A Appendix

A.1 Proofs

A.1.1 Explanation of Assumption 1

Assumption 1: For $i, j \in \{0, 1\}$, the classifiers $\widehat{Y} = f(X)$ and $\widehat{A} = h(X)$ satisfy

$$\frac{U_i}{\hat{r}_{i,j}} \leq \widehat{\alpha}_{i,j} \leq 1 - \frac{U_i}{\hat{r}_{i,j}}.$$

We now expand on the implications of this assumptions. Recall that $U_i = \mathbb{P}(\widehat{A} = i, A \neq i)$, $\hat{r}_{i,j} = \mathbb{P}(\widehat{A} = i, Y = j)$, and $\widehat{\alpha}_{i,j} = \mathbb{P}(\widehat{Y} = 1 \mid \widehat{A} = i, Y = j)$. Thus Assumption 1 states,

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = j) \leq \mathbb{P}(\widehat{A} = i, Y = j) - \mathbb{P}(\widehat{A} = i, A \neq i). \tag{10}$$

Immediately, it is clear that Eq. (10) implies

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{A} = i, Y = j) - \mathbb{P}(\widehat{A} = i, A \neq i) \tag{11}$$

$$\implies \mathbb{P}(\widehat{A} = i, A \neq i) \leq \frac{1}{2}\mathbb{P}(\widehat{A} = i, Y = j) \tag{12}$$

Now, the left inequality of Eq. (10) states

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = j) \tag{13}$$

and the right inequality of Eq. (10) states

$$\mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = j) \leq \mathbb{P}(\widehat{A} = i, Y = j) - \mathbb{P}(\widehat{A} = i, A \neq i)$$

which implies

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{A} = i, Y = j) - \mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = j) \tag{14}$$

$$= \mathbb{P}(\widehat{Y} = 0, \widehat{A} = i, Y = j) \tag{15}$$

If $j = 1$, then this implies

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = 1) \tag{16}$$

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = 0, \widehat{A} = i, Y = 1) \tag{17}$$

and if $j = 0$,

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = 0) \tag{18}$$

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = 0, \widehat{A} = i, Y = 0) \tag{19}$$

Any reasonable classifier $\widehat{Y}$ would have the properties

$$\mathbb{P}(\widehat{Y} = 0, \widehat{A} = i, Y = 1) \leq \mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = 1) \tag{20}$$

$$\mathbb{P}(\widehat{Y} = 1, \widehat{A} = i, Y = 0) \leq \mathbb{P}(\widehat{Y} = 0, \widehat{A} = i, Y = 0) \tag{21}$$

Thus, Assumption 1 is met when

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = j, \widehat{A} = i, Y \neq j) \tag{22}$$

A.1.2 Proof of Theorem 1

We only prove the result for $|\Delta_{\text{TPR}}(f)|$ as the proof for $|\Delta_{\text{FPR}}(f)|$ is completely analogous.

Proof. The rules of conditional probability and the law of total probability allow us to decompose $\alpha_{1,1}$ and $\alpha_{0,1}$ in the following manner,

$$\alpha_{1,1} = \mathbb{P}(\hat{Y} = 1 \mid A = 1, Y = 1) \tag{23}$$

$$= \frac{\mathbb{P}(\hat{Y} = 1, A = 1, Y = 1)}{\mathbb{P}(A = 1, Y = 1)} \tag{24}$$

$$= \frac{\sum_{i\in\{0,1\}} \mathbb{P}(\hat{Y} = 1, A = 1, Y = 1, \hat{A} = i)}{\sum_{j\in\{0,1\}} \sum_{i\in\{0,1\}} \mathbb{P}(\hat{Y} = j, A = 1, Y = 1, \hat{A} = i)} \tag{25}$$

