi(S, g(U))\right] - \mathbb{E}\left[\phi(S, g(U))\right]\right| > \epsilon\Big] \tag{55}$$

$$\leq 2\exp\left[-\frac{n}{2}\left(\epsilon - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right]$$

Next we revisit the quantity that we want to bound:

$$\left|\frac{\omega}{\bar{\omega}} - \frac{\hat{\omega}}{\hat{\bar{\omega}}}\right| \tag{56}$$

where $\omega = \mathbb{E}[f(X)\mu(X, A=1)|A=1]$ and $\bar{\omega} = \mathbb{E}[\mu(X, A=1)|A=1]$ and correspondingly for $\hat{\omega}$ and $\hat{\bar{\omega}}$. We will rewrite Expression 56 as a ratio of differences. We have

$$\left|\frac{\hat{\omega}}{\hat{\bar{\omega}}} - \frac{\omega}{\bar{\omega}}\right| = \left|\frac{\hat{\omega}\bar{\omega} - \hat{\bar{\omega}}\omega}{\hat{\bar{\omega}}\bar{\omega}}\right| \tag{57}$$

$$= \left|\frac{\bar{\omega}(\hat{\omega}-\omega) - \omega(\hat{\bar{\omega}}-\bar{\omega})}{\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2}\right| \tag{58}$$

(59)

By triangle inequality and union bound, we have

$$\mathbb{P}\Big[\left|\frac{\bar{\omega}(\hat{\omega}-\omega) - \omega(\hat{\bar{\omega}}-\bar{\omega})}{\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2}\right| \geq \frac{t}{\bar{\omega}^2/2}\Big]$$

$$< \mathbb{P}\Big[|\bar{\omega}(\hat{\omega}-\omega)| + |\omega(\hat{\bar{\omega}}-\bar{\omega})| \geq t\Big] + \mathbb{P}\Big[|(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2| \leq \frac{\bar{\omega}^2}{2}\Big]$$

$$< \mathbb{P}\big[|\bar{\omega}(\hat{\omega}-\omega)| \geq \frac{t}{2}\big] + \mathbb{P}\big[|\omega(\hat{\bar{\omega}}-\bar{\omega})| \geq \frac{t}{2}\big] + \mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2| \leq \frac{\bar{\omega}^2}{2}\Big]$$

Since $0 \leq \mu(X, A=1) \leq 1$, we can use a Hoeffding bound for the quantity $|(\hat{\bar{\omega}} - \bar{\omega})|$. Note that $0 \leq \omega \leq \bar{\omega} \leq 1$. Then applying Hoeffding's inequality gives us

$$\mathbb{P}\big[\,|\omega(\hat{\bar{\omega}} - \bar{\omega})| \geq \frac{t}{2}\big] \leq 2\exp\left[\frac{-nt^2}{4\omega^2}\right] \tag{60}$$

Next we bound the third term:

$$\mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2| \leq \frac{\bar{\omega}^2}{2}\Big] \leq \mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})| \geq \frac{\bar{\omega}^2}{2}\Big] \tag{61}$$

$$= \mathbb{P}\Big[|(\hat{\bar{\omega}}-\bar{\omega})| \geq \frac{\bar{\omega}}{2}\Big] \tag{62}$$

$$\leq 2\exp\left[\frac{-n\bar{\omega}^2}{4}\right] \tag{63}$$

where we again used Hoeffding's inequality for the last line.

We bound the first term using the restated Lemma in 55:

$$\mathbb{P}\big[|\bar{\omega}(\hat{\omega}-\omega)| \geq \frac{t}{2}\big] \leq 2\exp\left[-\frac{n}{2}\left(\frac{t}{2\bar{\omega}} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \tag{64}$$

Now we let $\tilde{\epsilon} = \frac{t}{\bar{\omega}^2/2}$ to get

$$\mathbb{P}\Big[\left|\frac{\hat{\omega}}{\hat{\bar{\omega}}} - \frac{\omega}{\bar{\omega}}\right| \geq \tilde{\epsilon}\Big] \tag{65}$$

$$\leq 2\exp\left[-\frac{n}{2}\left(\frac{\tilde{\epsilon}\bar{\omega}}{4} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + \exp\left[\frac{-n\tilde{\epsilon}^2\bar{\omega}^4}{16\omega^2}\right] + \exp\left[\frac{-n\bar{\omega}^2}{4}\right]$$

Now we turn to the general case. Recalling that we define $\beta = \max(|\beta_1, \beta_0|)$, we have

$$\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \Big(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\Big)\right| \geq \epsilon\right] \leq$$

$$\mathbb{P}\left[|\beta_1|\left|\frac{\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\omega_1}{\bar{\omega}_1}\right| + |\beta_0|\left|\frac{\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \frac{\omega_0}{\bar{\omega}_0}\right| \geq \epsilon\right] \leq$$

$$\mathbb{P}\left[\left|\frac{\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\omega_1}{\bar{\omega}_1}\right| \geq \frac{\epsilon}{2\beta}\right] + \mathbb{P}\left[\left|\frac{\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \frac{\omega_0}{\bar{\omega}_0}\right| \geq \frac{\epsilon}{2\beta}\right] \leq$$

$$2\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_1}{8\beta}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+\exp\left[\frac{-n\epsilon^2\bar{\omega}_1^4}{64\beta\omega_1^2}\right]+$$
$$\exp\left[\frac{-n\bar{\omega}_1^2}{4}\right]+2\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_0}{8\beta}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+$$
$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_0^4}{64\beta\omega_0^2}\right]+\exp\left[\frac{-n\bar{\omega}_0^2}{4}\right]\leq$$
$$4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right]$$
$$+2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]$$

where the first inequality holds by triangle inequality, the second inequality holds by the union bound, the third inequality applies (65) for $\tilde{\epsilon}=\frac{\epsilon}{2\beta}$, and the final inequality simplifies the bound using the notation $\omega_\vee=\max(\omega_1,\omega_0)$ and $\bar{\omega}_\wedge=\min(\bar{\omega}_1,\bar{\omega}_0)$. □

