A. Omitted Proofs

Proof of Prop. 1. Let $1-\rho = F_a^{Z=1}(\theta_a) = F_b^{Z=1}(\theta_b)$, which are equal by assumption. We then have that

$$\begin{aligned}
\epsilon_{a,b} &= (1 - F_a^{T=1}(\theta_a)) - (1 - F_b^{T=1}(\theta_b)) \\
&= F_b^{T=1}(\theta_b) - F_a^{T=1}(\theta_a) \\
&= ((1-\rho) - F_a^{T=1}(\theta_a)) - ((1-\rho) - F_b^{T=1}(\theta_b)) \\
&= (F_a^{Z=1}(\theta_a) - F_a^{T=1}(\theta_a)) - (F_b^{Z=1}(\theta_b) - F_b^{T=1}(\theta_b)) \\
&= \Delta_a(\theta_a) - \Delta_b(\theta_b).
\end{aligned}$$

□

Proof of Prop. 2. Let $\hat{Y} = \mathbb{I}[\hat{R} > \theta_A]$ be any derived equal opportunity classifier. By Prop. 1, $\epsilon_{a,b} = \Delta_a(\theta_a) - \Delta_b(\theta_b)$ and by assumption $\Delta_a(\theta_a) \geq 0$ while $\Delta_b(\theta_b) \leq 0$.

By assumption, at least one of $F_a^{Z=1} \neq F_a^{T=1}$ and $F_b^{Z=1} \neq F_b^{T=1}$ holds. Suppose that $F_a^{Z=1} \neq F_a^{T=1}$. Then there is θ_a such that $F_a^{Z=1}(\theta_a) \neq F_a^{T=1}(\theta_a)$. Then $\Delta_a(\theta_a) > 0$. Letting $\theta_b = (F_b^{Z=1})^{-1}(F_a^{Z=1}(\theta_a))$, we get that $\hat{Y} = \mathbb{I}[\hat{R} > \theta_A]$ is a derived equal opportunity classifier with $\epsilon_{a,b} > 0$. If instead $F_b^{Z=1} \neq F_b^{T=1}$ then we'd have θ_b with $\Delta_b(\theta_b) < 0$ and we'd let $\theta_a = (F_a^{Z=1})^{-1}(F_b^{Z=1}(\theta_b))$. □

Proof of Prop. 3. Let $\hat{Y} = \mathbb{I}[\hat{R} > \theta_A]$ be any nontrivial derived equal opportunity classifier. By Prop. 1, $\epsilon_{a,b} = \Delta_a(\theta_a) - \Delta_b(\theta_b)$ and by assumption $\Delta_a(\theta_a) > 0$ and $\Delta_b(\theta_b) < 0$. □

Proof of Prop. 4. Self-evident from Prop. 1. □

Proof of Prop. 5. Self-evident from Prop. 1 after noting that $F_a^{Z=1} \succeq F_b^{Z=1}$ necessarily implies that $\theta_a \geq \theta_b$. □

