A. Other related work

Fair classification. Many works have focused on formulating fair classification problems as constrained optimization problems, (Zafar et al., 2017b; Zhang et al., 2018; Menon & Williamson, 2018b; Goel et al., 2018; Celis et al., 2019), (Hardt et al., 2016; Zafar et al., 2017a; Menon & Williamson, 2018b; Celis et al., 2019), and developing algorithms for it. Another class of algorithms first learn an unconstrained optimal classifier and then shift the decision boundary according to the fairness requirement, e.g., (Fish et al., 2016; Hardt et al., 2016; Goh et al., 2016; Pleiss et al., 2017; Woodworth et al., 2017; Dwork et al., 2018). In contrast to our work, the assumption in all of these approaches is that the algorithm is given perfect information about the protected class.

Data correction. Cleaning raw data is a significant step in the pipeline, and efforts to correct for missing or inaccurately coded attributes have been studied in-depth for protected attributes, e.g., in the context of the census (Nobles, 2000). An alternate approach considers changing the composition of the dataset itself to correct for known biases in representation (Calders et al., 2009; Kamiran & Calders, 2009; 2012), (Gordaliza et al., 2019; Wang et al., 2019), (Calmon et al., 2017; Celis et al., 2020). In either case, the correction process, while important, can be imperfect and our work can help by starting with these improved yet imperfect datasets in order to build fair classifiers.

Unknown protected attributes. A related setting is when the information of some protected attributes is unknown. (Gupta et al., 2018; Chen et al., 2019; Kallus et al., 2020; Lahoti et al., 2020) considered this setting of unknown protected attributes and designed algorithms to improve fairness or assess disparity. In contrast, our approach aims to derive necessary information from the observed protected attributes to design alternate fairness constraints using the noisy attribute.

Classifiers robust to the choice of datasets. (Friedler et al., 2019) observed that fair classification algorithms may not be stable with respect to variations in the training dataset. (Hashimoto et al., 2018) proved that empirical risk minimization amplifies representation disparity over time. Towards this, certain variance reduction or stability techniques have been introduced; see e.g., (Huang & Vishnoi, 2019). However, their approach cannot be used to learn a classifier that is provably fair over the underlying dataset.

Noise in labels. Blum & Stangl (2020); Biswas & Mukherjee (2021); Roh et al. (2020); Jiang & Nachum (2020) study fair classification when the label in the input dataset is noisy. The main difference of these from our work is that they consider noisy *labels* instead of noisy protected attributes, which makes our denoised algorithms very different since the accuracy of protected attributes mainly relates to the fairness of the classifier but the accuracy of labels primarily affect the empirical loss.

B. Missing proofs in Section 3.3

In this section, we complete the missing proofs in Section 3.3. Let $\pi_{ij} := \Pr_{D,\widehat{D}}\left[\widehat{Z} = i \mid Z = j\right]$ for $i, j \in \{0,1\}$, $\mu_i := \Pr_D[Z = i]$ and $\widehat{\mu}_i := \Pr_{\widehat{D}}\left[\widehat{Z} = i\right]$ for $i \in \{0,1\}$.

B.1. Proof of Lemma 3.6

Proof: We first have the following simple observation.

Observation B.1 *1)* $\mu_0 + \mu_1 = 1$, $\widehat{\mu}_0 + \widehat{\mu}_1 = 1$, *and* $\pi_{0,i} + \pi_{1,i} = 1$ *holds for* $i \in \{0,1\}$*; 2) For any* $i, j \in \{0,1\}$, $\Pr\left[Z = i \mid \widehat{Z} = j\right] = \frac{\pi_{ji}\cdot\mu_i}{\widehat{\mu}_j}$*; 3) For any* $i \in \{0,1\}$, $\widehat{\mu}_i = \pi_{i,i}\mu_i + \pi_{i,1-i}\mu_{1-i}$.

Similar to Equation 36, we have

$$\begin{aligned} & \Pr\left[f = 1, \widehat{Z} = 0\right] \\ = \; & \Pr\left[\widehat{Z} = 0 \mid f = 1, Z = 0\right] \cdot \Pr\left[f = 1, Z = 0\right] \\ & + \Pr\left[\widehat{Z} = 0 \mid f = 1, Z = 1\right] \cdot \Pr\left[f = 1, Z = 1\right]. \end{aligned} \tag{3}$$

Similar to the proof of Lemma F.5, by the Chernoff bound (additive form) (Hoeffding, 1994), both

$$\Pr\left[\widehat{Z} = 1 \mid f = 1, Z = 0\right] \in \eta_0 \pm \frac{\varepsilon}{2\Pr\left[f = 1, Z = 0\right]}, \tag{4}$$

and

$$\Pr\left[\widehat{Z} = 0 \mid f = 1, Z = 1\right] \in \eta_1 \pm \frac{\varepsilon}{2\Pr\left[f = 1, Z = 1\right]}, \tag{5}$$

hold with probability at least

$$1 - 2e^{-\frac{\varepsilon^2 n}{12\eta\Pr[f=1,Z=0]}} - 2e^{-\frac{\varepsilon^2 n}{12\eta\Pr[f=1,Z=1]}}$$

which for $\eta \le 0.5$, is at least $1 - 2e^{-\varepsilon^2 n/6}$. Consequently,

we have

$$\begin{aligned}
& \Pr\left[f=1,\widehat{Z}=0\right] \\
= \;& \Pr\left[\widehat{Z}=0 \mid f=1,Z=0\right]\cdot\Pr\left[f=1,Z=0\right] \\
& +\Pr\left[\widehat{Z}=0 \mid f=1,Z=1\right]\cdot\Pr\left[f=1,Z=1\right] \\
& \text{(Eq. 3)} \\
\in \;& \left(1-\eta_0\pm\frac{\varepsilon}{2\Pr\left[f=1,Z=0\right]}\right)\cdot\Pr\left[f=1,Z=0\right] \\
& +\left(\eta_1\pm\frac{\varepsilon}{2\Pr\left[f=1,Z=1\right]}\right)\cdot\Pr\left[f=1,Z=1\right] \\
& \text{(Ineqs. 4 and 5)} \\
\in \;& (1-\eta_0)\Pr\left[f=1,Z=0\right] \\
& +\eta_1\Pr\left[f=1,Z=1\right]\pm\varepsilon,
\end{aligned} \tag{6}$$

and similarly,

$$\begin{aligned}
& \Pr\left[f=1,\widehat{Z}=1\right] \\
\in \;& \eta_0\Pr\left[f=1,Z=0\right] \\
& +(1-\eta_1)\Pr\left[f=1,Z=1\right]\pm\varepsilon.
\end{aligned} \tag{7}$$

By the above two inequalities, we conclude that

$$\begin{aligned}
& (1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right] \\
\in \;& (1-\eta_1)\big((1-\eta_0)\Pr\left[f=1,Z=0\right] \\
& +\eta_1\Pr\left[f=1,Z=1\right]\pm\varepsilon\big) \\
& -\eta_1\big(\eta_0\Pr\left[f=1,Z=0\right] \\
& +(1-\eta_1)\Pr\left[f=1,Z=1\right]\pm\varepsilon\big) \quad \text{(Ineqs. 6 and 7)} \\
\in \;& (1-\eta_0-\eta_1)\Pr\left[f=1,Z=0\right]\pm\varepsilon.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
& (1-\eta_0)\Pr\left[f=1,\widehat{Z}=1\right]-\eta_0\Pr\left[f=1,\widehat{Z}=0\right] \\
\in \;& (1-\eta_0-\eta_1)\Pr\left[f=1,Z=1\right]\pm\varepsilon.
\end{aligned}$$

This completes the proof of the first conclusion.

Next, we focus on the second conclusion. By assumption, $\min\left\{\Pr\left[f=1,Z=0\right],\Pr\left[f=1,Z=1\right]\right\}\geq\frac{\lambda}{2}$. Let $\varepsilon'=\frac{\varepsilon(1-\eta_0-\eta_1)\lambda}{20}$. By a similar argument as for the first conclusion, we have the following claim.

Claim B.2 *With probability at least* $1-4e^{-(\varepsilon')^2n/6}$*, we have*

$$\begin{cases}
& (1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right] \\
\in \;& (1-\eta_0-\eta_1)\Pr\left[f=1,Z=0\right]\pm\varepsilon', \\
& (1-\eta_0)\Pr\left[f=1,\widehat{Z}=1\right]-\eta_0\Pr\left[f=1,\widehat{Z}=0\right] \\
\in \;& (1-\eta_0-\eta_1)\Pr\left[f=1,Z=1\right]\pm\varepsilon', \\
& (1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1\in(1-\eta_0-\eta_1)\mu_0\pm\varepsilon', \\
& (1-\eta_0)\widehat{\mu}_1-\eta_0\widehat{\mu}_0\in(1-\eta_0-\eta_1)\mu_1\pm\varepsilon'.
\end{cases}$$

Now we assume Claim B.2 holds whose success probability is at least $1-4e^{-\frac{\varepsilon^2(1-\eta_0-\eta_1)^2\lambda^2n}{2400}}$ since $\varepsilon'=\frac{\varepsilon(1-\eta_0-\eta_1)\lambda}{20}$. Consequently, we have

$$\begin{aligned}
& (1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right] \\
\geq \;& (1-\eta_0-\eta_1)\Pr\left[f=1,Z=0\right]-\varepsilon' \quad \text{(Claim B.2)} \\
\geq \;& \frac{(1-\eta_0-\eta_1)\lambda}{2}-\varepsilon' \quad \text{(by assumption)} \\
\geq \;& 0.45\cdot(1-\eta_0-\eta_1)\lambda. \\
& (\varepsilon'=\frac{\varepsilon(1-\eta_0-\eta_1)\lambda}{20})
\end{aligned} \tag{8}$$

Similarly, we can also argue that

$$(1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1\geq 0.45\cdot(1-\eta_0-\eta_1)\lambda. \tag{9}$$

Then we have

$$\begin{aligned}
& \Pr\left[f=1\mid Z=0\right] \\
= \;& \frac{\Pr\left[f=1,Z=0\right]}{\mu_0} \\
\in \;& \frac{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]\pm\varepsilon'}{(1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1\pm\varepsilon'} \\
& \text{(Claim B.2)} \\
\in \;& \frac{\left((1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]\right)}{(1\pm\frac{\varepsilon'}{0.45\cdot(1-\eta_0-\eta_1)\lambda})\left((1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1\right)} \\
& \times(1\pm\frac{\varepsilon'}{0.45\cdot(1-\eta_0-\eta_1)\lambda}) \quad \text{(Ineq. 8)} \\
\in \;& (1\pm\frac{\varepsilon}{9})^2\cdot\Gamma_0(f). \quad \text{(Defns. of } \Gamma_0(f) \text{ and } \varepsilon')
\end{aligned}$$

Similarly, we can also prove that

$$\Pr\left[f=1\mid Z=1\right]\in(1\pm\frac{\varepsilon}{9})^2\cdot\Gamma_1(f).$$

By the above two inequalities, we have that with probability at least $1-4e^{-\frac{\varepsilon^2(1-\eta_0-\eta_1)^2\lambda^2n}{2400}}$,

$$\begin{aligned}
& \gamma^{\Delta}(f,S) \\
= \;& \min\left\{\frac{\Gamma_0(f)}{\Gamma_1(f)},\frac{\Gamma_1(f)}{\Gamma_0(f)}\right\} \\
\in \;& (1\pm\varepsilon)\times \\
& \min\left\{\frac{\Pr\left[f=1\mid Z=0\right]}{\Pr\left[f=1\mid Z=1\right]},\frac{\Pr\left[f=1\mid Z=1\right]}{\Pr\left[f=1\mid Z=0\right]}\right\} \\
\in \;& (1\pm\varepsilon)\cdot\gamma(f,S).
\end{aligned}$$

Combining with Claim B.2, we complete the proof of the second conclusion. □

B.2. Proof of Lemma 3.8

For preparation, we give the following definition.

Definition B.3 (**ε-nets**) *Given a family $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ of classifiers and $\varepsilon \in (0,1)$, we say $F \subseteq \mathcal{F}$ is an ε-net of $\mathcal{F}$ if for any $f, f' \in F$, $\Pr_D[f \neq f'] \geq \varepsilon$; and for any $f \in \mathcal{F}$, there exists $f' \in F$ such that $\Pr_D[f \neq f'] \leq \varepsilon$. We denote $M_\varepsilon(\mathcal{F})$ as the smallest size of an ε-net of $\mathcal{F}$.*

It follows from basic coding theory (Lint, 1998) that $M_\varepsilon(\{0,1\}^{\mathcal{X}}) = \Omega(2^{N - O(\varepsilon N \log N)})$. The size of an ε-net usually depends exponentially on the VC-dimension.