$$= \frac{\sum_{i\in\{0,1\}} \mathbb{P}(\hat{Y} = 1, A = 1, Y = 1 \mid \hat{A} = i) \cdot \mathbb{P}(\hat{A} = i)}{\sum_{j\in\{0,1\}} \sum_{i\in\{0,1\}} \mathbb{P}(\hat{Y} = j, A = 1, Y = 1 \mid \hat{A} = i) \cdot \mathbb{P}(\hat{A} = i)} \tag{26}$$

and

$$\alpha_{0,1} = \mathbb{P}(\hat{Y} = 1 \mid A = 0, Y = 1) \tag{27}$$

$$= \frac{\mathbb{P}(\hat{Y} = 1, A = 0, Y = 1)}{\mathbb{P}(A = 0, Y = 1)} \tag{28}$$

$$= \frac{\mathbb{P}(\hat{Y} = 1, Y = 1) - \mathbb{P}(\hat{Y} = 1, A = 1, Y = 1)}{\mathbb{P}(Y = 1) - \mathbb{P}(A = 1, Y = 1)} \tag{29}$$

$$= \frac{\mathbb{P}(\hat{Y} = 1, Y = 1) - \left[\sum_{i\in\{0,1\}} \mathbb{P}(\hat{Y} = 1, A = 1, Y = 1 \mid \hat{A} = i) \cdot \mathbb{P}(\hat{A} = i)\right]}{\mathbb{P}(Y = 1) - \left[\sum_{j\in\{0,1\}} \sum_{i\in\{0,1\}} \mathbb{P}(\hat{Y} = j, A = 1, Y = 1 \mid \hat{A} = i) \cdot \mathbb{P}(\hat{A} = i)\right]} \tag{30}$$

Therefore, $\Delta_{\text{TPR}}(f) = \alpha_{1,1} - \alpha_{0,1}$ is a function of the four probabilities given by

$$\mathbb{P}(\hat{Y} = j, A = 1, Y = 1 \mid \hat{A} = i) \tag{31}$$

which are unidentifiable in a demographically scarce regime and therefore not computable.

The Fréchet inequalities tell us that for $i, j \in \{0, 1\}$

$$\mathbb{P}(\hat{Y} = j, A = 1, Y = 1 \mid \hat{A} = i) \geq \max\{\mathbb{P}(\hat{Y} = j, Y = 1 \mid \hat{A} = i) - \mathbb{P}(A = 0 \mid \hat{A} = i), 0\} \tag{32}$$

$$\mathbb{P}(\hat{Y} = j, A = 1, Y = 1 \mid \hat{A} = i) \leq \min\{\mathbb{P}(\hat{Y} = j, Y = 1 \mid \hat{A} = i), \mathbb{P}(A = 1 \mid \hat{A} = i)\}. \tag{33}$$

Observe that, $\Delta_{\text{TPR}}(f)$ is an increasing function with respect to the two probabilities

$$\mathbb{P}(\hat{Y} = 1, A = 1, Y = 1 \mid \hat{A} = i) \tag{34}$$

and a decreasing one with respect to the two probabilities,

$$\mathbb{P}(\hat{Y} = 0, A = 1, Y = 1 \mid \hat{A} = i). \tag{35}$$

As a result, $\Delta_{\text{TPR}}(f)$ is maximal when $\mathbb{P}(\hat{Y} = 1, A = 1, Y = 1 \mid \hat{A} = i)$ achieve their maximum values and $\mathbb{P}(\hat{Y} = 0, A = 1, Y = 1 \mid \hat{A} = i)$ achieve their minimum values. On the other hand, $\Delta_{\text{TPR}}(f)$ is minimal when $\mathbb{P}(\hat{Y} = 1, A = 1, Y = 1 \mid \hat{A} = i)$ achieve their minimum values and $\mathbb{P}(\hat{Y} = 0, A = 1, Y = 1, \mid \hat{A} = i)$ achieve their maximum values. With these facts, we now provide the upper bound. Recall from Appendix A.1.1 that Assumption 1 implies