C.3. Auxiliary Lemmas for the Proof of Theorem 3

C.3.1. Iteration Complexity for Algorithm 4

Lemma 9. *Defining* $\rho := \max_{h\in\mathcal{H},\xi\in[0,B_\xi]}\max\{\widehat{\text{disp}}(h)-\xi,-\widehat{\text{disp}}(h)-\xi,\widehat{\text{cost}}(h)-\hat{\epsilon}\}$, *Algorithm 4 satisfies the inequality*

$$\nu_t\leq\frac{B_\lambda\log(3)}{\eta t}+\eta\rho^2B.$$

For $\eta=\frac{\nu}{2\rho^2B_\lambda}$, *Algorithm 4 will return a* ν*-approximate saddle point of* L *in at most* $\frac{4\rho^2B_\lambda^2\log(3)}{\nu^2}$ *iterations. Setting* $B_\xi=1$, *we observe* $\rho\leq1$, *and so the iteration complexity of Algorithm 4 is* $\frac{4B_\lambda^2\log(3)}{\nu^2}$.

Proof. Follows immediately from the proof of Theorem 3 in Agarwal et al. (2019) and the same argument given in the proof of Lemma 2. □

C.3.2. Solution Quality for Algorithm 4

Let $\Lambda=\{\lambda\in\mathbb{R}^3_+:\|\lambda\|\leq B_\lambda\}$. Assume we are given $\left(\hat{\xi},\hat{Q}_h,\hat{\lambda}\right)$, which is a ν-approximate saddle point satisfying $L(\hat{\xi},\hat{Q}_h,\hat{\lambda})\leq L(\xi,Q_h,\hat{\lambda})+\nu$ for all $Q_h\in\Delta(\mathcal{H}),\xi\in[0,B_\xi]$ and $L(\hat{\xi},\hat{Q}_h,\hat{\lambda})\geq L(\hat{\xi},\hat{Q}_h,\lambda)-\nu$ for all $\|\lambda\|\leq B_\lambda$. We extend Lemmas 3-5 to the problem of finding the absolute disparity minimizing model.

Lemma 10. $\left(\hat{\xi},\hat{Q}_h,\hat{\lambda}\right)$ *satisfies*

$$\hat{\lambda}_+(\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi})+\hat{\lambda}_-(-\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi})+\hat{\lambda}_{cost}\left(\widehat{\text{cost}}(\hat{Q}_h)-\hat{\epsilon}\right)$$
$$\geq B_\lambda\max\{\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi},-\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi},\widehat{\text{cost}}(\hat{Q}_h)-\hat{\epsilon}\}-\nu.$$

Proof. The argument is the same as the proof of Lemma 3. □

Lemma 11. *The value* $\hat{\xi}$ *satisfies*

$$\hat{\xi}\leq\xi+2\nu$$

for any ξ *such that there exists* Q_h *satisfying* $\widehat{\text{disp}}(Q_h)-\xi\leq0$, $-\widehat{\text{disp}}(Q_h)-\xi\leq0$ *and* $\widehat{\text{cost}}(Q_h)\leq\hat{\epsilon}$.

Proof. Assume the pair (ξ,Q_h) satisfies $\widehat{\text{disp}}(Q_h)-\xi\leq0$, $-\widehat{\text{disp}}(Q_h)-\xi\leq0$ and $\widehat{\text{cost}}(Q_h)\leq\hat{\epsilon}$. Since $\hat{\lambda}\geq0$, we have that $L(\xi,Q,\hat{\lambda})\leq\xi$. Moreover, the ν-approximate optimality conditions imply that $L(\hat{\xi},\hat{Q},\hat{\lambda})\leq L(\xi,Q,\hat{\lambda})+\nu$. Together, these inequalities imply that

$$L(\hat{\xi},\hat{Q},\hat{\lambda})\leq\xi+\nu.$$

Next, we can use Lemma 10 to construct a lower bound for $L(\hat{\xi},\hat{Q},\hat{\lambda})$. To do so, observe that

$$\begin{aligned}&L(\hat{\xi},\hat{Q},\hat{\lambda})\\&\geq\hat{\xi}+B_\lambda\max\{\widehat{\text{disp}}(\hat{Q})-\hat{\xi},-\widehat{\text{disp}}(\hat{Q})-\hat{\xi},\widehat{\text{cost}}(\hat{Q})-\hat{\epsilon}\}-\nu\\&\geq\hat{\xi}-\nu.\end{aligned}$$

By combining the inequalities, $L(\hat{\xi},\hat{Q},\hat{\lambda})\geq\hat{\xi}-\nu$ and $L(\hat{\xi},\hat{Q},\hat{\lambda})\leq\xi+\nu$, we arrive at the claim. □

Lemma 12. *Assume the empirical cost constraint* $\widehat{\text{cost}}Q_h\leq\hat{\epsilon}$ *and the slack variable constraints* $\widehat{\text{disp}}(Q_h)-\xi\leq0$ *and* $-\widehat{\text{disp}}(Q_h)-\xi\leq0$ *are feasible. Then, the pair* $(\hat{\xi},\hat{Q}_h)$ *satisfies*

$$\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi}\leq\frac{B_\xi+2\nu}{B_\lambda},$$
$$-\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi}\leq\frac{B_\xi+2\nu}{B_\lambda}.$$