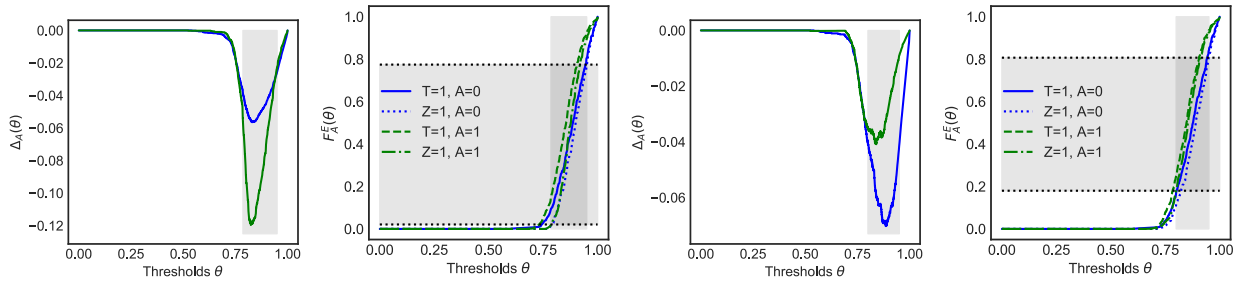
Proof of Prop. 6. We have that

$$\begin{aligned}
\mathbb{P}(\hat{Y} = \hat{y}, Y = y, A = a \mid T = 1) &= \mathbb{E}[\mathbb{P}(\hat{Y} = \hat{y}, Y = y, T = 1 \mid X, A)\mathbb{I}[A = a]]/\mathbb{P}(T = 1) \\
&= \mathbb{E}[\mathbb{P}(\hat{Y} = \hat{y}, Y = y \mid X, A, T = 1)\mathbb{P}(T = 1 \mid X, A)\mathbb{I}[A = a]]/\mathbb{P}(T = 1) \\
&= \mathbb{E}[\mathbb{P}(\hat{Y} = \hat{y} \mid X, A, T = 1)\mathbb{P}(Y = y \mid X, A, T = 1)\mathbb{P}(T = 1 \mid X, A)\mathbb{I}[A = a]]/\mathbb{P}(T = 1) \\
&= \mathbb{E}[\mathbb{P}(\hat{Y} = \hat{y} \mid X, A)\mathbb{P}(Y = y \mid X, A)\mathbb{P}(T = 1 \mid X, A)\mathbb{I}[A = a]]/\mathbb{P}(T = 1) \\
&= \mathbb{E}[\mathbb{P}(\hat{Y} = \hat{y} \mid X, A, Z = 1)\mathbb{P}(Y = y \mid X, A, Z = 1)\mathbb{P}(T = 1 \mid X, A)\mathbb{I}[A = a]]/\mathbb{P}(T = 1) \\
&= \mathbb{E}[\mathbb{P}(\hat{Y} = \hat{y}, Y = y \mid X, A, Z = 1)\mathbb{P}(T = 1 \mid X, A)\mathbb{I}[A = a]]/\mathbb{P}(T = 1) \\
&= \mathbb{E}[\mathbb{P}(\hat{Y} = \hat{y}, Y = y, Z = 1 \mid X, A)\tfrac{\mathbb{P}(T=1|X,A)}{\mathbb{P}(Z=1|X,A)}\mathbb{I}[A = a]]/\mathbb{P}(T = 1) \\
&= \mathbb{E}[\mathbb{I}[\hat{Y} = \hat{y}, Y = y, Z = 1, A = a]p(X, A)]/\mathbb{P}(T = 1).
\end{aligned}$$

The rest follows by Bayes law. □

B. Weak Disparate Benefit of the Doubt in Credit Card Data

We consider the data from Greene (1992), which contains individual-level data on credit card acceptance, default on payments or not (if accepted), information about individual income, derogatory reports on accounts, and other features of creditworthiness such as self-employment indicators. The dataset also includes age, which has been a concern regarding the fairness of credit scoring models (Board of Governors of the Federal Reserve System, 1997). We construct a protected class by defining $A = \mathbb{I}[X_{age} < F_{X_{age}}^{-1}(0.5)]$ as the indicator for being below the median age (31.67). To illustrate how the direction of disparities can change depending on the $Z = 1$ policy we consider two scenarios. In both, we consider the target population $T = 1$ to actually consist of all accepted applicants (rather than all applicants). First, we consider $Z = \mathbb{I}[T = 1, X_i > F_{X_{inc}}^{-1}(.1)]$ where X_{inc} is income so that we *further* censor the lowest-income individuals from the available data. Second, we also consider $Z = \mathbb{I}[T = 1, X_i > F_{X_{inc\_per}}^{-1}(.1)]$ where X_{inc_per} is the income per dependent. In $T = 1$, income is somewhat correlated with membership in the protected class (being young), with a correlation coefficient $\rho = -0.32$. However, income per dependent is very weakly correlated with being young, with a correlation coefficient



(a) Δ_a, censoring on X_{inc} (b) F_A^E, censoring on X_{inc} (c) Δ_a, censoring on $X_{\text{inc\_per}}$ (d) F_A^E, censoring on $X_{\text{inc\_per}}$

Figure 6: Illustration of score disparities between censored and full training data. The dataset is from credit card applications, where we additionally censor the population of accepted cardholders by income or income per dependent.

$\rho = 0.03$. This is intuitive as the correlation of greater income by age is canceled out by the correlation of greater household size with age.

In Fig. 6, we plot the FNRs $F_A^{T=1}, F_A^{Z=1}$ in training and in target and the discrepancies Δ_A between them for both censoring cases. First we study the case of censoring on X_{inc}. By inspecting the FNRs in Fig. 6b, we see that $F_0^{Z=1} \succeq F_1^{Z=1}$. Therefore, we can apply Prop. 5. In Fig. 6a we shade a region of thresholds that satisfies eq. (4). In particular, because of the relaxation for disparately endowed groups, we can extend the region farther right than would be possible under eq. (3). In Fig. 6b we shade the corresponding ranges of FNRs that, per Prop. 5, would lead to spuriously-fairness-adjusted classifiers that actually induce an inequity of opportunity disadvantages the younger group $A = 1$.