Theorem B.4 (Relation between VC-dimension and ε-nets (Haussler, 1995)) *Suppose the VC-dimension of $(S, \mathcal{F})$ is t. For any $\varepsilon \in (0,1)$, $M_\varepsilon(\mathcal{F}) = O(\varepsilon^{-t})$.*

We define the capacity of bad classifiers based on ε-nets.

Definition B.5 (Capacity of bad classifiers) *Let $\varepsilon_0 = \frac{(1-\eta_0-\eta_1)\lambda - 2\delta}{5}$. Let $\varepsilon_i = \frac{1.01^{2^{i-1}}\delta}{5}$ for $i \in [T]$ where $T = \lceil 232 \log\log \frac{2(\tau-3\delta)}{\lambda} \rceil$. Given $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, we denote the capacity of bad classifiers by $\Phi(\mathcal{F}) := 2e^{-\varepsilon_0^2 n/6} M_{\varepsilon_0}(\mathcal{G}_0) + 4\sum_{i\in[T]} e^{-\frac{\varepsilon_i^2(1-\eta_0-\eta_1)^2\lambda^2 n}{2400}} M_{\varepsilon_i(1-\eta_0-\eta_1)\lambda/10}(\mathcal{G}_i)$.*

Actually, we can prove $\Phi(\mathcal{F})$ is an upper bound for the probability that there exists a bad classifier that is feasible for Program **DFair**, which is a generalized version of Lemma 3.8. Roughly, the factor $2e^{-\varepsilon_0^2 n/6}$ is an upper bound of the probability that a bad classifier $f \in \mathcal{G}_0$ violates Constraint (2), and the factor $4e^{-\varepsilon_i^2\lambda^2\delta^2 n}$ is an upper bound of the probability that a bad classifier $f \in \mathcal{G}_i$ violates Constraint (2). We prove if all bad classifiers in the nets of $\mathcal{G}_i$ ($0 \leq i \leq T$) are not feasible for Program **DFair**, then all bad classifiers should violate Constraint (2). Note that the scale of $\Phi(\mathcal{F})$ depends on the size of ε-nets of $\mathcal{F}$, which can be upper bounded by Theorem B.4 and leads to the success probability of Theorem 3.3.

Proof: We first claim that Lemma 3.8 holds with probability at least $1 - \Phi(\mathcal{F})$. We discuss $\mathcal{G}_0$ and $\mathcal{G}_i$ ($i \in [T]$) separately.

Bad classifiers in $\mathcal{G}_0$. Let G_0 be an ε_0-net of $\mathcal{G}_0$ of size $M_{\varepsilon_0}(\mathcal{G}_0)$. Consider an arbitrary classifier $g \in G_0$. By Lemma 3.6, with probability at least $1 - 2e^{-\varepsilon_0^2 n/6}$, we have

$$\begin{aligned} & (1-\eta_1)\Pr\left[g=1,\widehat{Z}=0\right] - \eta_1 \Pr\left[g=1,\widehat{Z}=1\right] \\ \leq\ & (1-\eta_0-\eta_1)\Pr[g=1,Z=0] + \varepsilon_0 \\ <\ & \frac{(1-\eta_0-\eta_1)\lambda}{2} + \varepsilon_0, \quad \text{(Defn. of } \mathcal{G}_0\text{)} \end{aligned} \tag{10}$$

and

$$\begin{aligned} & (1-\eta_0)\Pr\left[g=1,\widehat{Z}=1\right] - \eta_0 \Pr\left[g=1,\widehat{Z}=0\right] \\ <\ & \frac{(1-\eta_0-\eta_1)\lambda}{2} + \varepsilon_0. \end{aligned} \tag{11}$$

By the union bound, all classifiers $g \in G_0$ satisfy Inequalities 10 and 11 with probability at least $1 - 2e^{-\varepsilon_0^2 n/6} M_{\varepsilon_0}(\mathcal{G}_0)$. Suppose this event happens. We consider an arbitry classifier $f \in \mathcal{G}_0$. W.l.o.g., we assume $\Pr[f=1, Z=0] < \frac{\lambda}{2}$. By Definition B.3, there must exist a classifier $g \in G_0$ such that $\Pr[f \neq g] \leq \varepsilon_0$. Then we have

$$\begin{aligned} & (1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right] - \eta_1 \Pr\left[f=1,\widehat{Z}=1\right] \\ \leq\ & (1-\eta_1)(\Pr\left[g=1,\widehat{Z}=0\right] + \varepsilon_0) \\ & - \eta_1(\Pr\left[g=1,\widehat{Z}=1\right] - \varepsilon_0) \quad (\Pr[f\neq g] \leq \varepsilon_0) \\ \leq\ & \frac{(1-\eta_0-\eta_1)\lambda}{2} + 2\varepsilon_0 \quad \text{(Ineq. 10)} \\ \leq\ & \frac{(1-\eta_0-\eta_1)\lambda}{2} + \frac{(1-\eta_0-\eta_1)\lambda - 2\delta}{2} \\ & \text{(Defn. of } \varepsilon_0\text{)} \\ =\ & (1-\eta_0-\eta_1)\lambda - \delta, \end{aligned}$$

Thus, we conclude that all classifiers $f \in \mathcal{G}_0$ violate Constraint 2 with probability at least $1 - 2e^{-\varepsilon_0^2 n/6} M_{\varepsilon_0}(\mathcal{G}_0)$.

Bad classifiers in $\mathcal{G}_i$ for $i \in [T]$. We can assume that $\tau - 3\delta \geq \lambda/2$. Otherwise, all $\mathcal{G}_i$ for $i \in [T]$ are empty, and hence, we complete the proof. Consider an arbitry $i \in [T]$ and let G_i be an ε_i-net of $\mathcal{G}_i$ of size $M_{\varepsilon_i(1-\eta_0-\eta_1)\lambda/10}(\mathcal{G}_i)$. Consider an arbitrary classifier $g \in G_i$. By the proof of Lemma 3.6, with probability at least $1 - 4e^{-\frac{\varepsilon_i^2(1-\eta_0-\eta_1)^2\lambda^2 n}{2400}}$, we have

$$\begin{cases} (1-\eta_1)\Pr\left[g=1,\widehat{Z}=0\right] - \eta_1 \Pr\left[g=1,\widehat{Z}=1\right] \\ \in (1-\eta_0-\eta_1)\Pr[g=1,Z=0] \pm \frac{\varepsilon_i(1-\eta_0-\eta_1)\lambda}{20}, \\ (1-\eta_0)\Pr\left[g=1,\widehat{Z}=1\right] - \eta_0 \Pr\left[g=1,\widehat{Z}=0\right] \\ \in (1-\eta_0-\eta_1)\Pr[g=1,Z=1] \pm \frac{\varepsilon_i(1-\eta_0-\eta_1)\lambda}{20}, \\ \gamma^{\Delta}(f,\widehat{S}) \in (1\pm\varepsilon_i)\cdot\gamma(f,S). \end{cases} \tag{12}$$

Moreover, we have

$$\begin{aligned} \gamma^{\Delta}(g,\widehat{S}) \leq\ & (1+\varepsilon_i)\cdot\gamma(g,S) \\ <\ & (1+\varepsilon_i)\cdot\frac{\tau-3\delta}{1.01^{2^{i-1}}}. \quad \text{(Defn. of } \mathcal{G}_i\text{)} \end{aligned} \tag{13}$$

By the union bound, all classifiers $g \in G_i$ satisfy Inequality 13 with probability at least

$$1 - 4e^{-\frac{\varepsilon_i^2(1-\eta_0-\eta_1)^2\lambda^2 n}{2400}} M_{\varepsilon_i(1-\eta_0-\eta_1)\lambda/10}(\mathcal{G}_i).$$

Suppose this event happens. We consider an arbitry classifier $f \in \mathcal{G}_i$. By Definition B.3, there must exist a classifier $g \in G_i$ such that $\Pr[f \neq g] \leq \varepsilon_i(1-\eta_0-\eta_1)\lambda/10$. By Inequality 12 and a similar argument as that for Inequality 8, we have

$$\begin{aligned} & (1-\eta_1)\Pr\left[g=1,\widehat{Z}=0\right] - \eta_1 \Pr\left[g=1,\widehat{Z}=1\right] \\ \geq\ & 0.45\cdot(1-\eta_0-\eta_1)\lambda. \end{aligned} \tag{14}$$

$$
\begin{aligned}
&\Gamma_0(f)\\
=&\frac{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]}{(1-\eta)\widehat{\mu}_0-\eta\widehat{\mu}_1}\\
\in&\frac{(1-\eta_1)\left(\Pr\left[g=1,\widehat{Z}=0\right]\pm\frac{\varepsilon_i(1-\eta_0-\eta_1)\lambda}{10}\right)}{(1-\eta)\widehat{\mu}_0-\eta\widehat{\mu}_1}\\
&-\frac{\eta_1\left(\Pr\left[g=1,\widehat{Z}=1\right]\pm\frac{\varepsilon_i(1-\eta_0-\eta_1)\lambda}{10}\right)}{(1-\eta)\widehat{\mu}_0-\eta\widehat{\mu}_1}\\
&(\Pr[f\neq g]\leq\varepsilon_i(1-\eta_0-\eta_1)\lambda/10)\\
\in&\frac{(1-\eta_1)\Pr\left[g=1,\widehat{Z}=0\right]-\eta_1\Pr\left[g=1,\widehat{Z}=1\right]}{(1-\eta)\widehat{\mu}_0-\eta\widehat{\mu}_1}\\
&\pm\frac{\frac{\varepsilon_i(1-\eta_1-\eta_1)\lambda}{5}}{(1-\eta)\widehat{\mu}_0-\eta\widehat{\mu}_1}\\
\in&\frac{(1-\eta)\Pr\left[g=1,\widehat{Z}=0\right]-\eta\Pr\left[g=1,\widehat{Z}=1\right]}{(1-\eta)\widehat{\mu}_0-\eta\widehat{\mu}_1}\times\\
&(1\pm0.45\varepsilon_i)\quad\text{(Ineq. 14)}\\
\in&(1\pm0.45\varepsilon_i)\cdot\Gamma_0(g).
\end{aligned}
\tag{15}
$$

Similarly, we can also prove

$$
\Gamma_1(f)\in(1\pm0.45\varepsilon_i)\cdot\Gamma_1(g). \tag{16}
$$

Thus, we conclude that

$$
\begin{aligned}
\gamma^{\Delta}(f,\widehat{S})=&\quad\min\left\{\frac{\Gamma_0(f)}{\Gamma_1(f)},\frac{\Gamma_1(f)}{\Gamma_0(f)}\right\}\\
\leq&\quad\frac{1+0.45\varepsilon_i}{1-0.45\varepsilon_i}\cdot\min\left\{\frac{\Gamma_0(g)}{\Gamma_1(g)},\frac{\Gamma_1(g)}{\Gamma_0(g)}\right\}\\
&\text{(Ineqs. 15 and 16)}\\
<&\quad\frac{1+0.45\varepsilon_i}{1-0.45\varepsilon_i}\cdot(1+\varepsilon_i)\cdot\frac{\tau-3\delta}{1.01^{2^i-1}}\\
&\text{(Ineq. 13)}\\
\leq&\quad\frac{1+0.45\varepsilon_1}{1-0.45\varepsilon_1}\cdot(1+\varepsilon_1)\cdot(\tau-3\delta)\quad\text{(Defn. of }\varepsilon_i)\\
\leq&\quad\tau-\delta.\quad(\varepsilon_1=\frac{1.01\delta}{5})
\end{aligned}
$$

It implies that all classifiers $f\in\mathcal{G}_i$ violate Constraint 2 with probability at least

$$
1-4e^{-\frac{\varepsilon_i^2(1-\eta_0-\eta_1)^2\lambda^2 n}{2400}}M_{\varepsilon_i(1-\eta_0-\eta_1)\lambda/10}(\mathcal{G}_i).
$$

By the union bound, we complete the proof of Lemma 3.8 for $\delta\in(0,0.1\lambda)$.