Proof. Let ξ be a feasible value of the slack variable such that there exists Q_h satisfying $\widehat{\text{cost}}(Q_h)\leq\hat{\epsilon}$ and the slack variable constraints $\widehat{\text{disp}}(Q_h)-\xi\leq0$, $-\widehat{\text{disp}}(Q_h)-\xi\leq0$. Recall from the Proof of Lemma 11, we showed that

$$\hat{\xi}+B_\lambda\max\{\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi},-\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi},\widehat{\text{cost}}(\hat{Q}_h)-\hat{\epsilon}\}$$
$$-\nu\leq L(\hat{\xi},\hat{Q}_h,\hat{\lambda})\leq\xi+\nu.$$

Therefore, it is immediate that

$$B_\lambda\max\{\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi},-\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi},\widehat{\text{cost}}(\hat{Q}_h)-\hat{\epsilon}\}$$
$$\leq\left(\xi-\hat{\xi}\right)+2\nu.$$

and so

$$B_\lambda\left(\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi}\right)\leq\left(\xi-\hat{\xi}\right)+2\nu,$$
$$B_\lambda\left(-\widehat{\text{disp}}(\hat{Q}_h)-\hat{\xi}\right)\leq\left(\xi-\hat{\xi}\right)+2\nu.$$

Since $\xi \in [0, B_\xi]$, we can bound $\xi - \hat{\xi}$ by B_ξ. The result follows. □

Lemma 13. *Assume the empirical cost constraint* $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$ *and the slack variable constraints* $\widehat{\text{disp}}(Q_h) - \xi \leq 0$, $-\widehat{\text{disp}}(Q_h) - \xi \leq 0$ *are feasible. Then the distribution* $\hat{Q}_h$ *satisfies*

$$\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon} \leq \frac{B_\xi + 2\nu}{B_\lambda}.$$

Proof. The proof is analogous to the proof of Lemma 12. □

Lemma 14. *Suppose that* (ξ, Q_h) *is a feasible solution to the empirical version of (13). Then, the solution* $(\hat{\xi}, \hat{Q}_h)$ *returned by Algorithm 4 satisfies*

$$\begin{aligned}
&\hat{\xi} \leq \xi + 2\nu, \\
&\widehat{\text{disp}}(\hat{Q}_h) - \hat{\xi} \leq \frac{B_\xi + 2\nu}{B_\lambda}, \\
&-\widehat{\text{disp}}(\hat{Q}_h) - \hat{\xi} \leq \frac{B_\xi + 2\nu}{B_\lambda} \\
&\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon} \leq \frac{B_\xi + 2\nu}{B_\lambda}.
\end{aligned}$$

Proof. The proof follows from Lemmas 12-13. If the algorithm returns $null$, then these inequalities are vacuously satisfied. □

D. Additional Experimental Details and Results

In this section, we present additional details on our experimental setup as well as additional results for both experiments presented in the main paper.

D.1. Recidivism Risk Prediction: Additional Results

ProPublica's COMPAS recidivism data (Angwin et al., 2016) contains 7,214 examples. We randomly split this data 50%-50% into a train and test set. We evaluate models using logistic regression loss, defined as $l(y, f(x)) = \log(1 + e^{-C(2y-1)(2f(x)-1)})/(\log(1 + e^C))$ for $C = 5$. We ran the exponentiated gradient algorithm for at most 500 iterations on a fixed discretization grid, $\mathcal{Z}_\alpha = \{1/40, 2/40, \ldots, 1\}$. Letting $n = 3,607$, we set the parameters of the exponentiated gradient algorithm to be $B = \sqrt{n}/2$ for minimization problems, $B = \sqrt{n}$ for maximization problems, $\nu = 1/\sqrt{n}$ and $\eta = 2$. We report the average run time results for a single run of the exponentiated gradient algorithm to solve the minimization and maximization problems for each disparity measure in Table 2 below. These experiments were conducted on a 2012 MacBook Pro with a 2.3 GHz Quad-Core Intel Core i7.

Table 2. Timing for the recidivism risk prediction experiment on the ProPublic COMPAS dataset. We report the average time for the exponentiated gradient algorithm to complete at most 500 iterations on the train set ($n_{train} = 3,607$) in computing the disparity minimizing model (Min. Disp.) and the disparity maximizing model (Max. Disp.). Timing is reported in minutes. See § 6 for details.

	Timing (in minutes)	
	Min. Disp.	Max. Disp.
SP	7.29	24.10
BFPC	8.45	24.18
BFNC	22.24	23.64

Table 3. The disparity minimizing and disparity maximizing models over the set of good models (performing within 1% of COMPAS's training loss) achieve comparable test loss to COMPAS. The first panel (SP) displays the test loss for the models that minimize (Min. Disp.) and maximize (Max. Disp.) the disparity in average predictions for black versus white defendants (Def. 1). The second panel (BFPC) analyzes the test loss for the models that minimize and maximize the disparity in average predictions for black versus white defendants in the positive class, and the third panel examines the test loss for the models that minimize and maximize the disparity in average predictions for black versus white defendants in the negative class (Def. 2). Standard errors are reported in parentheses. See § 6 for details.

	Test loss		
	Min. Disp.	Max. Disp.	COMPAS
SP	0.095 (0.001)	0.067 (0.002)	0.102 (0.003)
BFPC	0.099 (0.003)	0.085 (0.002)	0.102 (0.003)
BFNC	0.094 (0.004)	0.073 (0.001)	0.102 (0.003)

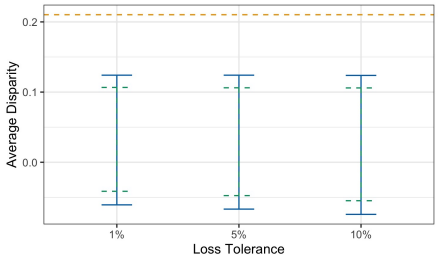
D.1.1. Test Loss

Table 3 reports the test loss of COMPAS and the test losses of the disparity minimizing and disparity maximizing models over the set of good models. The disparity minimizing and disparity maximizing models achieve comparable and in some cases lower test loss than COMPAS.