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \frac{1}{2}\mathbb{P}(\widehat{A} = i, Y = j) \tag{36}$$

$$\mathbb{P}(\widehat{A} = i, A \neq i) \leq \mathbb{P}(\widehat{Y} = j, \widehat{A} = i, Y \neq j) \leq \mathbb{P}(\widehat{Y} = j, \widehat{A} = i, Y = j) \tag{37}$$

$$\tag{38}$$

With Assumption 1 we first provide the values of $\min\left[\mathbb{P}(\hat{Y}=0, A=1, Y=1, \mid \hat{A}=i)\right]$. First,

$$\mathbb{P}(\hat{Y}=0, Y=1, \mid \hat{A}=1) - \mathbb{P}(A=0 \mid \hat{A}=1) = \frac{\mathbb{P}(\hat{Y}=0, \hat{A}=1, Y=1)}{P(\hat{A}=1)} - \frac{U_1}{\mathbb{P}(\hat{A}=1)} \tag{39}$$

$$\geq 0 \tag{40}$$

because $\mathbb{P}(\hat{Y}=0, \hat{A}=1, Y=1) - U_1 \geq 0$. Second,

$$\mathbb{P}(\hat{Y}=0, Y=1, \mid \hat{A}=0) - \mathbb{P}(A=0 \mid \hat{A}=0) = \frac{\mathbb{P}(\hat{Y}=0, \hat{A}=0, Y=1)}{P(\hat{A}=0)} - \frac{\mathbb{P}(A=0, \hat{A}=0)}{\mathbb{P}(\hat{A}=0)} \tag{41}$$

$$= \frac{\mathbb{P}(\hat{Y}=0, \hat{A}=0, Y=1)}{P(\hat{A}=0)} - \frac{\mathbb{P}(\hat{A}=0) - U_0}{\mathbb{P}(\hat{A}=0)} \tag{42}$$

$$= \frac{\mathbb{P}(\hat{Y}=0, \hat{A}=0, Y=1)}{P(\hat{A}=0)} - \frac{\mathbb{P}(\hat{A}=0) - U_0}{\mathbb{P}(\hat{A}=0)} \tag{43}$$

$$= \frac{\mathbb{P}(\hat{Y}=0, \hat{A}=0, Y=1)}{P(\hat{A}=0)} + \frac{U_0}{\mathbb{P}(\hat{A}=0)} - 1 \tag{44}$$

Now note that,

$$\mathbb{P}(\hat{Y}=0, \hat{A}=0, Y=1) = \mathbb{P}(\hat{A}=0, Y=1) - \mathbb{P}(\hat{Y}=1, \hat{A}=0, Y=1) \tag{45}$$

$$\leq \mathbb{P}(\hat{A}=0, Y=1) - U_0 \tag{46}$$

where the second equality is due to Assumption 1. As a result

$$\frac{\mathbb{P}(\hat{Y}=0, \hat{A}=0, Y=1)}{P(\hat{A}=0)} + \frac{U_0}{\mathbb{P}(\hat{A}=0)} - 1 \leq \frac{\mathbb{P}(\hat{A}=0, Y=1) - U_0}{P(\hat{A}=0)} + \frac{U_0}{\mathbb{P}(\hat{A}=0)} - 1 \tag{47}$$

$$= \frac{\mathbb{P}(\hat{A}=0, Y=1)}{P(\hat{A}=0)} - 1 \leq 0 \tag{48}$$

Therefore,

$$\min\left[\mathbb{P}(\hat{Y}=0, A=1, Y=1 \mid \hat{A}=1)\right] = \mathbb{P}(\hat{Y}=0, Y=1, \mid \hat{A}=1) - \mathbb{P}(A=0 \mid \hat{A}=1) \tag{49}$$

$$\min\left[\mathbb{P}(\hat{Y}=0, A=1, Y=1 \mid \hat{A}=0)\right] = 0 \tag{50}$$

Now we provide the values of $\max\left[\mathbb{P}(\hat{Y}=1, A=1, Y=1, \mid \hat{A}=i)\right]$. First,