For general $\delta\in(0,1)$, each bad classifier violates Constraint 2 with probability at most $4e^{-\frac{\varepsilon_1^2(1-\eta_0-\eta_1)^2\lambda^2 n}{2400}}$ by the above argument. By Definition 3.7, $|M_{\varepsilon_0}(\mathcal{G}_0)|+\sum_{i\in[T]}|M_{\varepsilon_i(1-\eta_0-\eta_1)\lambda/10}(\mathcal{G}_i)|\leq|M_{\varepsilon_1(1-\eta_0-\eta_1)\lambda/10}(\mathcal{F})|$. Then by the definition of $\Phi(\mathcal{F})$ and Theorem B.4, the probability that there exists a bad classifier violating Constraint 2 is at most $\Phi(\mathcal{F})=O\left(e^{-\frac{(1-\eta_0-\eta_1)^2\lambda^2\delta^2 n}{60000}+t\ln(\frac{50}{(1-\eta_0-\eta_1)\lambda\delta})}\right)$. This completes the proof of Lemma 3.8. □

B.3. Proof of Theorem 3.3 for $p=2$ and statistical rate

Proof: We first upper bound the probability that $\gamma^{\Delta}(f^{\Delta},\widehat{S})\geq\tau-3\delta$. Let $\mathcal{F}_b=\{f\in\mathcal{F}:\gamma(f,S)<\tau-3\delta\}$. If all classifiers in $\mathcal{F}_b$ violate Constraint (2), we have that $\gamma^{\Delta}(f^{\Delta},\widehat{S})\geq\tau-3\delta$. Note that if $\min_{i\in\{0,1\}}\Pr[f=1,Z=i]\geq\frac{\lambda}{2}$, then $\gamma(f,S)\geq\frac{\lambda}{2}$ holds by definition. Also, $\frac{\lambda-3\delta}{1.01^{2^{T+1}-1}}\leq\frac{\lambda}{2}$. Thus, we conclude that $\mathcal{F}_b\subseteq\cup_{i=0}^{T}\mathcal{G}_i$. Then if all bad classifiers violate Constraint (2), we have $\gamma^{\Delta}(f^{\Delta},\widehat{S})\geq\tau-3\delta$. By Lemma 3.8, $\gamma^{\Delta}(f^{\Delta},\widehat{S})\geq\tau-3\delta$ holds with probability at least $1-O\left(e^{-\frac{(1-\eta_0-\eta_1)^2\lambda^2\delta^2 n}{60000}+t\ln(\frac{50}{(1-\eta_0-\eta_1)\lambda\delta})}\right)$.

Next, we upper bound the probability that $f^\star$ is feasible for Program **DFair**, which implies $\frac{1}{N}\sum_{a\in[N]}L(f^{\Delta},s_a)\leq\frac{1}{N}\sum_{a\in[N]}L(f^\star,s_a)$. Letting $\varepsilon=\delta$ in Lemma 3.6, we have that with probability at least $1-2e^{-\delta^2 n/6}-4e^{-\frac{(1-\eta_0-\eta_1)^2\lambda^2\delta^2 n}{2400}}$,

$$
\begin{cases}
&(1-\eta)\Pr\left[f^\star=1,\widehat{Z}=0\right]-\eta\Pr\left[f^\star=1,\widehat{Z}=1\right]\\
&\geq(1-\eta_0-\eta_1)\Pr[f^\star=1,Z=0]-\delta,\\
&(1-\eta)\Pr\left[f^\star=1,\widehat{Z}=1\right]-\eta\Pr\left[f^\star=1,\widehat{Z}=0\right]\\
&\geq(1-\eta_0-\eta_1)\Pr[f^\star=1,Z=1]-\delta,\\
&\gamma^{\Delta}(f^\star,\widehat{S})\geq(1-\delta)\gamma(f,S)\geq\gamma(f,S)-\delta.
\end{cases}
$$

It implies that $f^\star$ is feasible for Program **DFair** with probability at least $1-2e^{-\delta^2 n/6}-4e^{-\frac{(1-\eta_0-\eta_1)^2\lambda^2\delta^2 n}{2400}}$. This completes the proof. □

C. Analysis of the influences of estimation errors

We discuss the influences of estimation errors by considering a simple setting as in Section 3.3, say $p=2$ with statistical rate. Recall that we assume η_0 and η_1 are given in Theorem 3.3. However, we may only have estimations for η_0 and η_1 in practice, say η_0' and η_1' respectively. Define $\zeta:=\max\{|\eta_0-\eta_0'|,|\eta_1-\eta_1'|\}$ to be the additive estimation error. We want to understand the influences of ζ on the performance of our denoised program.

Since η_0 and η_1 are unknown now, we can not directly compute $\Gamma_0(f)$ and $\Gamma_1(f)$ in Definition 3.1. Instead, we can compute

$$
\Gamma_0'(f):=\frac{(1-\eta_1')\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1'\Pr\left[f=1,\widehat{Z}=1\right]}{(1-\eta_1')\widehat{\mu}_0-\eta_1'\widehat{\mu}_1},
$$

$$
\Gamma_1(f):=
$$

$$\frac{(1-\eta_0')\Pr\left[f=1,\widehat{Z}=1\right]-\eta_0'\Pr\left[f=1,\widehat{Z}=0\right]}{(1-\eta_0')\widehat{\mu}_1-\eta_0'\widehat{\mu}_0}.$$

Then we have

$$\begin{aligned}
&\Gamma_0'(f)\\
=\;&\frac{(1-\eta_1')\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1'\Pr\left[f=1,\widehat{Z}=1\right]}{(1-\eta_1')\widehat{\mu}_0-\eta_1'\widehat{\mu}_1}\\
=\;&\frac{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]}{(1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1+(\eta_1-\eta_1')}\\
&+\frac{(\eta_1-\eta_1')\Pr\left[f=1\right]}{(1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1+(\eta_1-\eta_1')} \qquad (17)\\
\in\;&\frac{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]}{(1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1}\\
&\pm\frac{\zeta\cdot\Pr\left[f=1\right]}{(1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1}\quad\text{(Defn. of }\zeta\text{)}\\
\in\;&\Gamma_0(f)\pm\frac{\zeta\cdot\Pr\left[f=1\right]}{(1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1}.\quad\text{(Defn. of }\Gamma_0(f)\text{)}
\end{aligned}$$

Symmetrically, we have

$$\Gamma_1'(f)\in\Gamma_1(f)\pm\frac{\zeta\cdot\Pr\left[f=1\right]}{(1-\eta_0)\widehat{\mu}_1-\eta_0\widehat{\mu}_0}. \qquad (18)$$

By a similar argument, we can also prove that

$$\begin{aligned}
&\frac{1}{\Gamma_0'(f)}\in\frac{1}{\Gamma_0(f)}\\
&\pm\frac{\zeta}{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]}. \qquad (19)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{\Gamma_1'(f)}\in\frac{1}{\Gamma_1(f)}\\
&\pm\frac{\zeta}{(1-\eta_0)\Pr\left[f=1,\widehat{Z}=1\right]-\eta_0\Pr\left[f=1,\widehat{Z}=0\right]}. \qquad (20)
\end{aligned}$$

Then by the denoised constraint on η_0' and η_1', i.e.,

$$\min\left\{\frac{\Gamma_1'(f)}{\Gamma_0'(f)},\frac{\Gamma_0'(f)}{\Gamma_1'(f)}\right\}\geq\tau-\delta, \qquad (21)$$

we conclude that

$$\begin{aligned}
&\frac{\Gamma_1(f)}{\Gamma_0(f)}\\
&\geq\left(\Gamma_1'(f)-\frac{\zeta\cdot\Pr\left[f=1\right]}{(1-\eta_0)\widehat{\mu}_1-\eta_0\widehat{\mu}_0}\right)\times\Big(\frac{1}{\Gamma_0'(f)}-\\
&\frac{\zeta}{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]}\Big)\\
&\text{(Ineqs. 18 and 19)}\\
&\geq\frac{\Gamma_1'(f)}{\Gamma_0'(f)}-\zeta\Big(\frac{\Pr\left[f=1\right]}{\Gamma_0'(f)\left((1-\eta_0)\widehat{\mu}_1-\eta_0\widehat{\mu}_0\right)}+\\
&\frac{\Gamma_1'(f)}{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]}\Big)\\
&\geq\tau-\delta-\zeta\alpha_1,\quad\text{(Ineqs. 21)}
\end{aligned}$$

where $\alpha_1=\frac{\Pr[f=1]}{\Gamma_0'(f)((1-\eta_0)\widehat{\mu}_1-\eta_0\widehat{\mu}_0)}+\frac{\Gamma_1'(f)}{(1-\eta_1)\Pr\left[f=1,\widehat{Z}=0\right]-\eta_1\Pr\left[f=1,\widehat{Z}=1\right]}$. Similarly, by Inequalities 17 and 20, we have

$$\frac{\Gamma_0(f)}{\Gamma_1(f)}\geq\tau-\delta-\zeta\alpha_0,$$

where $\alpha_0=\frac{\Pr[f=1]}{\Gamma_1'(f)((1-\eta_1)\widehat{\mu}_0-\eta_1\widehat{\mu}_1)}+\frac{\Gamma_0'(f)}{(1-\eta_0)\Pr\left[f=1,\widehat{Z}=1\right]-\eta_0\Pr\left[f=1,\widehat{Z}=0\right]}$. Thus, we have

$$\gamma^{\Delta}(f,\widehat{S})\geq\tau-\delta-\zeta\cdot\max\left\{\alpha_0,\alpha_1\right\}.$$

The influence of the above inequality is that the fairness guarantee of Theorem 3.3 changes to be

$$\gamma(f^{\Delta},S)\geq\tau-3(\delta+\zeta\cdot\max\left\{\alpha_0,\alpha_1\right\}),$$

i.e., the estimation errors will weaken the fairness guarantee of our denoised program. Also, observe that the influence becomes smaller as ζ goes to 0.

D. Proof of Theorem 3.3

In this section, we prove Theorem 3.3 and show how to extend the theorem to multiple protected attributes and multiple fairness constraints (Remark D.7). Denote $\mathcal{Q}_{\text{linf}}$ to be the collection of all group performance functions. Denote $\mathcal{Q}_{\text{lin}}\subseteq\mathcal{Q}_{\text{linf}}$ to be the collection of linear group performance functions.

Remark D.1 *The fairness metric considered in (Awasthi et al., 2020), i.e., equalized odds, can also be captured by $\mathcal{Q}_{\text{linf}}$; equalized odds simply requires equal false positive and true positive rates across the protected types. The fairness metrics used in (Lamy et al., 2019), on the other hand, are somewhat different; they work with statistical parity and equalized odds for binary protected attributes, however, while we define disparity Ω_q as the ratio between the minimum and maximum q_i, (Lamy et al., 2019) define the disparity using the additive difference of q_i across the protected types. It is not apparent how to extend their method for improving additive metrics to linear-fractional fairness metrics as they counter the noise by scaling the tolerance of their constraints, and it is unclear how to compute these scaling parameters prior to the optimization step when the group performance function q is conditioned on the classifier prediction. On the other hand, our method can handle additive metrics by using the difference of altered q_i across the noisy protected attribute to form fairness constraints.*

Similar to Eq (3), we first have for each $i \in [p]$

$$\Pr\left[\xi'(f), \widehat{Z} = i\right] = \sum_{j\in[p]} \Pr\left[\widehat{Z} = i \mid \xi'(f), Z = j\right] \Pr\left[\xi'(f), Z = j\right].$$

By Definition 2.3 and a similar argument as in the proof of Lemma 3.6, we have the following lemma.

Lemma D.2 (Relation between $\Pr\left[\xi'(f), \widehat{Z} = i\right]$ and $\Pr\left[\xi'(f), Z = j\right]$) *Let $\varepsilon \in (0,1)$ be a fixed constant. With probability at least $1 - 2pe^{-\varepsilon^2 n/6}$, we have for each $i \in [p]$,*

$$\Pr\left[\xi'(f), \widehat{Z} = i\right] \in \textstyle\sum_{j\in[p]} H_{ji} \cdot \Pr\left[\xi'(f), Z = j\right] \pm \varepsilon.$$

Define

$$w(f) := \left(\Pr\left[\xi'(f), Z = 1\right], \ldots, \Pr\left[\xi'(f), Z = p\right]\right),$$

and recall that

$$\widehat{w}(f) := \left(\Pr\left[\xi'(f), \widehat{Z} = 1\right], \ldots, \Pr\left[\xi'(f), \widehat{Z} = p\right]\right).$$

By Lemma D.2, we directly obtain the following lemma.

Lemma D.3 (Approximation of $\Pr\left[\xi'(f), Z = i\right]$) *With probability at least $1 - 2pe^{-\varepsilon^2 n/6}$, for each $i \in [p]$,*

$$w(f)_i \in (H^\top)_i^{-1}\widehat{w}(f) \pm \varepsilon\|(H^\top)_i^{-1}\|_1 \in (H^\top)_i^{-1}\widehat{w}(f) \pm \varepsilon M.$$

Thus, we use $(H^\top)_i^{-1}\widehat{w}(f)$ to estimate $\Pr\left[\xi'(f), Z = i\right]$. Similarly, we define

$$u(f) := \left(\Pr\left[\xi(f), \xi'(f), Z = i\right]\right)_{i\in[p]},$$

and recall that

$$\widehat{u}(f) := \left(\Pr\left[\xi(f), \xi'(f), \widehat{Z} = i\right]\right)_{i\in[p]}.$$

Once again, we use $(H^\top)_i^{-1}\widehat{u}(f)$ to estimate $\Pr\left[\xi(f), \xi'(f), Z = i\right]$ and to estimate constraint $\min_{i\in[p]} \Pr\left[\xi(f), \xi(f), \xi'(f), Z = i\right] \geq \lambda$, we construct the following constraint:

$$(H^\top)^{-1}\widehat{u}(f) \geq (\lambda - \varepsilon M)\mathbf{1}, \tag{22}$$

which is the first constraint of Program (**DFair**).