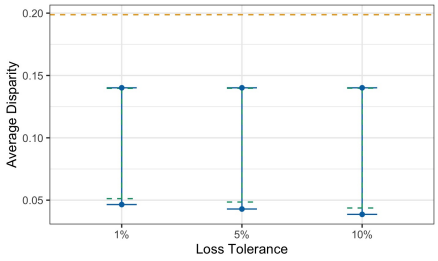
D.1.2. Train Set Performance

Figure 2 plots the range of predictive disparities over the train set when the parameter ϵ is calibrated using COMPAS. We report the train set performance for various choices of the loss tolerance parameter, setting $\epsilon = 1\%, 5\%, 10\%$ of COMPAS' training loss. The blue error bars plot the relative disparities associated with the linear program reduction (§ A.2), the green error bars plot the relative disparities associated with the stochastic prediction function returned by

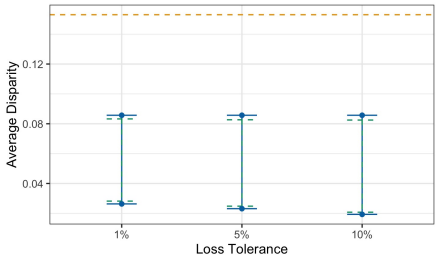
Algorithm 3 and the orange dashed line plots the relative disparity associated with COMPAS. The range of disparities produced by the linear program reduction closely track the range of disparities produced by the stochastic prediction function returned by Algorithm 3 in the train set, confirming the quality of the linear programming reduction.



(a) Statistical Parity



(b) Balance for the Positive Class



(c) Balance for the Negative Class

Figure 2. The minimal and maximal predictive disparities between black defendants ($A_i = 1$) and white defendants ($A_i = 0$) over the set of good models in the train set. We set the loss tolerance as $\epsilon = 1\%, 5\%, 10\%$ of COMPAS' training loss. The blue error bars plot the relative disparities associated with the linear program reduction (§ A.2), the green error bars plot the relative disparities associated with the stochastic prediction function returned by Algorithm 3 and the orange dashed line plots the predictive disparity associated with COMPAS. See § 6 and § D.1.2 for details.

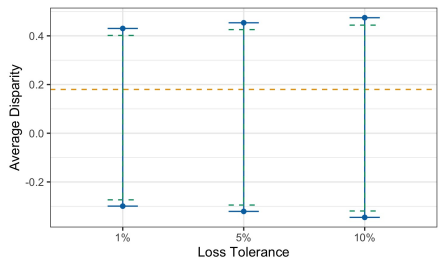
D.1.3. Results for Predictive Disparities across Young and Older Defendants

We also examine the range of predictive disparities between defendants that are younger than 25 years old ($A_i = 1$) and defendants older than 25 years old ($A_i = 0$), focusing on the range of predictive disparities that could be generated by a risk score that is constructed using logistic regression on a quadratic polynomial of the defendant's age and number of prior offenses. We calibrate the loss tolerance parameter ϵ such that (2) constructs the fairness frontier over all models that achieve a logistic regression loss within 1% of COMPAS's training loss. We provide the results for the statistical parity, balance for the positive class, and balance for the negative class disparity measures (Def. 1 and Def 2). Table 4 summarizes the range of predictive disparities over the test set when the parameter ϵ is calibrated using COMPAS' training loss. While COMPAS lies within the range of possible disparities for each measure, notice that there exists a predictive model that produces strictly smaller disparities between young and older defendants than the COMPAS risk assessment at minimal cost to predictive performance. The disparity minimizing and disparity maximizing models over the set of good models achieve a test loss that is comparable to COMPAS (see Table 5).

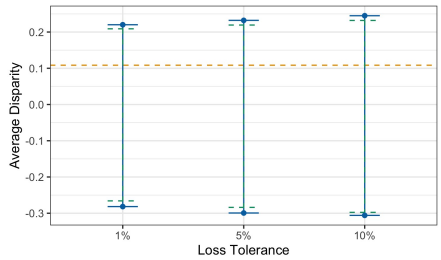
Figure 3 plots the range of predictive disparities over the train set when the parameter ϵ is calibrated using COMPAS. We report the train set performance for various choices of the loss tolerance parameter, setting $\epsilon = 1\%, 5\%, 10\%$ of COMPAS' training loss. The blue error bars plot the relative disparities associated with the linear program reduction (Section A.2), the green error bars plot the relative disparities associated with the stochastic prediction function returned by Algorithm 3 and the orange dashed line plots the relative disparity associated with COMPAS. We again find that the range of disparities produced by the linear program reduction closely track the range of disparities produced by the stochastic prediction function returned by Algorithm 3.

Table 4. The minimal and maximal disparities between young defendants ($A_i = 1$) and older defendants ($A_i = 0$) over the set of good models (performing within 1% of COMPAS' training loss) on the test set. The first panel (SP) displays the disparity in average predictions for young versus older defendants (Def. 1). The second panel (BFPC) displaces the disparity in average predictions for young versus old defendants in the positive class, and the third panel examines the disparity in average predictions for young versus older defendants in the negative class (Def. 2). Standard errors are reported in parentheses. See § D.1.3 of the Supplement for details.