Next we study the case of censoring on $X_{\text{inc\_per}}$. We first note that the disparate benefit of the doubt induced is now going in the *opposite* direction (Fig. 6c) – so the spuriously-fairness-adjusted classifiers will disadvantage the older group rather than the younger group. Although we have the same ordering of FNRs as before, $F_0^{Z=1} \succeq F_1^{Z=1}$ (Fig. 6d), the ordering of Δ_A's is opposite and therefore eq. (4) does not offer a relaxation over eq. (3). We therefore apply the standard weak disparate benefit of the doubt. In Fig. 6c we shade a region of thresholds that satisfies eq. (3). In Fig. 6b we shade the corresponding ranges of FNRs, per Prop. 4, would lead to spuriously-fairness-adjusted classifiers that actually induce an inequity of opportunity disadvantages the older group $A = 0$.

C. Residual Unfairness Under MAR

In this section we study several implications for residual unfairness under the MAR assumption. First we show that in rare cases prejudice, if applied purely and directly on protected attribute alone, can actually be perfectly corrected for based on training-data-based fairness adjustment. Second we study how disparities in importance weights can be used to characterize the presence residual unfairness.

C.1. No Residual Unfairness if Inclusion Depends Only on Protected Attributes

We next show that if the censoring mechanism depends *only on the protected attribute* A that is to be adjusted for fairness, then in fact there will be no residual unfairness and the true positive rates remain the same in the target population as in the training population.

Proposition 7. *Suppose* $\mathbb{P}(Z = 1 \mid X, A) = \mathbb{P}(Z = 1 \mid A)$ *and* $\mathbb{P}(T = 1 \mid X, A) = \mathbb{P}(T = 1 \mid A)$. *Then, under Assumption 1, inequity of opportunity on training is the same as inequity of opportunity on target, i.e.,* $\epsilon_{a,b}^{Z=1} = \epsilon_{a,b}^{T=1}$.

Thus, residual unfairness occurs only when biased inclusion is heterogeneous based on covariates X, *e.g.*, as in the case of SQF where the application of stops differs based on both precinct and race. This is typical of the application areas where fairness is of concern: censoring is usually disparate in large part via proxies for protected attributes.

C.2. Characterizing Residual Unfairness in Terms of Propensity Ratios Disparities

Under Assumption 1, we can characterize residual unfairness in terms of the reweighting estimates.

Proposition 8. $\Pr[\hat{Y}=1 \mid {}^{Y=1}_{A=a,T=1}] > \Pr[\hat{Y}=1 \mid {}^{Y=1}_{A=a,Z=1}]$ *if and only if*

$$\mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=a}_{Y=1,\hat{Y}=0}\right] < \mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=a}_{Y=1,\hat{Y}=1}\right]$$

This says that the TPR will be in actuality higher in the target population if the average ratio weights in the group of true positives included in the dataset is greater than in the group of false negatives included. Intuitively, the predictor will be more accurate in the target population if, due to censoring, positive examples that the predictor will be correct on were less likely to appear in the training data.

We can use to characterize exactly when a classifier that satisfies equal opportunity on training will have residual unfairness on target.

Corollary 9. *Suppose $\hat{Y}$ satisfies equal opportunity wrt $Z=1$. Then $\epsilon^{T=1}_{a,b} > 0$ if and only if*

$$\frac{\mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=a}_{Y=1,\hat{Y}=0}\right]}{\mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=b}_{Y=1,\hat{Y}=0}\right]} < \frac{\mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=a}_{Y=1,\hat{Y}=1}\right]}{\mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=b}_{Y=1,\hat{Y}=1}\right]}$$

This characterization follows from Prop. 8 and the fact that the true positive rates in training are the same under equality of opportunity.

C.3. Proofs

Proof of Prop. 7. The result follows from applying Prop. 6 with $\Pr(Z=1 \mid X,A) = \Pr(Z=1 \mid A)$ and iterating the expectation over X:

$$\begin{aligned}
&\mathbb{P}(\hat{Y}=1 \mid Y=y, A=a, T=1) \\
&= \frac{\mathbb{E}[\frac{1}{\Pr[Z=1|A]}\mathbb{E}[\mathbb{E}[\mathbb{I}\{\hat{Y}=1, Y=1, Z=1\} \mid X]]]\mathbb{I}\{A=a\}]}{\sum_{\hat{y}\in\{0,1\}}\mathbb{E}[\frac{1}{\Pr[Z=1|A]}\mathbb{E}[\mathbb{E}[\mathbb{I}\{\hat{Y}=\hat{y}, Y=1, Z=1\} \mid X]]\mathbb{I}\{A=a\}]} \\
&= \frac{\mathbb{E}[\frac{1}{\Pr[Z=1|A]}]\mathbb{E}[\mathbb{E}[\mathbb{E}[\mathbb{I}\{\hat{Y}=1, Y=1, Z=1\} \mid X]]]\mathbb{I}\{A=a\}]}{\mathbb{E}[\frac{1}{\Pr[Z=1|A]}]\sum_{\hat{y}\in\{0,1\}}\mathbb{E}[\mathbb{E}[\mathbb{E}[\mathbb{I}\{\hat{Y}=\hat{y}, Y=1, Z=1\} \mid X]]\mathbb{I}\{A=a\}]} \\
&= \mathbb{P}(\hat{Y}=1 \mid Y=y, A=a, Z=1)
\end{aligned}$$