To provide the performance guarantees on the solution of the above program, once again we define the following general notions of bad classifiers and the corresponding capacity.

Definition D.4 (Bad classifiers in general) *Given a family $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, we call $f \in \mathcal{F}$ a bad classifier if f belongs to at least one of the following sub-families:*

- $\mathcal{G}_0 := \left\{f \in \mathcal{F} : \min_{i\in[p]} \Pr\left[\xi(f), \xi'(f), Z = i\right] < \frac{\lambda}{2}\right.$;
- *Let $T = \lceil 232 \log\log \frac{2(\tau-3\delta)}{\lambda} \rceil$. For $i \in [T]$, define*

$$\mathcal{G}_i := \left\{f \in \mathcal{F} \setminus \mathcal{G}_0 : \Omega_q(f, S) \in [\tfrac{\tau-3\delta}{1.01^{2^{i+1}-1}}, \tfrac{\tau-3\delta}{1.01^{2^i-1}})\right\}.$$

Note that Definition 3.7 is a special case of the above definition by letting $p = 2$, $M = 10$, $\xi(f) = (f = 1)$ and $\xi'(f) = \emptyset$. We next propose the following definition of the capacity of bad classifiers.

Definition D.5 (Capacity of bad classifiers in general) *Let $\varepsilon_0 = \frac{\lambda-2\delta}{5M}$. Let $\varepsilon_i = \frac{1.01^{2^{i-1}}\delta}{5}$ for $i \in [T]$ where $T = \lceil 232 \log\log \frac{2(\tau-3\delta)}{\lambda} \rceil$. Given a family $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, we denote the capacity of bad classifiers by*

$$\Phi(\mathcal{F}) := 2pe^{-\varepsilon_0^2 n/6} M_{\varepsilon_0}(\mathcal{G}_0) + 4p \sum_{i\in[T]} e^{-\frac{\varepsilon_i^2\lambda^2 n}{2400M^2}} \cdot M_{\varepsilon_i\lambda/10M}(\mathcal{G}_i).$$

By a similar argument as in Lemma 3.8, we can prove that $\Phi(\mathcal{F})$ is an upper bound of the probability that there exists a bad classifier feasible for Program **DFair**. Now we are ready to prove Theorem 3.3. Actually, we prove the following generalized version.

Theorem D.6 (Performance of Program DFair) *Suppose the VC-dimension of $(S, \mathcal{F})$ is $t \geq 1$. Given any non-singular matrix $H \in [0,1]^{p\times p}$ with $\sum_{j\in[p]} H_{ij} = 1$ for each $i \in [p]$ and $\lambda \in (0, 0.5)$, let $f^\Delta \in \mathcal{F}$ denote an optimal fair classifier of Program **DFair**. With probability at least $1 - \Phi(\mathcal{F}) - 4pe^{-\frac{\lambda^2\delta^2 n}{2400M^2}}$, the following properties hold*

- $\frac{1}{N}\sum_{a\in[N]} L(f^\Delta, s_a) \leq \frac{1}{N}\sum_{a\in[N]} L(f^\star, s_a)$;
- $\Omega_q(f^\Delta, S) \geq \tau - 3\delta$.

Specifically, if the VC-dimension of $(S, \mathcal{F})$ is t and $\delta \in (0,1)$, the success probability is at least $1 - O(pe^{-\frac{\lambda^2\delta^2 n}{60000M^2} + t\ln(50M/\lambda\delta)})$.

The proof is almost the same as in Theorem 3.3: we just need to replace $\frac{1}{1-\eta_0-\eta_1}$ by M everywhere. For multiple fairness constraints, the success probability of Theorem 3.3 changes to be

$$1 - O(kpe^{-\frac{\lambda^2\delta^2 n}{60000M^2} + t\ln(50M/\lambda\delta)}).$$

Proof: Note that the term $4pe^{-\frac{\lambda^2\delta^2 n}{2400M^2}}$ is an upper bound of the probability that $f^\star$ is not feasible for Program **DFair**.

The idea comes from Lemma D.3 by letting $\varepsilon = \frac{\lambda\delta}{20M}$ such that for each $i \in [p]$,

$$w(f^\star)_i \in (1 \pm \frac{\delta}{10})(H^\top)_i^{-1}\widehat{w}(f^\star) \text{ and}$$

$$u(f^\star)_i \in (1 \pm \frac{\delta}{10})(H^\top)_i^{-1}\widehat{u}(f^\star).$$

Consequently, $\frac{1}{N}\sum_{a\in[N]} L(f^\Delta, s_a) \leq \frac{1}{N}\sum_{a\in[N]} L(f^\star, s_a)$. Since $\Phi(\mathcal{F})$ is an upper bound of the probability that there exists a bad classifier feasible for Program **DFair**, we complete the proof. □

Remark D.7 (Generalization to multiple protected attributes and multiple fairness metrics) *For the general case that $m, k \geq 1$, i.e., there exists m protected attributes $Z_1 \in [p_1], \ldots, Z_m \in [p_m]$ and k group performance functions $q^{(1)}, \ldots, q^{(l)}$ together with a threshold vector $\tau \in [0,1]^k$ where each $q^{(l)}$ is on some protected attribute. In this case, we need to make a generalized assumption of Assumption 1, i.e., there exists constant $\lambda \in (0, 0.5)$ such that for any $l \in [k]$,*

$$\min_{i\in[p]} \Pr_D\left[\xi^{(l)}(f^\star), (\xi')^{(l)}(f^\star), Z = i\right] \geq \lambda.$$

The arguments are almost the same except that for each group performance function $q^{(i)}$, we need to construct corresponding denoised constraints and have an individual capacity $\phi^{(i)}(\mathcal{F})$. Consequently, the success probability of Theorem D.6 becomes $1 - O\left(\sum_{i\in[m]} \phi^{(i)}(\mathcal{F})\right)$.

E. Other empirical details and results

We state the exact empirical form of the constraints used for our simulations in this section and then present additional empirical results.

E.1. Implementation of our denoised algorithm.

As a use case, we solve Program **DFair** for logistic regression. Let $\mathcal{F}' = \left\{f'_\theta \mid \theta \in \mathbb{R}^d\right.$ be the family of logistic regression classifiers where for each sample $s = (x, z, y)$, $f'_\theta(x) := \frac{1}{1+e^{-\langle x,\theta\rangle}}$. We learn a classifier $f'_\theta \in \mathcal{F}'$ and then round each $f'_\theta(\widehat{x}_i)$ to $f_\theta(\widehat{x}_i) := \mathbf{I}\left[f(\widehat{x}_i) \geq 0.5\right]$.[9]

We next show how to implement the Program **DFair** for any general fairness constraints. Let $\xi(f)$ and $\xi'(f)$ denote the relevant events to measure the group performances. The constraints use the group-conditional probabilities of these events, i.e. $\widehat{u}(f) := \left(\Pr\left[\xi(f), \xi'(f), \widehat{Z} = i\right]\right)_{i\in[p]}$ and $\widehat{w}(f) := \left(\Pr\left[\xi'(f), \widehat{Z} = i\right]\right)_{i\in[p]}$. Let $N = |S|$ and let $u'(f)$, $w'(f)$ denote the empirical approximation of $\widehat{u}(f)$, $\widehat{w}(f)$ respectively; i.e.,

$$u'(f) := \left(\frac{1}{N}\sum_{\alpha\in[N],\hat{Z}=i} \mathbf{1}\left[\xi(f(x_\alpha)), \xi'(f(x_\alpha))\right]\right)_{i\in[p]},$$

$$w'(f) := \left(\frac{1}{N}\sum_{\alpha\in[N],\hat{Z}=i} \mathbf{1}\left[\xi'(f(x_\alpha))\right]\right)_{i\in[p]}.$$

Let $\Gamma'_i(f) := \left((H^\top)^{-1}u'(f)\right)_i / \left((H^\top)^{-1}w'(f)\right)_i$, for each $i \in [p]$ and $M := \max_{i\in[p]} \|(H^\top)_i^{-1}\|_1$. Then, given $\tau \in [0,1]$ and $\lambda, \delta > 0$, the empirical implementation in Program **DFair** use the following constraints.

$$\begin{cases} \Gamma'_i(f) \geq (\tau - \delta)\cdot\Gamma'_j(f), \forall i, j \in [p]\times[p], \\ \left((H^\top)^{-1}u'(f)\right)_i \geq (\lambda - M\delta), \forall i \in [p]. \end{cases} \tag{23}$$

The program **DLR** simply implements the following optimization problem.

$$\min_{\theta\in\mathbb{R}^d} -\frac{1}{N}\sum_{a\in[N]} \left(y_a \log f_\theta(x_a) + (1-y_a)\log(1 - f_\theta(x_a))\right)$$

$s.t.$ Constraints (23) are satisfied.

(DLR)

Program DFair for statistical rate metric (DLR-SR). For statistical rate metric, simply set $\xi(f_\theta(x_\alpha)) = (f_\theta(x_\alpha) = 1)$ and $\xi'(f_\theta(x_\alpha)) = \emptyset$, and compute the empirical constraints in Eqns 23.

Program DFair for false positive rate metric (DLR-FPR). For false positive rate metric, set $\xi(f_\theta(x_\alpha)) = (f_\theta(x_\alpha) = 1)$ and $\xi'(f_\theta(x_\alpha)) = (Y = 0)$, and compute the empirical constraints in Eqns 23.

Program DFair for false discovery rate metric (DLR-FDR). For false discovery rate metric, simply set $\xi(f_\theta(x_\alpha)) = (Y = 0)$ and $\xi'(f_\theta(x_\alpha)) = (f_\theta(x_\alpha) = 1)$, and compute the empirical constraints in Eqns 23.

If required, one can also append a regularization term $C \cdot \|\theta\|_2^2$ to the above loss function where $C \geq 0$ is a given regularization parameter.

E.2. SLSQP parameters

We use standard constrained optimization packages to solve this program, such as SLSQP (Kraft, 1988) (implemented using python *scipy* package). For each optimization problem, we run the SLSQP algorithm for 500 iterations, starting with a randomly chosen point and with parameters ftol=1e-3 and eps=1e-3.

[9] The extension to non-linear classifiers, such as kernel SVMs, can be done by changing the formulation of f_θ accordingly. For instance, we can extend to kernel SVM by letting $f_\theta(\widehat{x}_i) = I[K(\theta, \widehat{x}_i) \geq 0]$ where K is some non-linear kernel function.

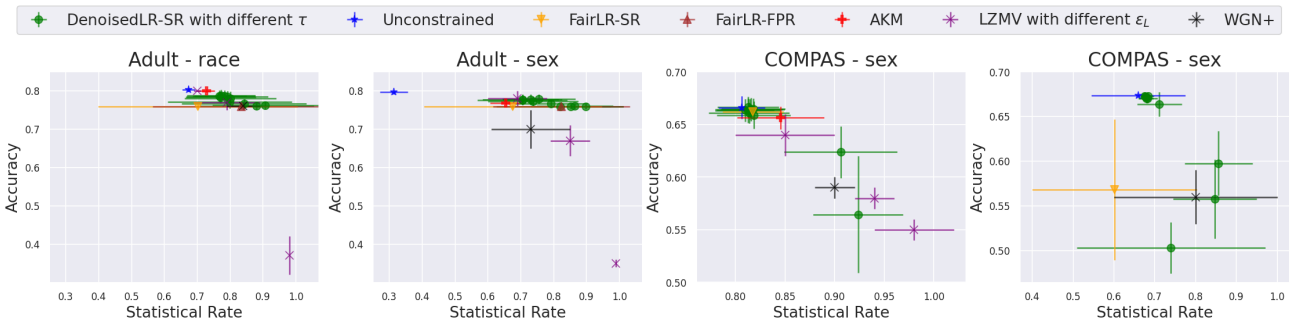


Figure 1. Performance of **DLR-SR** and baselines with respect to statistical rate and accuracy for different combinations of dataset and protected attribute. For **DLR-SR**, the performance for different τ is presented, while for **LZMV** the input parameter ε_L is varied. The plots shows that for all settings **DLR-SR** can attain a high statistical rate, often with minimal loss in accuracy.