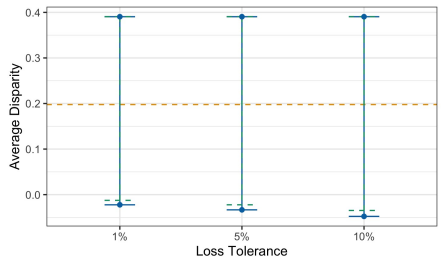
	Min. Disp.	Max. Disp.	COMPAS
SP	-0.296 (0.019)	0.433 (0.008)	0.173 (0.014)
BFPC	-0.207 (0.010)	0.260 (0.008)	0.101 (0.019)
BFNC	-0.040 (0.038)	0.329 (0.008)	0.200 (0.022)



(a) Statistical Parity



(b) Balance for the Positive Class



(c) Balance for the Negative Class

Figure 3. The minimal and maximal disparities between young defendants ($A_i = 1$) and older defendants ($A_i = 0$) over the set of good models on the train set. We set the loss tolerance as $\epsilon = 1\%, 5\%, 10\%$ of COMPAS' training loss. The blue error bars plot the relative disparities associated with the linear program reduction (§ A.2), the green error bars plot the relative disparities associated with the stochastic prediction function returned by Algorithm 3 and the orange dashed line plots the predictive disparity associated with COMPAS. See § D.1.3 of the Supplement for details.

Table 5. The disparity minimizing and disparity maximizing models over the set of good models (performing within 1% of COMPAS's training loss) achieve comparable test loss to COMPAS. The first panel (SP) displays the test loss for the models that minimize (Min. Disp.) and maximize (Max. Disp.) the disparity in average predictions for young versus older defendants (Def. 1). The second panel (BFPC) analyzes the test loss for the models that minimize and maximize the disparity in average predictions for young versus older defendants in the positive class, and the third panel examines the test loss for the models that minimize and maximize the disparity in average predictions for young versus older defendants in the negative class (Def. 2). Standard errors are reported in parentheses. See § 6 for details.

	Test loss		
	Min. Disp.	Max. Disp.	COMPAS
SP	0.096 (0.004)	0.097 (0.003)	0.102 (0.003)
BFPC	0.098 (0.002)	0.098 (0.003)	0.102 (0.003)
BFNC	0.094 (0.016)	0.093 (0.002)	0.102 (0.003)

D.2. Consumer Lending: Additional Data Details

Construction of IRSD for SA4 Regions As discussed in § 7, we focus our analysis on predictive disparities across SA4 geographic regions within Australia. We use the Australian Bureau of Statistics' Index of Relative Socioeconomic Disadvantage (IRSD) to define socioeconomically disadvantaged SA4 regions. The IRSD is calculated for SA2 regions, which are more granular statistical areas used by the ABS, by aggregating sixteen variables that were collected in the 2016 Australian census. These variables include, for example, the fraction of households making less than AU$26,000, the fraction of households with no internet access, and the fraction of residents who do not speak English well. Higher scores on the IRSD are associated with less socioeconomically disadvantaged regions, and conversely, lower scores on the IRSD are associated with more socioeconomically disadvantaged regions. The full list of variables that are included in the IRSD and complete details on how the IRSD is constructed is provided in Australian Bureau of Statistics (2016).

Because the IRSD is constructed for SA2 regions, we first aggregate this index to SA4 regions. We construct an aggregated IRSD for each SA4 region by constructing a population-weighted average of the IRSD for all SA2 regions that fall within each SA4 region. This delivers a quantitative measure of which SA4 regions are the most and least socioeconomically disadvantaged. For example, the bottom ventile (i.e., the 20th ventile) of SA4 regions based upon the population-weighted average IRSD (i.e., the least socioeconomically disadvantaged SA4 regions) are regions

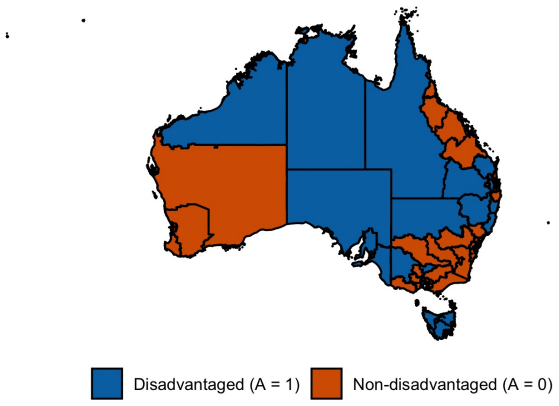


Figure 4. SA4 regions in Australia. We classify SA4 regions as being "socioeconomically disadvantaged" (red) and "non-socioeconomically disadvantaged" (blue) based on the Index of Relative Socioeconomic Disadvantage (IRSD).

associated with Sydney and Perth. The top ventile (i.e., the 1st ventile) of SA4 regions based upon the population-weighted average IRSD (i.e., the most socioeconomically disadvantaged SA4 regions) are regions associated with the Australian outback such as the Northern territory outback and the Southern Australia outback. Figure 4 provides a map of SA4 regions in Australia, in which colors SA4 regions classified as socioeconomically disadvantaged in blue.

D.3. Consumer Lending: Additional Experimental Details

We performed experiments on a random 2% sample of over 360,000 loan applications submitted from July 2017 to July 2019 by customers who did not have a prior financial relationship with CommBank, yielding our experimental sample of 7414 applications. We did a 2:1 train-test split, resulting in 4906 applications in our training set and 2508 applications in our test set.

In order to evaluate our methods on the full population (including applications that are not funded), we generate synthetic funding decisions D_i and outcomes $\tilde{Y}_i^*$ from the observed application features. On a 20% sample of the full 360,000 applicants, we train a classifier $\pi(x)$ to predict the observed funding decision D_i using the application features X_i, and we train a classifier $\mu(x)$ on funded applicants to predict the observed default outcome Y_i using the application features X_i. In other words, $\pi(x)$ estimates $P(D_i = 1|X_i = x)$ and $\mu(x)$ estimates $P(Y_i = 1|D_i = 1, X_i = x)$. For both models we use probability forests from the R package `ranger` with the default hyperparameters: 500 trees, $mtry = \sqrt{[dim(X)]} = 6$, min node size equal 10, and max depth equal to 0. To learn μ, we use bootstrap sampling of the $(0, 1)$ classes with probabilities $(0.01, 1)$, respectively, in order to down-sample the applicants who repaid the loans because we have significant class imbalance: Only 2.0% of applicants have default outcomes $= 1$.