□

Proof of Prop. 8.

$$\begin{aligned}
\Delta^{TPR}_a &= \frac{\mathrm{TPR}^*}{\mathrm{TPR}} = \frac{\Pr[\hat{Y}=1 \mid Y=1, A=a, T=1]}{\Pr[\hat{Y}=1 \mid Y=1, A=a, Z=1]} \\
&= \frac{\frac{\Pr[\hat{Y}=1, y=1|A=a,T=1]/\Pr[Y-1|A=a,T=1]}{\sum_{\hat{y}\in\{0,1\}}\Pr[\hat{Y}=1,y=1|A=a,T=1]}}{\frac{\Pr[\hat{Y}=1, y=1|A=a,T=1]/\Pr[Y=1|A=a,Z=1]}{\sum_{\hat{y}\in\{0,1\}}\Pr[\hat{Y}=1,y=1|A=a,Z=1]}} = \frac{\frac{1}{1+\frac{\Pr[Y=1|A=a,T=1]}{\Pr[\hat{Y}=1,y=1|A=a,T=1]}}}{\frac{1}{1+\frac{\Pr[Y=1|A=a,Z=1]}{\Pr[\hat{Y}=1,y=1|A=a,Z=1]}}}
\end{aligned}$$

So

$$\Delta^{TPR}_a > 1 \iff \frac{\Pr[Y=1,\hat{Y}=0 \mid A=a,T=1]}{\Pr[Y=1,\hat{Y}=1 \mid A=a,T=1]} < \frac{\Pr[Y=1,\hat{Y}=0 \mid A=a,Z=1]}{\Pr[Y=1,\hat{Y}=1 \mid A=a,Z=1]}$$

We can apply Prop. 6:

$$\Delta^{TPR}_a > 1 \iff \frac{\Pr[Y=1,\hat{Y}=1 \mid A=a,Z=1]}{\mathbb{E}[\mathbb{I}[Y=1,\hat{Y}=1]p(X,A) \mid {}^{Z=1}_{A=a}]} < \frac{\Pr[Y=1,\hat{Y}=0 \mid A=a,Z=1]}{\mathbb{E}[\mathbb{I}[Y=1,\hat{Y}=0]p(X,A) \mid {}^{Z=1}_{A=a}]}$$

and the identification that $\mathbb{E}[\mathbb{I}\{Y=1,\hat{Y}=0\}p(X,A) \mid {}^{Z=1}_{A=a}] = \mathbb{E}\left[p(x,A) \mid {}^{Z=1,A=a}_{Y=1,\hat{Y}=0}\right]\Pr[Y=1,\hat{Y}=0 \mid {}^{Z=1}_{A=a}]$:

$$\Delta^{TPR}_a > 1 \iff \frac{1}{\mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=a}_{Y=1,\hat{Y}=1}\right]} < \frac{1}{\mathbb{E}\left[p(X,A) \mid {}^{Z=1,A=a}_{Y=1,\hat{Y}=0}\right]}$$

□

D. Information on Stop, Question and Frisk

Stop, Question, and Frisk is a program which allows police officers to stop citizens in public, question, and possibly search them, under reasonable suspicion of a crime but not enough probable cause for an arrest (Goel et al., 2017). Around 600,000 people were stopped in 2011, and around 90% of stops led to no evidence of a crime (Keefe, 2011). Each officer is required to file an individual report after a stop detailing individual characteristics (including physical attributes of the suspect and location) and reasons for the stop, leading to relatively rich context about each individual decision (NYCLU, 2017).

SQF was studied by statistical researchers and adjudicated in the court case Floyd v. City of New York for discrimination on the basis of race and national origin. The program has been controversial since the demographic makeup of stops in the data systematically misrepresents the population of NYC at large due to disparate patrol levels and implementation of SQF by NYPD precinct, which correlates with demographics, as well as the potential for racial biases at the individual level. These demographic imbalances have been studied and analyzed judicially, discussed alongside evidence of administrative and structural deviation in application of SQF practices (Goel et al., 2017; Gelman et al., 2007). Some of the covariates themselves may reflect proxy indicators for discrimination. In the SQF data, for example, recorded reasons for stop include whether the suspect was actually engaging in a crime, was a known criminal, or exhibited "furtive movement". The potential for some of these reasons to be proxies for discrimination was noted by Judge Scheindlin in the court case of Floyd, et al. v. City of New York.