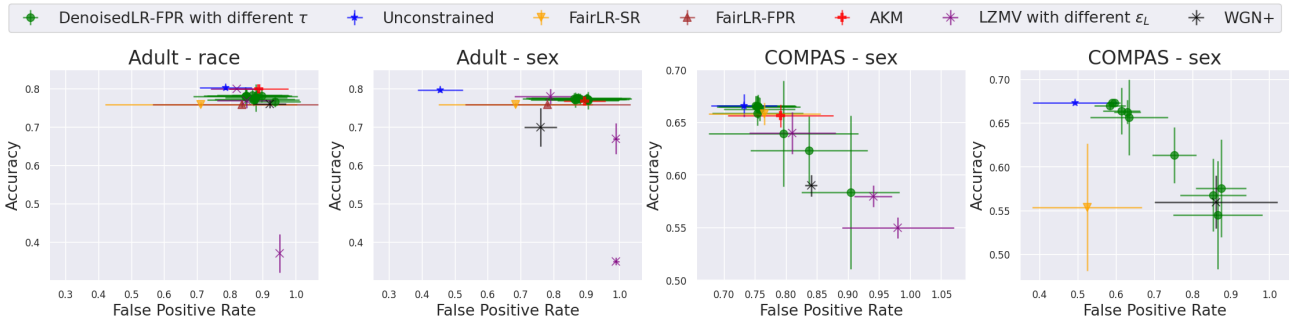


Figure 2. Performance of **DLR-FPR** and baselines with respect to false positive rate and accuracy for different combinations of dataset and protected attribute. For **DLR-FPR**, the performance for different τ is plotted to present the entire fairness-accuracy tradeoff picture. Similarly, for **LZMV** the input parameter ε_L is varied. The plots shows that for all settings **FPR** can attain a high false positive rate, often with minimal loss in accuracy.

Table 2. The performance of all algorithms over test datasets with respect to false discovery rate fairness metric - average and standard error (in brackets) of accuracy and false discovery rate. Our method **DLR-FDR**, with $\tau = 0.9$, achieves higher false discovery rate than baselines in almost every setting, at a minimal cost to accuracy.

	Adult				**COMPAS**			
	sex (binary)		race (binary)		sex (binary)		race (non-binary)	
	acc	FDR	acc	FDR	acc	FDR	acc	FDR
LR-SR	.76 (.01)	.55 (.45)	.76 (.01)	.56 (.46)	.67 (.01)	.66 (0)	.58 (.05)	.73 (.06)
LR-FPR	.76 (.01)	.54 (.45)	.76 (0)	.35 (.43)	.67 (.01)	.75 (.09)	.56 (.05)	.72 (.05)
LZMV $\varepsilon_L = .01$	.35 (.01)	0 (0)	.37 (.05)	0 (0)	.55 (.01)	.74 (.04)	-	-
LZMV $\varepsilon_L = .04$	.67 (.04)	0 (0)	.77 (.03)	0 (0)	.58 (.01)	.74 (.04)	-	-
LZMV $\varepsilon_L = .10$	.78 (.02)	.47 (.01)	.80 (0)	.76 (.05)	.64 (.02)	.83 (.04)	-	-
AKM	.77 (0)	.55 (.17)	.80 (0)	.71 (.01)	.69 (.01)	.75 (.03)	-	-
WGN+	.59 (0)	.54 (.02)	.67 (0)	.65 (.01)	.54 (.01)	.72 (.05)	.56 (.03)	.68 (.07)
DLR-FDR $\tau = .7$	.73 (.04)	.66 (.07)	.80 (.02)	.76 (.06)	.64 (.03)	.75 (.11)	.67 (.02)	.79 (.03)
DLR-FDR $\tau = .9$	.75 (.01)	.87 (.08)	.76 (.02)	.89 (.09)	.60 (.07)	.77 (.10)	.54 (.13)	.79 (.07)

E.3. Baselines' parameters

LZMV: For this algorithm of Lamy et al. (2019), we use the implementation from https://github.com/AIasd/noise_fairlearn . The constraints are with respect to additive statistical rate. The fairness tolerance parameter ε (referred to as ε_L in our empirical results to avoid confusion) are chosen to be $\{0.01, 0.04, 0.10\}$ to present the range of performance of the algorithm. See the paper (Lamy et al., 2019) for descriptions of these parameters. The base classifier used is the algorithm of Agarwal et al. (2018), and the noise parameters are provided as input to the **LZMV** algorithm.

AKM: For this algorithm, we use the implementation from https://github.com/matthklein/equalized_odds_under_perturbation. The constraints are with respect to additive false positive rate parity. Once again, the algorithm takes noise parameters as input and uses the base classifier of Hardt et al. (2016).

WGN+: For this algorithm, we use the implementation from https://github.com/wenshuoguo/robust-fairness-code. Once again, the constraints here are additive false positive rate constraints using the soft-group assignments. See the paper (Wang et al., 2020) for descriptions of these parameters. The learning rate parameters used for this algorithm are $\eta_\theta \in \{.001, 0.01, 0.1\}$, $\eta_\lambda \in \{0.5, 1.0, 2.0\}$, and $\eta_W \in \{0.01, 0.1\}$. These parameters are same as the one the authors suggest in their paper and code. We run their algorithm for all combinations of the above parameters and select and report the test performance of the model that has the best training objective value, while satisfying the program constraints.

E.4. Other results

In this section, we present other empirical results to complement the arguments made in Section 4. First, we present the plot for comparison of all methods with respect to statistical rate, Figure 1, and false positive rate, Figure 2.

E.4.1. Performance with respect to false discovery rate

We also present the empirical performance of our algorithm, compared to baselines, when the fairness metric in consideration is false discovery rate (a linear-fractional metric). Table 2 presents the results. For most combinations of datasets and protected attributes, our method **DLR-FDR**, with $\tau = 0.9$, achieves a higher false discovery rate than baselines, at a minimal cost to accuracy.

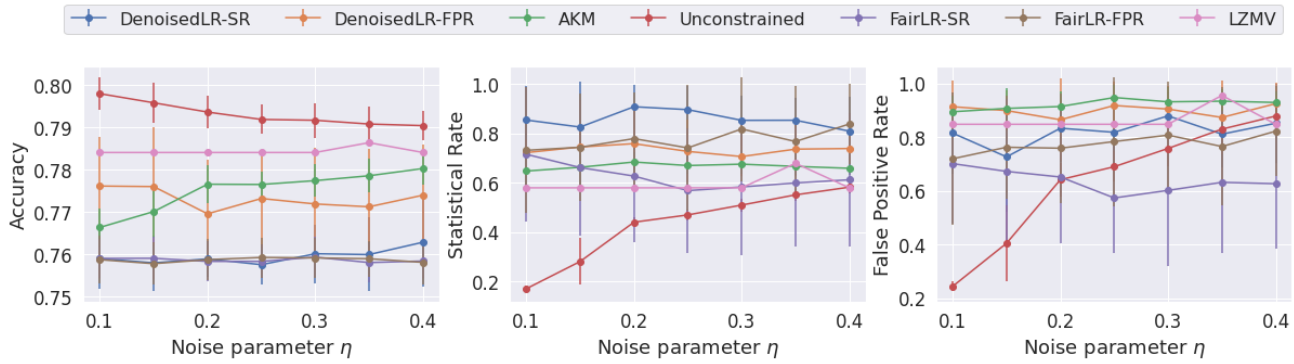
E.4.2. Variation of noise parameter

We also investigate the performances of algorithms w.r.t. varying η_0, η_1. We consider $\eta_0 = \eta_1 = \eta \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$ for the binary case, and $H_{i,j} \in \{0.05, \cdots, 0.25\}$, for $i \neq j$, in the non-binary case. Other settings are the same as in the main text. We select $\tau = 0.9$ for **FairLR** and **DLR**. The performance on Adult dataset is presented in Figure 3 when sex is the protected attribute and in Figure 4 when race is the protected attribute. The performance on COMPAS dataset is presented in Figure 5 when sex is the protected attribute and in Figure 6 when race is the protected attribute.

E.4.3. Error in noise parameter estimation

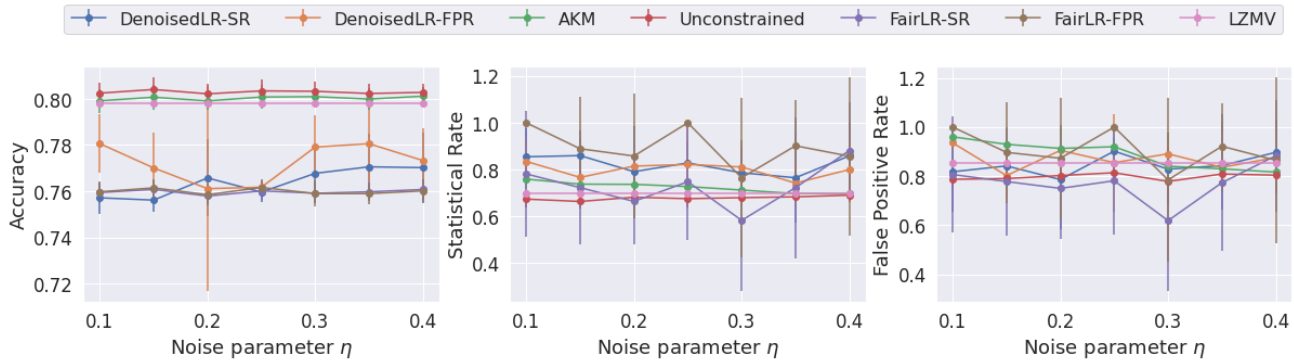
As discussed at the end of Section 3.2, the scale of error in the noise parameter estimation can affect the fairness guarantees. In this section, we empirically look at the impact of estimation error on the statistical rate of the generated classifier.

We set the true noise parameters $\eta_0 = \eta_1 = 0.3$, in case of binary protected attribute, and $H = \begin{bmatrix} 0.70 & 0.15 & 0.15 \\ 0.15 & 0.70 & 0.15 \\ 0.15 & 0.15 & 0.70 \end{bmatrix}$, in case of non-binary protected attribute. The estimated noise parameter ranges η' ranges from 0.1 to 0.3. In case of non-binary protected attribute, the noise is distributed equally amongst all different protected attribute values (e.g., when $\eta' = 0.1$, $Z = 0$ flips to $Z = 1$ with probability



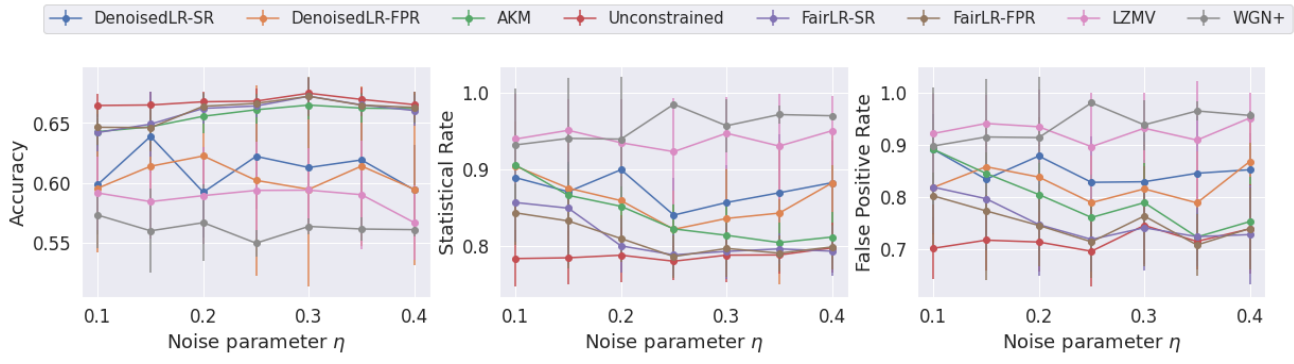
(a) Accuracy vs η (b) Statistical Rate vs η (c) False Positive Rate vs η

Figure 3. Performance of **DLR-SR**, **DLR-FPR** ($\tau = 0.9$) and baselines with respect to statistical rate, false positive rate and accuracy for different noise parameters η. The dataset used is **Adult** and the protected attribute is sex.



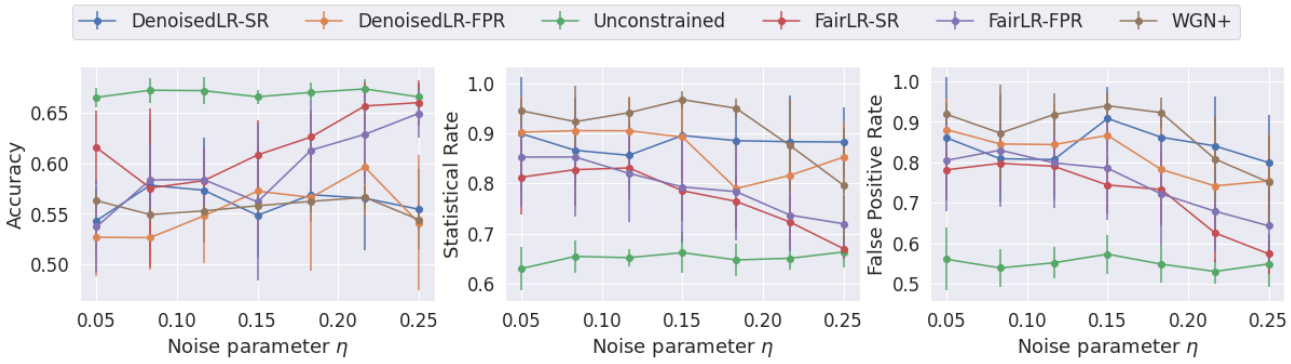
(a) Accuracy vs η (b) Statistical Rate vs η (c) False Positive Rate vs η

Figure 4. Performance of **DLR-SR**, **DLR-FPR** ($\tau = 0.9$) and baselines with respect to statistical rate, false positive rate and accuracy for different noise parameters η. The dataset used is **Adult** and the protected attribute is race.