We generate synthetic funding decisions $\tilde{D}_i$ according to $\widetilde{D}_i \mid X_i \sim Bernoulli(\pi(X_i))$ and synthetic default outcomes $\widetilde{Y}_i^*$ according to $\widetilde{Y}_i^* \mid X_i \sim Bernoulli(\mu(X_i))$. We then proceed with our learning as if we only had access to labels $\widetilde{Y}_i^*$ for applicants with $\widetilde{D}_i = 1$. We estimate $\hat{\mu}(x) := \hat{P}(\widetilde{Y}_i = 1|X_i = x, \widetilde{D}_i = 1)$ using random forests with the same hyperparameters as above and use $\hat{\mu}(x)$ to construct the pseudo-outcomes used by the IE and RIE approaches. The KGB, IE, and RIE approaches use linear regression. Our FaiRS algorithm ran the exponentiated gradient algorithm for at most 500 iterations on a fixed discretization grid, $\mathcal{Z}_\alpha = \{1/40, 2/40, \ldots, 1\}$ with parameters $B = \sqrt{n}$ and $\nu = 1/\sqrt{n}$ and $\eta = 2$. These choices were guided by our theoretical results as well as prior work (Agarwal et al., 2019). The average runtime for a single error tolerance ϵ was 26.4 minutes. The experiments were conducted on a machine with one Intel Xeon E5-2650 v2 processor with 2.60 GHz and 16 cores.

Our comparison against prior work used the fairlearn[8] API with logistic regression, using parameters parameters $C = 10$ and maximum iterations $= 10,000$. We ran fairlearn using both grid search and exponentiated gradient algorithm, but we report only the grid search algorithm since it traced out a larger fairness-performance tradeoff curve than the exponentiated gradient algorithm. We used a grid size of 41 with a grid limit of 2.

We also compared against the Target-Fair Covariate Shift method in Coston et al. (2019). To construct our covariate shift weights, we first estimated the propensity scores $P(D = 1 \mid X = x)$ by regressing $D \sim X$, yielding propensity estimates $\hat{\pi}(x)$. Our propensity model used `ranger` probability forests that with the default hyperparameters: 500 trees, $mtry = \sqrt{[dim(X)]} = 6$, min node size equal 10, and max depth equal to 0. We used $\max(\hat{\pi}(X), 50)$ as covariate shift weights. We ran the method for $\lambda = \{0, 10, 1000, 20000, 50000\}$ using step size $\eta = 0.01$. We terminated the algorithm when the L1 distance in the weight vector $\leq 1e - 7$ or after 500 iterations (whichever came first).

D.4. Consumer Lending Risk Scores: Additional Results

Figure 5 provides an extended version of Figure 1 that reports models over a range of hyperparameters (e.g. loss tolerance for FaiRS) to show the range of possible fairness-performance combinations.

[8] See Fairlearn Github for code.

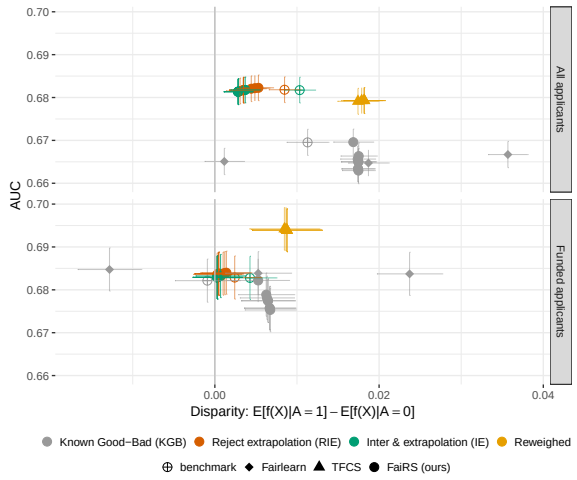


Figure 5. Area under the ROC curve (AUC) with respect to the synthetic outcome against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs advantaged ($A_i = 0$) groups. FaiRS reduces disparities for the RIE and IE approaches while maintaining AUCs comparable to the benchmark models (first row). Evaluation on only funded applicants (second row) overestimates the performance of TFCS and KGB models and underestimates disparities for all models. Error bars show the 95% confidence intervals. See § 7 of the main paper for details.

We next consider the implications of FaiRS for the credit applicants from the sensitive group. One might hope that encouraging statistical parity will increase access to credit for the sensitive group, and indeed we see some evidence for this. Figure 6 shows the distribution of risk scores for the disadvantaged group for the KGB, RIE, and IE models for the benchmark models (first row) and for FaiRS with 1% loss tolerance (second row). The 75% percentile score is given as a dashed line. FaiRS shifts the 75% percentile KGB score to the left (left column). FaiRS therefore reduces the predicted risk of the sensitive group, thereby expanding access to credit. We see a smaller shift for the RIE and IE approaches, which have lower risk distributions than the KGB model. As we have seen elsewhere, evaluation on the funded only applicants (right column) lends misleading conclusions, e.g., underestimating both the difference in distributions between the KGB and RIE/IE approaches as well as differences between the benchmark and FaiRS variants.