$$\mathbb{P}(A=1 \mid \hat{A}=1) = \frac{\mathbb{P}(A=1, \hat{A}=1)}{\mathbb{P}(\hat{A}=1)} \tag{51}$$

$$= \frac{\mathbb{P}(\hat{A}=1) - U_1}{\mathbb{P}(\hat{A}=1)} \tag{52}$$

$$\geq \frac{\mathbb{P}(\hat{A}=1, Y=1) - U_1}{\mathbb{P}(\hat{A}=1)} \tag{53}$$

$$\geq \frac{\mathbb{P}(\hat{A}=1, Y=1) - \mathbb{P}(\hat{Y}=0, \hat{A}=1, Y=1)}{\mathbb{P}(\hat{A}=1)} \tag{54}$$

$$= \frac{\mathbb{P}(\hat{Y}=1, \hat{A}=1, Y=1)}{\mathbb{P}(\hat{A}=1)} = \mathbb{P}(\hat{Y}=1, Y=1 \mid \hat{A}=1) \tag{55}$$

Second,

$$\mathbb{P}(A=1 \mid \hat{A}=0) = \frac{\mathbb{P}(A=1, \hat{A}=0)}{\mathbb{P}(\hat{A}=0)} \tag{56}$$

$$\leq \frac{\mathbb{P}(\hat{Y}=1, \hat{A}=0, Y=1)}{\mathbb{P}(\hat{A}=0)} = \mathbb{P}(\hat{Y}=1, Y=1 \mid \hat{A}=0) \tag{57}$$

Therefore,

$$\max\left[\mathbb{P}(\hat{Y}=1, A=1, Y=1 \mid \hat{A}=1)\right] = \mathbb{P}(\hat{Y}=1, Y=1 \mid \hat{A}=1) \tag{58}$$

$$\max\left[\mathbb{P}(\hat{Y}=1, A=1, Y=1 \mid \hat{A}=0)\right] = \mathbb{P}(A=1 \mid \hat{A}=0) \tag{59}$$

Plugging these 4 values into Δ_{TPR} will yield the upper bound,

$$B_1 + C_{0,1} = \frac{\hat{r}_{1,1}}{\hat{r}_{1,1} + \Delta U}\widehat{\alpha}_{1,1} - \frac{\hat{r}_{0,1}}{\hat{r}_{0,1} - \Delta U}\widehat{\alpha}_{0,1} + U_0\left(\frac{1}{\hat{r}_{1,1} + \Delta U} + \frac{1}{\hat{r}_{0,1} - \Delta U}\right) \tag{60}$$

One can similarly use the assumptions to derive the lower bound,

$$B_1 - C_{1,1} = \frac{\hat{r}_{1,1}}{\hat{r}_{1,1} + \Delta U}\widehat{\alpha}_{1,1} - \frac{\hat{r}_{0,1}}{\hat{r}_{0,1} - \Delta U}\widehat{\alpha}_{0,1} - U_1\left(\frac{1}{\hat{r}_{1,1} + \Delta U} + \frac{1}{\hat{r}_{0,1} - \Delta U}\right) \tag{61}$$

and thus $|\Delta_{\text{TPR}}| \leq \max\{|B_1 + C_{0,1}|, |B_1 - C_{1,1}|\}$. One can use same arguments to derive the upper bound for $|\Delta_{\text{FPR}}|$. □

A.1.3 Proof of Theorem 2

We prove the result for $|\Delta_{\text{TPR}}(f)|$. We first start by proving the existence part of the theorem.