(a) Accuracy vs η (b) Statistical Rate vs η (c) False Positive Rate vs η

Figure 5. Performance of **DLR-SR**, **DLR-FPR** ($\tau = 0.9$) and baselines with respect to statistical rate, false positive rate and accuracy for different noise parameters η. The dataset used is **COMPAS** and the protected attribute is sex.



(a) Accuracy vs η (b) Statistical Rate vs η (c) False Positive Rate vs η

Figure 6. Performance of **DLR-SR**, **DLR-FPR** ($\tau = 0.9$) and baselines with respect to statistical rate, false positive rate and accuracy for different noise parameters η. The dataset used is **COMPAS** and the protected attribute is race.

0.05 and to $Z = 2$ with probability 0.05). The variation of accuracy and statistical rate with noise parameter estimate of **DenoisedLR-SR** for COMPAS and Adult datasets is presented in Figure 7a,b. The plots show that, in most settings, the best statistical rate (close to the desired guarantee of 0.90) is achieved when the estimate is close to the true noise parameter value. However, even for estimates that are considerably lower than the true estimate (for instance, $\eta' < 0.2$), the average statistical rate is still quite high (> 0.75).

The results show that if the error in the noise parameter estimate is reasonable, the framework ensures that the fairness of the generated classifier is still high.

E.4.4. Performance using predicted protected attribute

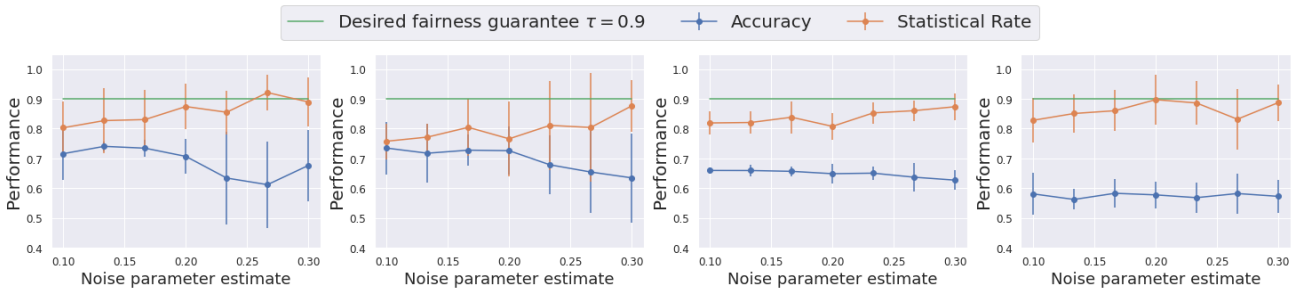
The primary empirical results consider the setting when the noise in the protected attribute is i.i.d. While this assumption is necessary for our theoretical analysis, it may not be satisfied in many real-world settings, for example, when the protected attribute is predicted using other features. In this section, we present the empirical performance of our approach when the protected attribute is partially predicted using other non-protected features.

Methodology. We randomly split a given dataset into three parts (40-40-20 split). The first partition is treated as an auxiliary dataset for which the underlying protected attributes are known and is used to train a protected attribute prediction model. Using this auxiliary partition, we construct a simple 2-layer multi-perceptron protected attribute classification model g. To predict the protected attribute of any new sample, we return the true label of the sample with probability 0.5 and return the label predicted using g with probability 0.5 (this way we ensure that the requirement of less than 50% corrupted samples - on average - for any protected attribute in Definition 2.3 is satisfied).

The normalized confusion matrix for this prediction process on the auxiliary dataset is used as the noise matrix H for the rest of our analysis. The above prediction model is then used to predict the protected attributes of the second and third partition, and the predicted protected attributes are treated as the noisy protected attributes. The second partition is employed as the train partition for the denoised fair classification algorithms and the third partition is the test partition to evaluate the performance. The rest of the parameters are kept to be the same as the simulations in the main body. We repeat this process multiple times with multiple random splits of both Adult and COMPAS datasets and report the mean and standard error of the accuracy and fairness of the returned classifier over the test dataset.

Results. The results are presented in Table 3. Despite the fact that the noise in this case is non-identical, **DLR** with $\tau = 0.9$ is still able to achieve high values of fairness. In all settings, the mean of the fairness of the output classifier is greater than ≥ 0.79.

In the case of statistical rate metric, **DLR-SR** can achieve higher fairness than baselines in all cases except **Adult** with race as the protected attribute. In the case of false positive rate metric, **DLR-FPR** can achieve higher fairness than baselines in all cases except **Adult** with sex as the protected attribute. Both methods, for all settings, achieve high fairness at certain cost to accuracy, showing that our approach can indeed handle settings where protected attribute is partially predicted.



(a) Adult - sex (b) Adult - race (c) COMPAS - sex (d) COMPAS - race

Figure 7. Performance of **DLR-SR** ($\tau = 0.9$) with respect to statistical rate and accuracy for different noise parameter estimate η'. The true noise parameters are $\eta_0 = \eta_1 = 0.3$.

Table 3. The performance on accuracy and fairness metrics of all algorithms over the test datasets when the protected attribute is partially predicted using other non-protected attributes; we report the average and standard error (in parenthesis) across multiple random splits of the dataset.

	Adult						**COMPAS**					
		sex (binary)			race (binary)			sex (binary)			race (non-binary)	
	acc	SR	FPR	acc	SR	FPR	acc	SR	FPR	acc	SR	FPR
Unconstrained	.80 (.01)	.33 (.02)	.49 (.03)	.80 (.01)	.52 (.06)	.52 (.09)	.66 (.01)	.78 (.04)	.70 (.07)	.66 (.01)	.62 (.05)	.56 (.09)
LR-SR	.75 (.03)	.79 (.10)	.83 (.13)	.74 (.06)	.82 (.11)	.87 (.12)	.61 (.05)	.87 (.05)	.91 (.03)	.56 (.05)	.83 (.12)	.77 (.13)
LR-FPR	.73 (.11)	.74 (.09)	.89 (.06)	.78 (.02)	.67 (.20)	.69 (.21)	.62 (.03)	.89 (.06)	.93 (.04)	.56 (.07)	.74 (.19)	.72 (.18)
DLR-SR τ=.9	.75 (.02)	.85 (.11)	.82 (.17)	.73 (.05)	.79 (.10)	.85 (.09)	.63 (.02)	.87 (.02)	.94 (.02)	.55 (.03)	.95 (.04)	.92 (.05)
DLR-FPR τ=.9	.76 (.02)	.70 (.15)	.80 (.22)	.77 (.03)	.76 (.08)	.83 (.10)	.63 (.04)	.85 (.07)	.93 (.05)	.55 (.05)	.89 (.12)	.90 (.09)

F. Discussion of initial attempts

We first discuss two natural ideas including randomized labeling (Section F.1) and solving Program ConFair that only depends on $\widehat{S}$ (Section F.2). For simplicity, we consider the same setting as in Section 3.3: $p = 2$ with statistical rate, and assume $\eta = \eta_1 = \eta_2 \in (0, 0.4)$. We also discuss their weakness on either the empirical loss or the fairness constraints. This section aims to show that directly applying the same fairness constraints on $\widehat{S}$ may introduce bias on S and, hence, our modifications to the constraints (Definition 3.1) are necessary; see Section F in the Supplementary Material for a discussion.

F.1. Randomized labeling

A simple idea is that for each sample $s_a \in S$, i.i.d. draw the label $f(s_a)$ to be 0 with probability α and to be 1 with probability $1 - \alpha$ ($\alpha \in [0, 1]$). This simple idea leads to a fair classifier by the following lemma.

Lemma F.1 (A random classifier is fair) *Let $f \in \{0, 1\}^{\mathcal{X}}$ be a classifier generated by randomized labeling. With probability at least $1 - 2e^{-\frac{\alpha\lambda N}{1.2\times 10^5}}$, $\gamma(f, S) \geq 0.99$.*

Proof: Let $A = \{a \in [N] : z_a = 0\}$ be the collection of samples with $Z = 0$. By Assumption 1, we know that $|A| \geq \lambda N$. For $a \in A$, let X_a be the random variable where $X_a = f(s_a)$. By randomized labeling, we know that $\Pr[X_i = 1] = \alpha$. Also,

$$\Pr[f = 1 \mid Z = 0] = \frac{\sum_{i\in A} X_i}{|A|}. \tag{24}$$

Since all X_i ($i \in A$) are independent, we have

$$\begin{aligned} &\Pr\left[\sum_{i\in A} X_i \in (1 \pm 0.005) \cdot \alpha|A|\right] \\ \geq\ & 1 - 2e^{-\frac{0.005^2\alpha|A|}{3}} \quad \text{(Chernoff bound)} \\ \geq\ & 1 - 2e^{-\frac{\alpha\lambda N}{1.2\times 10^5}}. \quad (|A| \geq \lambda N) \end{aligned} \tag{25}$$

Thus, with probability at least $1 - 2e^{-\frac{\alpha\lambda N}{1.2\times 10^5}}$,

$$\begin{aligned} &\Pr[f = 1 \mid Z = 0] \\ =\ & \frac{\sum_{i\in A} X_i}{|A|} \quad \text{(Eq. 24)} \\ \in\ & (1 \pm 0.005) \cdot \frac{\alpha|A|}{|A|} \quad \text{(Ineq. 25)} \\ \in\ & (1 \pm 0.005)\alpha. \end{aligned}$$

Similarly, we have that with probability at least $1 - 2e^{-\frac{\alpha\lambda N}{1.2\times 10^5}}$,

$$\Pr[f = 1 \mid Z = 1] \in (1 \pm 0.005)\alpha.$$

By the definition of $\gamma(f, S)$, we complete the proof. □

However, there is no guarantee for the empirical risk of randomized labeling. For instance, consider the loss function $L(f, s) := \mathbf{I}[f(s) = y]$ where $\mathbf{I}[\cdot]$ is the indicator function, and suppose there are $\frac{N}{2}$ samples with $y_a = 0$. In this setting, the empirical risk of $f^\star$ may be close to 0, e.g., $f^\star = Y$. Meanwhile, the expected empirical risk of randomized labeling is

$$\frac{1}{N}\left((1-\alpha)\cdot\frac{N}{2} + \alpha\cdot\frac{N}{2}\right) = \frac{1}{2},$$

which is much larger than that of $f^\star$.

F.2. Replacing S by $\widehat{S}$ in Program TargetFair

Another idea is to solve the following program which only depends on $\widehat{S}$, i.e., simply replacing S by $\widehat{S}$ in Program **TargetFair**.

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{a\in[N]} L(f, \widehat{s}_a) \quad s.t. \quad \gamma(f, \widehat{S}) \geq \tau. \tag{ConFair}$$

Remark F.2 *Similar to Section 4, we can design an algorithm that solves Program ConFair by logistic regression.*

$$\begin{aligned}
&\min_{\theta\in\mathbb{R}^d} -\frac{1}{N}\sum_{a\in[N]} \left(y_a \log f_\theta(s_a) + (1-y_a)\log(1-f_\theta(s_a))\right)\\
s.t.\ &\widehat{\mu}_1 \cdot \sum_{a\in[N]:\widehat{Z}=0} \mathbf{I}[\langle x_a, \theta\rangle \geq 0]\\
\geq &\tau\widehat{\mu}_0 \cdot \sum_{a\in[N]:\widehat{Z}=1} \mathbf{I}[\langle x_a, \theta\rangle \geq 0],\\
&\widehat{\mu}_0 \cdot \sum_{a\in[N]:\widehat{Z}=1} \mathbf{I}[\langle x_a, \theta\rangle \geq 0]\\
\geq &\tau\widehat{\mu}_1 \cdot \sum_{a\in[N]:\widehat{Z}=0} \mathbf{I}[\langle x_a, \theta\rangle \geq 0].
\end{aligned} \tag{FairLR}$$

Let $\widehat{f}^\star$ denote an optimal solution of Program ConFair. Ideally, we want to use $\widehat{f}^\star$ to estimate $f^\star$. Since Z is not used for prediction, we have that for any $f \in \mathcal{F}$,

$$\sum_{a\in[N]} L(f, s_a) = \sum_{a\in[N]} L(f, \widehat{s}_a).$$

Then if $\widehat{f}^\star$ satisfies $\gamma(\widehat{f}^\star, S) \geq \tau$, we conclude that $\widehat{f}^\star$ is also an optimal solution of Program **TargetFair**. However, due to the flipping noises, $\widehat{f}^\star$ may be far from $f^\star$ (Example F.3). More concretely, it is possible that $\gamma(\widehat{f}^\star, S) \ll \tau$ (Lemma F.4). Moreover, we discuss the range of $\Omega(f^\star, \widehat{S})$ (Lemma F.5). We find that $\Omega(f^\star, \widehat{S}) < \tau$ may hold which implies that $f^\star$ may not be feasible for Program ConFair. We first give an example showing that $\widehat{f}^\star$ can perform very bad over S with respect to the fairness metric.