We present results for the benchmarks and FaiRS models with respect to the loss they were trained to minimize, mean-squared error. Figure 7 shows the mean square error (MSE) against predictive disparity for the KGB, RIE, IE benchmarks and FaiRS variants on held-out test data. The qualitative patterns are the same as Figure 1 in § 7 of the main text. Evaluation on all applicants shows that FaiRS with reject extrapolation (RIE and IE) reduces disparities without impacting MSE. The RIE and IE methods achieve lower disparity and lower MSE than the KGB model trained only on funded data, highlighting the importance of adjusting for selective labels. We again observe that evaluation on only funded applications is misleading as it suggests that the KGB models have comparable MSE and it drastically underestimates predictive disparities for all models.

Figure 1 in § 7 of the main text and Figure 7 shows that the FaiRS KGB model appears to produce larger predictive disparities than the benchmark KGB model. This is likely due to generalization error on the held-out test data. To verify this hypothesis, Figure 8 shows the MSE against predictive disparity for the KGB, RIE, IE benchmarks and FaiRS variants on the training data. Indeed among funded applicants in the train data, FaiRS-KGB models produce smaller absolute predictive disparities than the benchmark KGB model (second row).

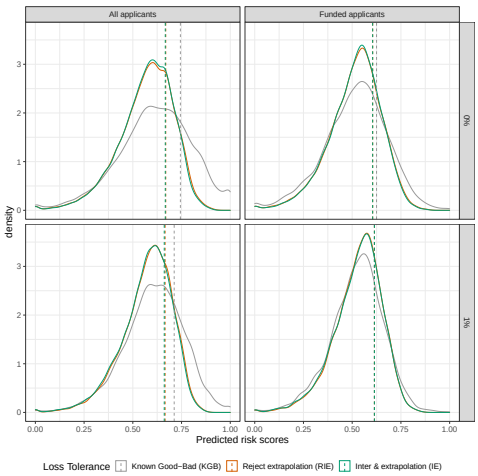


Figure 6. Predicted risk distributions for disadvantaged group $A_i = 1$ for FaiRS algorithm using KGB, RIE and IE approaches. The first row shows the benchmark model risk scores. The second row shows our FaiRS's risk scores for a loss tolerance of 1%. The left and right columns show risk scores on all applicants and funded applicants from the disadvantaged group respectively. The dashed line indicates the 75-percentile score. The RIE and IE methods predict lower rates of default for the disadvantaged group than the KGB method. The densities for the funded applicants (right column) underestimate the differences in risk scores across the KGB, RIE, and IE methods (compare to left column). See § 7 for details.

Another approach to expanding credit access for applicants from geographically disadvantaged regions is to target a prediction model that reduces score disparities among those applicants who would repay the loan if approved. This notion is related to balance for the positive/negative class (See Def. 2). To avoid confusion that may result from the definition of the positive class as those having the adverse outcome (e.g., default) in a risk assessment context, we will use *would-default class* to denote those who would have

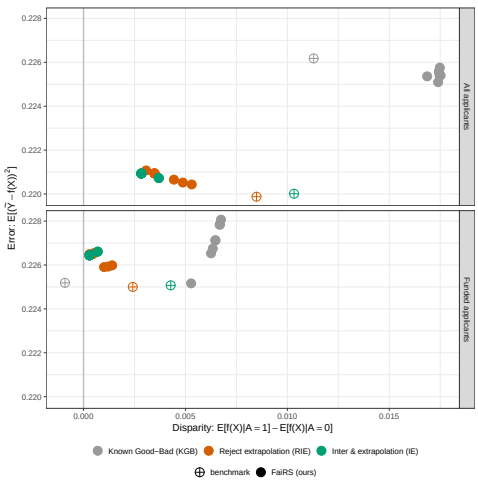


Figure 7. Mean square error (MSE) with respect to the synthetic outcome $\tilde{Y}_i$ against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs. advantaged ($A_i = 0$) groups in held-out test data. The first row evaluates each method on all applicants and and the second row evaluates each method on funded applicants only. See § 7 and § D.4 for details.

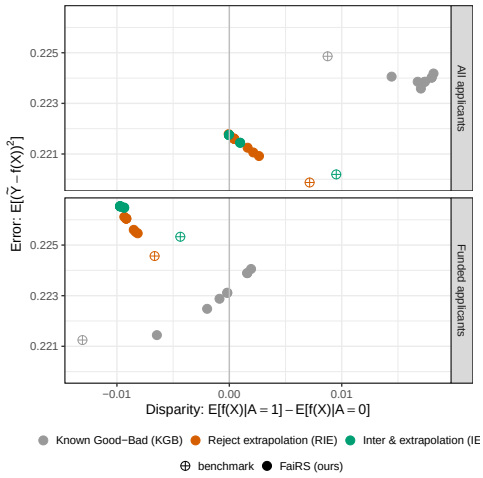


Figure 8. Mean square error (MSE) with respect to the synthetic outcome $\tilde{Y}_i$ against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs. advantaged ($A_i = 0$) groups in the training data. The first row evaluates each method on all applicants and and the second row evaluates each method on funded applicants only. See § 7 and § D.4 for details.

defaulted had the loan been funded and *would-repay class* to denote those who would have repaid had the loan been funded. Balance for the would-repay class is related to the notion of equality of opportunity (Hardt et al., 2016), but unlike standard applications of equality of opportunity, our target notion explicitly address selective labels. We compare our method to fairlearn models learned using true positive rate and false positive rate parity. We do not present the TFCS models as we did for statistical parity in the main paper because TFCS does not offer a method for balance parities. In certain settings where affirmative action-type policies are desirable, instead of targeting balance in the classes, the decision-maker may instead like to know whether there exists in the set of good models a prediction model for which the average score among those who would repay if funded is lower for the disadvantaged class versus the privileged class. To answer this question, we focus on characterizing the relative disparities (Problem 2).

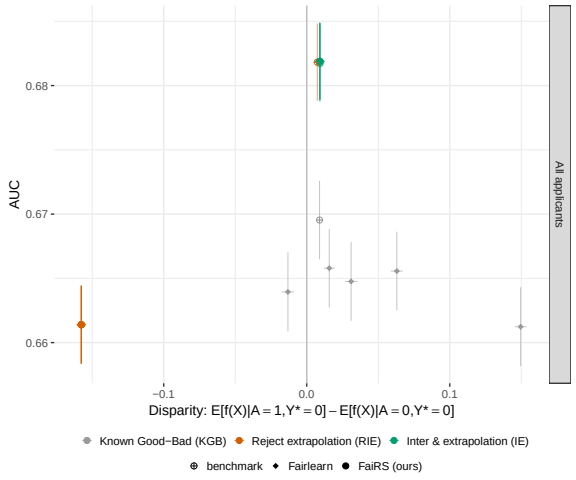


Figure 9. Area under the ROC curve (AUC) with respect to the synthetic outcome against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs advantaged ($A_i = 0$) groups among those who would repay the loan if funded ($Y^* = 0$). The benchmark IE, RIE, and FaiRS with IE models achieve the highest AUC. They also yield low disparities, but the disparities still favor the advantaged $A = 0$ applicants. FaiRS with RIE yields a model that favors the disadvantaged $A = 1$ applicants, but this model must sacrifice performance in order to do so. Error bars show the 95% confidence intervals. See § D.4 for details.

Figure 9 presents the results for disparities on the would-repay class. We present the AUC and disparities for the KGB, RIE, and IE benchmarks and their FaiRS variants as well as the fairlearn models. Our IE approach yields a prediction model with low disparities and a higher AUC than the fairlearn or other KGB approaches that do not adjust for selective labels. Using FaiRS with RIE, it is possible to achieve an affirmative-action type prediction model that has a substantially lower average score among the would-repay

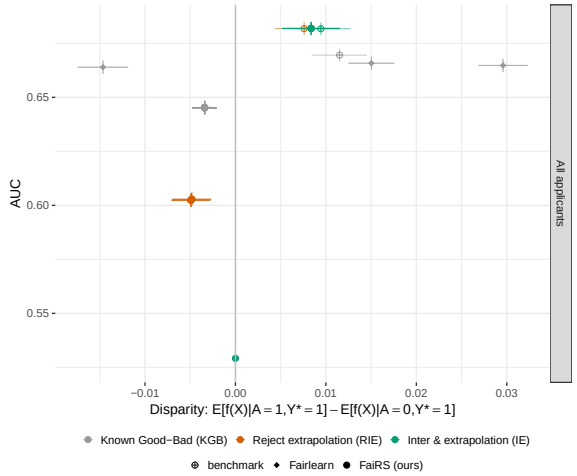


Figure 10. Area under the ROC curve (AUC) with respect to the synthetic outcome against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs advantaged ($A_i = 0$) groups among those who would default if funded ($Y^* = 1$). The benchmark IE and RIE models as well as a FaiRS with IE model achieve the highest AUC with low disparities that slightly favor the advantaged $A = 0$ applicants. A fairlearn model as well as FaiRS with RIE and with KGB yield predictions that favor the disadvantaged $A = 1$ applicants, but these models have lower AUC. Increasing the error allowance for FaiRS with IE yields a constant model that achieves a disparity of zero and a low AUC. Error bars show the 95% confidence intervals. See § D.4 for details.

for those with $A = 1$ versus $A = 0$. The AUC of this model is lower than the benchmark RIE model but is comparable to the AUCs of fairlearn and FaiRS with KGB.

Figure 10 presents the results for disparities on the would-default class. The results are similar for the would-repay class except here it is a fairlearn model that most favors the disadvantaged class, albeit at a lower AUC than that achieved by the benchmark or FaiRS with IE models.

D.5. Regression Experiments: Communities & Crime Dataset

The Communities & Crime dataset (Dua & Graff, 2017) contains 1,994 examples. We randomly split this data 50%-50% into a train and test set. We train models to predict the violent crime rate within each community (the number of violent crimes per 100,000 people), which is a continuous outcome. We evaluate models using least squares loss, define the benchmark model to be the loss-minimizing linear regression and focus on the statistical parity measure of predictive disparities between communities that are majority white vs. majority non-white. We use FaiRS to search for the predictive disparity minimizing linear regression that achieves a loss that is comparable to the benchmark (loss tolerance $\epsilon = 1\%, 5\%, 10\%$ of the loss-minimizing linear regression). In this dataset, there is no selective labels problem, so we construct the FaiRS model following approach detailed in § 4 and Supplement § A.3.

Table 6. The FaiRS models over the set of good models (performing within 1%, 5%, and 10% of the loss-minimizing linear regression's training loss) achieve comparable performance to the test loss of the loss-minimizing linear regression and produce lower absolute predictive disparities. The first column reports the disparity in average predictions between majority white and majority non-white communities (Def. 1). The second column reports the test losses for each model. Standard errors are reported in parentheses. See § D.5 for details.

	LOSS	DISP.
BENCHMARK	0.0101	-0.3386
	(0.0007)	(0.0135)
FAIRS		
$\epsilon = 1\%$	0.0103	-0.2989
	(0.0008)	(0.0130)
$\epsilon = 5\%$	0.0105	-0.2856
	(0.0008)	(0.0129)
$\epsilon = 10\%$	0.0108	-0.2658
	(0.0008)	(0.0127)

Table 6 summarizes both the predictive disparities and least squares losses over the test set of the $FaiRS$ models and the benchmark linear regression. The FaiRS models achieve comparable performance to the test loss of the benchmark loss-minimizing linear regression while producing lower predictive disparities. These results highlight that our proposed methods perform as desired in a regression setting.