Example F.3 *Our example is shown in Figure 8. We assume that $\mu_0 = 1/3$ and $\mu_1 = 2/3$. Let $\eta = 1/3$ be the noise parameter and we assume $\pi_{20} = \pi_{01} = 1/3$. Consequently, we have that*

$$\widehat{\mu}_0 = 1/3 \times 2/3 + 2/3 * 1/3 = 4/9.$$

Then we consider the following simple classifier $f \in \{0,1\}^{\mathcal{X}}$: $\widehat{f}^\star = Z$. We directly have that $\Pr\left[\widehat{f}^\star = 1 \mid Z = 0\right] = 0$ and $\Pr\left[\widehat{f}^\star = 1 \mid Z = 1\right] = 1$, which implies that $\gamma(\widehat{f}^\star, S) = 0$. We also have that

$$\begin{aligned}
&\Pr\left[\widehat{f}^\star = 1 \mid \widehat{Z} = 0\right]\\
=\ &\Pr\left[Z = 1 \mid \widehat{Z} = 0\right] \quad (\widehat{f}^\star = Z)\\
=\ &\frac{\pi_{01}\cdot\mu_1}{\widehat{\mu}_0} \quad (\textit{Observation B.1})\\
=\ &0.5,
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\left[\widehat{f}^\star = 1 \mid \widehat{Z} = 1\right]\\
=\ &\Pr\left[Z = 1 \mid \widehat{Z} = 1\right] \quad (\widehat{f}^\star = Z)\\
=\ &\frac{\pi_{11}\cdot\mu_1}{\widehat{\mu}_1} \quad (\textit{Observation B.1})\\
=\ &0.8,
\end{aligned}$$

which implies that $\gamma(\widehat{f}^\star, \widehat{S}) = 0.625$. Hence, there is a gap between $\gamma(\widehat{f}^\star, S)$ and $\gamma(\widehat{f}^\star, \widehat{S})$, say 0.625, in this example. Consequently, $\widehat{f}^\star$ can be very unfair over S, and hence, is far from $f^\star$.

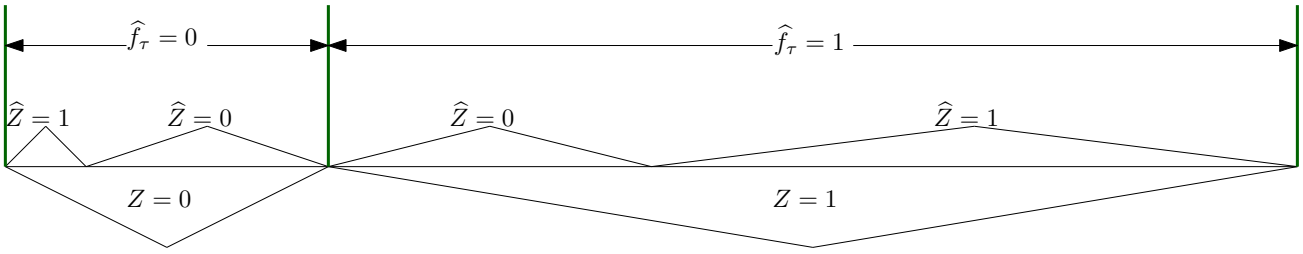
Next, we give some theoretical results showing the weaknesses of Program ConFair.

An upper bound for $\gamma(f, S)$. More generally, given a classifier $f \in \{0,1\}^{\mathcal{X}}$, we provide an upper bound for $\gamma(f, S)$ that is represented by $\gamma(f, \widehat{S})$; see the following lemma.

Lemma F.4 (An upper bound for $\gamma(f, S)$) *Suppose we have*

1. $\Pr\left[f = 1 \mid \widehat{Z} = 0\right] \leq \Pr\left[f = 1 \mid \widehat{Z} = 1\right]$;
2. $\Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 0\right] \leq \alpha_0 \cdot \Pr\left[f = 1, Z = 1 \mid \widehat{Z} = 0\right]$ *for some* $\alpha_0 \in [0, 1]$;

Figure 8. An example showing that $\gamma(f, S)$ and $\gamma(f, \widehat{S})$ can differ by a lot. The detailed explanation can be found in Example F.3.



3. $\Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 1\right] \leq \alpha_1 \cdot \Pr\left[f = 1, Z = 1 \mid \widehat{Z} = 1\right]$ *for some* $\alpha_1 \in [0, 1]$.

Let $\beta_{ij} = \frac{\widehat{\mu}_i}{\mu_j}$ *for* $i, j \in \{0, 1\}$. *The following inequality holds*

$$\begin{aligned}\gamma(f, S) \leq & \frac{\alpha_0(1+\alpha_1)\beta_{00}\cdot\gamma(f,\widehat{S})+\alpha_1(1+\alpha_0)\beta_{10}}{(1+\alpha_1)\beta_{01}\cdot\gamma(f,\widehat{S})+(1+\alpha_0)\beta_{11}} \\ \leq & \max\{\alpha_0, \alpha_1\} \cdot \frac{\mu_1}{\mu_0}.\end{aligned}$$

The intuition of the first assumption is that the statistical rate for $Z = 0$ is at most that for $Z = 1$ over the noisy dataset $\widehat{S}$. The second and the third assumptions require the classifier f to be less positive when $Z = 0$. Intuitively, f is restricted to induce a smaller statistical rate for $Z = 0$ over both S and $\widehat{S}$. Specifically, if $\alpha_0 = \alpha_1 = 0$ as in Example F.3, we have $\gamma(f, S) = 0$. Even if $\alpha_0 = \alpha_1 = 1$, we have $\gamma(f, S) \leq \frac{\mu_1}{\mu_0}$ which does not depend on $\gamma(f, \widehat{S})$.

Proof: [Proof of Lemma F.4] By the first assumption, we have

$$\gamma(f, \widehat{S}) = \frac{\Pr\left[f = 1 \mid \widehat{Z} = 0\right]}{\Pr\left[f = 1 \mid \widehat{Z} = 1\right]}. \tag{26}$$

By the second assumption, we have

$$\begin{aligned}& \Pr\left[f = 1, Z = 1 \mid \widehat{Z} = 0\right] \\ = & \frac{(1+\alpha_0)\cdot\Pr\left[f = 1, Z = 1 \mid \widehat{Z} = 0\right]}{1+\alpha_0} \\ \geq & \frac{\Pr\left[f = 1, Z = 1 \mid \widehat{Z} = 0\right]}{1+\alpha_0} \\ & + \frac{\Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 0\right]}{1+\alpha_0} \\ = & \frac{1}{1+\alpha_0}\cdot\Pr\left[f = 1 \mid \widehat{Z} = 0\right].\end{aligned} \tag{27}$$

Similarly, we have the following

$$\begin{aligned}& \Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 0\right] \\ \leq & \frac{\alpha_0}{1+\alpha_0}\Pr\left[f = 1 \mid \widehat{Z} = 0\right].\end{aligned} \tag{28}$$

Also, by the third assumption, we have

$$\begin{aligned}& \Pr\left[f = 1, Z = 1 \mid \widehat{Z} = 1\right] \\ \geq & \frac{1}{1+\alpha_1}\Pr\left[f = 1 \mid \widehat{Z} = 1\right],\end{aligned} \tag{29}$$

and

$$\begin{aligned}& \Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 1\right] \\ \leq & \frac{\alpha_1}{1+\alpha_1}\Pr\left[f = 1 \mid \widehat{Z} = 1\right].\end{aligned} \tag{30}$$

Then

$$\begin{aligned}& \Pr\left[f = 1 \mid Z = 0\right] \\ = & \Pr\left[f = 1, \widehat{Z} = 0 \mid Z = 0\right] \\ & + \Pr\left[f = 1, \widehat{Z} = 1 \mid Z = 0\right] \\ = & \Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 0\right]\cdot\frac{\widehat{\mu}_0}{\mu_0} \\ & + \Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 1\right]\cdot\frac{\widehat{\mu}_1}{\mu_0} \\ = & \Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 0\right]\cdot\beta_{00} \\ & + \Pr\left[f = 1, Z = 0 \mid \widehat{Z} = 1\right]\cdot\beta_{10} \\ & \text{(Defn. of } \beta_{00} \text{ and } \beta_{10}\text{)} \\ \leq & \frac{\alpha_0\beta_{00}}{1+\alpha_0}\cdot\Pr\left[f = 1 \mid \widehat{Z} = 0\right] \\ & + \frac{\alpha_1\beta_{10}}{1+\alpha_1}\cdot\Pr\left[f = 1 \mid \widehat{Z} = 1\right]. \\ & \text{(Ineqs. 28 and 30)}\end{aligned} \tag{31}$$

By a similar argument, we have

$$\begin{aligned} & \Pr[f=1 \mid Z=1] \\ = \; & \Pr\left[f=1, Z=1 \mid \widehat{Z}=0\right] \cdot \beta_{01} \\ & + \Pr\left[f=1, Z=1 \mid \widehat{Z}=1\right] \cdot \beta_{11} \\ & \text{(Defn. of } \beta_{01} \text{ and } \beta_{11}) \\ \geq \; & \frac{\beta_{01}}{1+\alpha_0} \cdot \Pr\left[f=1 \mid \widehat{Z}=0\right] \\ & + \frac{\beta_{11}}{1+\alpha_1} \cdot \Pr\left[f=1 \mid \widehat{Z}=1\right]. \\ & \text{(Ineqs. 27 and 29)} \end{aligned} \tag{32}$$

Thus, we have

$$\begin{aligned} & \gamma(f, S) \\ \leq \; & \frac{\Pr[f=1 \mid Z=0]}{\Pr[f=1 \mid Z=1]} \quad \text{(Defn. of } \gamma(f,S)) \\ \leq \; & \frac{\frac{\alpha_0\beta_{00}}{1+\alpha_0}\Pr\left[f=1 \mid \widehat{Z}=0\right] + \frac{\alpha_1\beta_{10}}{1+\alpha_1}\Pr\left[f=1 \mid \widehat{Z}=1\right]}{\frac{\beta_{01}}{1+\alpha_0}\Pr\left[f=1 \mid \widehat{Z}=0\right] + \frac{\beta_{11}}{1+\alpha_1}\Pr\left[f=1 \mid \widehat{Z}=1\right]} \\ & \text{(Ineqs. 31 and 32)} \\ = \; & \frac{\alpha_0(1+\alpha_1)\beta_{00} \cdot \gamma(f, \widehat{S}) + \alpha_1(1+\alpha_0)\beta_{10}}{(1+\alpha_1)\beta_{01} \cdot \gamma(f, \widehat{S}) + (1+\alpha_0)\beta_{11}} \quad \text{(Eq. 26)} \\ \leq \; & \max\left\{\alpha_0 \cdot \frac{\beta_{00}}{\beta_{01}}, \alpha_1 \cdot \frac{\beta_{10}}{\beta_{11}}\right\} \\ = \; & \max\{\alpha_0, \alpha_1\} \cdot \frac{\mu_1}{\mu_0}, \quad \text{(Defn. of } \beta_{ij}) \end{aligned}$$

which completes the proof. □

$f^\star$ may not be feasible in Program ConFair. We consider a simple case that $\eta_1 = \eta_2 = \eta$. Without loss of generality, we assume that $\Pr[f^\star = 1 \mid Z = 0] \leq \Pr[f^\star = 1 \mid Z = 1]$, i.e., the statistical rate of $Z = 0$ is smaller than that of $Z = 1$ over S. Consequently, we have

$$\gamma(f^\star, S) = \frac{\Pr[f^\star = 1 \mid Z = 0]}{\Pr[f^\star = 1 \mid Z = 1]}.$$

Lemma F.5 (Range of $\Omega(f^\star, \widehat{S})$) *Let $\varepsilon \in (0, 0.5)$ be a given constant and let*

$$\Gamma = \frac{\eta\mu_0 + (1-\eta)(1-\mu_0)}{(1-\eta)\mu_0 + \eta(1-\mu_0)} \times \frac{(1-\eta)\mu_0\gamma(f^\star, S) + \eta(1-\mu_0)}{\eta\mu_0\gamma(f^\star, S) + (1-\eta)(1-\mu_0)}.$$

With probability at least $1 - 4e^{-\frac{\varepsilon^2\eta\lambda N}{192}}$, the following holds

$$\gamma(f^\star, \widehat{S}) \in (1 \pm \varepsilon) \cdot \min\left\{\Gamma, \frac{1}{\Gamma}\right\}.$$

For instance, if $\mu_0 = 0.5$, $\gamma(f^\star, S) = 0.8 = \tau$ and $\eta = 0.2$, we have

$$\gamma(f^\star, \widehat{S}) \approx 0.69 < \tau.$$

Then $f^\star$ is not a feasible solution of Program ConFair. Before proving the lemma, we give some intuitions.

Discussion F.6 *By assumption, we have that for a given classifier $f^\star \in \mathcal{F}$,*

$$\Pr\left[\widehat{Z}=1 \mid Z=0\right] \approx \Pr\left[\widehat{Z}=0 \mid Z=1\right] \approx \eta \tag{33}$$

Moreover, the above property also holds when conditioned on a subset of samples with $Z = 0$ or $Z = 1$. Specifically, for $i \in \{0, 1\}$,

$$\begin{aligned} & \Pr\left[\widehat{Z}=1 \mid f^\star=1, Z=0\right] \\ \approx \; & \Pr\left[\widehat{Z}=0 \mid f^\star=1, Z=1\right] \approx \eta \end{aligned} \tag{34}$$

Another consequence of Property 33 is that for $i \in \{0, 1\}$,

$$\begin{aligned} \widehat{\mu}_i = \; & \pi_{i,i}\mu_i + \pi_{i,1-i}\mu_{1-i} \quad (\textit{Observation B.1}) \\ \approx \; & (1-\eta)\mu_i + \eta\mu_{1-i}. \quad (\textit{Property 33}) \end{aligned} \tag{35}$$

Then we have

$$\begin{aligned}
&\Pr\left[f^\star=1\mid\widehat{Z}=0\right]\\
=\;&\Pr\left[f^\star=1,Z=0\mid\widehat{Z}=0\right]\\
&+\Pr\left[f^\star=1,Z=1\mid\widehat{Z}=0\right]\\
=\;&\Pr\left[Z=0\mid\widehat{Z}=0\right]\cdot\Pr\left[f^\star=1\mid Z=0,\widehat{Z}=0\right]\\
&+\Pr\left[Z=1\mid\widehat{Z}=0\right]\cdot\Pr\left[f^\star=1\mid Z=1,\widehat{Z}=0\right]\\
=\;&\frac{\pi_{00}\mu_0}{\widehat{\mu}_0}\cdot\Pr\left[f^\star=1\mid Z=0,\widehat{Z}=0\right]\\
&+\frac{\pi_{01}\mu_1}{\widehat{\mu}_0}\cdot\Pr\left[f^\star=1\mid Z=1,\widehat{Z}=0\right]\\
&(\textit{Observation B.1})\\
\approx\;&\frac{(1-\eta)\mu_0}{(1-\eta)\mu_0+\eta\mu_1}\cdot\Pr\left[f^\star=1\mid Z=0,\widehat{Z}=0\right]\\
&+\frac{\eta\mu_1}{(1-\eta)\mu_0+\eta\mu_1}\cdot\Pr\left[f^\star=1\mid Z=1,\widehat{Z}=0\right]\\
&(\textit{Properties 33 and 35})\\
=\;&\frac{(1-\eta)\mu_0}{(1-\eta)\mu_0+\eta(1-\mu_0)}\times\\
&\frac{\Pr\left[f^\star=1\mid Z=0\right]\cdot\Pr\left[\widehat{Z}=0\mid f^\star=1,Z=0\right]}{\Pr\left[\widehat{Z}=0\mid Z=0\right]}\\
&+\frac{\eta\mu_1}{(1-\eta)\mu_0+\eta(1-\mu_0)}\times\\
&\frac{\Pr\left[f^\star=1\mid Z=1\right]\cdot\Pr\left[\widehat{Z}=0\mid f^\star=1,Z=1\right]}{\Pr\left[\widehat{Z}=0\mid Z=1\right]}\\
\approx\;&\frac{(1-\eta)\mu_0}{(1-\eta)\mu_0+\eta(1-\mu_0)}\cdot\Pr\left[f^\star=1\mid Z=0\right]\\
&+\frac{\eta\mu_1}{(1-\eta)\mu_0+\eta(1-\mu_0)}\cdot\Pr\left[f^\star=1\mid Z=1\right].\\
&(\textit{Properties 33 and 34})
\end{aligned}$$

Similarly, we can represent

$$\begin{aligned}
&\Pr\left[f^\star=1\mid\widehat{Z}=1\right]\\
\approx\;&\frac{\eta\mu_0}{\eta\mu_0+(1-\eta)(1-\mu_0)}\Pr\left[f^\star=1\mid Z=0\right]\\
&+\frac{(1-\eta)\mu_1}{\eta\mu_0+(1-\eta)(1-\mu_0)}\Pr\left[f^\star=1\mid Z=1\right].
\end{aligned}$$

Applying the approximate values of $\Pr\left[f^\star=1\mid\widehat{Z}=0\right]$ *and* $\Pr\left[f^\star=1\mid\widehat{Z}=1\right]$ *to compute* $\gamma(f^\star,S)$*, we have Lemma F.5.*

Proof: [Proof of Lemma F.5] By definition, we have

$$\gamma(f^\star,\widehat{S})\le\frac{\Pr\left[f^\star=1\mid\widehat{Z}=0\right]}{\Pr\left[f^\star=1\mid\widehat{Z}=1\right]}.$$

Thus, it suffices to provide an upper bound for $\Pr\left[f^\star=1\mid\widehat{Z}=0\right]$ and a lower bound for $\Pr\left[f^\star=1\mid\widehat{Z}=1\right]$. Similar to Discussion F.6, we have

$$\begin{aligned}
&\Pr\left[f^\star=1\mid\widehat{Z}=0\right]\\
=\;&\frac{\Pr\left[Z=0\right]\cdot\Pr\left[f^\star=1\mid Z=0\right]}{\Pr\left[\widehat{Z}=0\right]}\times\\
&\Pr\left[\widehat{Z}=0\mid f^\star=1,Z=0\right]\\
&+\frac{\Pr\left[Z=1\right]\cdot\Pr\left[f^\star=1\mid Z=1\right]}{\Pr\left[\widehat{Z}=0\right]}\times\\
&\Pr\left[\widehat{Z}=0\mid f^\star=1,Z=1\right]\\
=\;&\frac{\mu_0\cdot\Pr\left[f^\star=1\mid Z=0\right]}{\pi_{00}\mu_0+\pi_{01}(1-\mu_0)}\times\\
&\Pr\left[\widehat{Z}=0\mid f^\star=1,Z=0\right]\\
&+\frac{\mu_1\cdot\Pr\left[f^\star=1\mid Z=1\right]}{\pi_{00}\mu_0+\pi_{01}(1-\mu_0)}\times\\
&\Pr\left[\widehat{Z}=0\mid f^\star=1,Z=1\right],
\end{aligned}\tag{36}$$

and

$$\begin{aligned}
&\Pr\left[f^\star=1\mid\widehat{Z}=1\right]\\
=\;&\frac{\Pr\left[Z=0\right]\cdot\Pr\left[f^\star=1\mid Z=0\right]}{\Pr\left[\widehat{Z}=1\right]}\times\\
&\Pr\left[\widehat{Z}=1\mid f^\star=1,Z=0\right]\\
&+\frac{\Pr\left[Z=1\right]\cdot\Pr\left[f^\star=1\mid Z=1\right]}{\Pr\left[\widehat{Z}=1\right]}\times\\
&\Pr\left[\widehat{Z}=1\mid f^\star=1,Z=1\right]\\
=\;&\frac{\mu_0\cdot\Pr\left[f^\star=1\mid Z=0\right]}{\pi_{11}(1-\mu_0)+\pi_{20}\mu_0}\times\\
&\Pr\left[\widehat{Z}=1\mid f^\star=1,Z=0\right]\\
&+\frac{\mu_1\cdot\Pr\left[f^\star=1\mid Z=1\right]}{\pi_{11}(1-\mu_0)+\pi_{20}\mu_0}\times\\
&\Pr\left[\widehat{Z}=1\mid f^\star=1,Z=1\right],
\end{aligned}\tag{37}$$

We then analyze the right side of the Equation 36. We take the term $\Pr\left[\widehat{Z}=0\mid f^\star=1,Z=1\right]$ as an example. Let $A=\{a\in[N]:f^\star(s_a)=1,z_a=0\}$. By Assumption 1, we have $|A|\ge\lambda N$. For $i\in A$, let X_i be the random variable where $X_i=1-\widehat{z}_i$. By Definition 2.3, we know

that $\Pr[X_i = 1] = \eta$. Also,

$$\Pr\left[\widehat{Z} = 0 \mid f^\star = 1, Z = 1\right] = \frac{\sum_{i\in A} X_i}{|A|}. \tag{38}$$

Since all X_i ($i \in A$) are independent, we have

$$\begin{aligned} & \Pr\left[\sum_{i\in A} X_i \in (1 \pm \frac{\varepsilon}{8}) \cdot \eta|A|\right] \\ \geq\ & 1 - 2e^{-\frac{\varepsilon^2\eta|A|}{192}} \quad \text{(Chernoff bound)} \\ \geq\ & 1 - 2e^{-\frac{\varepsilon^2\eta\lambda N}{192}}. \quad (|A| \geq \lambda N) \end{aligned} \tag{39}$$

Thus, with probability at least $1 - 2e^{-\frac{\varepsilon^2\eta\lambda N}{192}}$,

$$\begin{aligned} & \Pr\left[\widehat{Z} = 0 \mid f^\star = 1, Z = 1\right] \\ =\ & \frac{\sum_{i\in A} X_i}{|A|} && \text{(Eq. 38)} \\ \in\ & (1 \pm \frac{\varepsilon}{8}) \cdot \frac{\eta|A|}{|A|} && \text{(Ineq. 39)} \\ \in\ & (1 \pm \frac{\varepsilon}{8})\eta. \end{aligned}$$

Consequently, we have

$$\begin{aligned} & \Pr\left[\widehat{Z} = 1 \mid f^\star = 1, Z = 1\right] \\ =\ & 1 - \Pr\left[\widehat{Z} = 0 \mid f^\star = 1, Z = 1\right] \\ \in\ & 1 - (1 \pm \frac{\varepsilon}{8})\eta && \text{(Ineq. 40)} \\ \in\ & (1 \pm \frac{\varepsilon}{8})(1 - \eta) && (\eta < 0.5) \end{aligned}$$

Similarly, we can prove that with probability at least $1 - 4e^{-\frac{\varepsilon^2\eta\lambda N}{192}}$,

- $\pi_{01}, \pi_{20}, \Pr\left[\widehat{Z} = 1 \mid f^\star = 1, Z = 0\right]$, $\Pr\left[\widehat{Z} = 0 \mid f^\star = 1, Z = 1\right] \in (1 \pm \frac{\varepsilon}{8})\eta$;
- $\pi_{00}, \pi_{11}, \Pr\left[\widehat{Z} = 0 \mid f^\star = 1, Z = 0\right]$, $\Pr\left[\widehat{Z} = 1 \mid f^\star = 1, Z = 1\right] \in (1 \pm \frac{\varepsilon}{8})(1 - \eta)$.

Applying these inequalities to Equations 36 and 37, we have that with probability at least $1 - 4e^{-\frac{\varepsilon^2\eta\lambda N}{192}}$,

$$\begin{aligned} & \frac{\Pr\left[f^\star = 1 \mid \widehat{Z} = 0\right]}{\Pr\left[f^\star = 1 \mid \widehat{Z} = 1\right]} \\ \in\ & (1 \pm \varepsilon) \cdot \frac{\eta\mu_0 + (1-\eta)(1-\mu_0)}{(1-\eta)\mu_0 + \eta(1-\mu_0)} \times \\ & \frac{(1-\eta)\mu_0\gamma(f^\star, S) + \eta(1-\mu_0)}{\eta\mu_0\gamma(f^\star, S) + (1-\eta)(1-\mu_0)} \\ \in\ & (1 \pm \varepsilon) \cdot \Gamma, \end{aligned}$$

and

$$\frac{\Pr\left[f^\star = 1 \mid \widehat{Z} = 1\right]}{\Pr\left[f^\star = 1 \mid \widehat{Z} = 0\right]} \in (1 \pm \varepsilon) \cdot \frac{1}{\Gamma}.$$

By the definition of $\gamma(f^\star, \widehat{S})$, we complete the proof. □