Let $\widehat{A} = h(X)$ be a sensitive attribute classifier with errors U_0 and U_1 that produces rates $\hat{r}_{i,j} = \mathbb{P}(\widehat{A} = i, Y = j)$. Let $\mathcal{F}$ be the set of classifiers for Y such that $\forall f \in \mathcal{F}$, f and h satisfy Assumption 1. Consider any $f \in \mathcal{F}$ with group conditional probabilities, $\widehat{\alpha}_{i,j} = \mathbb{P}(\widehat{Y} = 1 \mid \widehat{A} = i, Y = j)$. Since we are only proving the result for $|\Delta_{\text{TPR}}(f)|$, set $j = 1$. Consider the xy plane, with the x-axis being $\widehat{\alpha}_{0,1}$ and the y-axis being $\widehat{\alpha}_{1,1}$. We know,

$$\frac{U_i}{\hat{r}_{i,1}} \leq \widehat{\alpha}_{i,1} \leq 1 - \frac{U_i}{\hat{r}_{i,1}} \tag{62}$$

which implies

$$\frac{U_i}{\hat{r}_{i,1}} \leq \frac{1}{2}. \tag{63}$$

The two equations above define a rectangular region in the xy plane with a center $(\frac{1}{2}, \frac{1}{2})$, meaning any classifier $f \in \mathcal{F}$, has $\widehat{\alpha}_{i,j}$ that are in this region.

Now, denote $\bar{\mathcal{F}}$ to be a the set of classifiers for Y, with group conditional probabilities $\underline{\widehat{\alpha}}_{i,1}$, that satisfy the condition,

$$\frac{\hat{r}_{0,1}}{\hat{r}_{0,1} - \Delta U}\underline{\widehat{\alpha}}_{0,1} - \frac{\hat{r}_{1,1}}{\hat{r}_{1,1} + \Delta U}\underline{\widehat{\alpha}}_{1,1} = \frac{\Delta U}{2}\left(\frac{1}{\hat{r}_{1,1} + \Delta U} + \frac{1}{\hat{r}_{0,1} - \Delta U}\right). \tag{64}$$

This condition defines a line in the xy plane meaning any classifier in $\bar{\mathcal{F}}$ has $\underline{\widehat{\alpha}}_{i,1}$ that are on this line. Now observe that the classifier $\bar{f} \in \bar{\mathcal{F}}$ with $\underline{\widehat{\alpha}}_{i,1} = \frac{1}{2}$, satisfy the above condition because,

$$\begin{aligned}
\frac{\hat{r}_{0,1}}{\hat{r}_{0,1} - \Delta U}\left(\frac{1}{2}\right) - \frac{\hat{r}_{1,1}}{\hat{r}_{1,1} + \Delta U}\left(\frac{1}{2}\right) &= \frac{1}{2}\left(\frac{\hat{r}_{0,1}}{\hat{r}_{0,1} - \Delta U} - \frac{\hat{r}_{1,1}}{\hat{r}_{1,1} + \Delta U}\right) && (65)\\
&= \frac{1}{2}\left(\frac{\hat{r}_{0,1} - \Delta U + \Delta U}{\hat{r}_{0,1} - \Delta U} - \frac{\hat{r}_{1,1} + \Delta U - \Delta U}{\hat{r}_{1,1} + \Delta U}\right) && (66)\\
&= \frac{1}{2}\left(1 + \frac{\Delta U}{\hat{r}_{0,1} - \Delta U} - 1 + \frac{\Delta U}{\hat{r}_{1,1} + \Delta U}\right) && (67)\\
&= \frac{\Delta U}{2}\left(\frac{1}{\hat{r}_{1,1} + \Delta U} + \frac{1}{\hat{r}_{0,1} - \Delta U}\right) && (68)
\end{aligned}$$

This implies that the line defined by Eq. (64) intersects the rectangular region that Assumption 1 defines. As a result, $\mathcal{F} \cap \bar{\mathcal{F}}$ is not empty, meaning there exists a classifier $\bar{f} \in \mathcal{F}$ with group conditional probabilities $\underline{\widehat{\alpha}}_{i,1}$ that also satisfies the condition,

$$\frac{\hat{r}_{0,1}}{\hat{r}_{0,1} - \Delta U}\underline{\widehat{\alpha}}_{0,1} - \frac{\hat{r}_{1,1}}{\hat{r}_{1,1} + \Delta U}\underline{\widehat{\alpha}}_{1,1} = \frac{\Delta U}{2}\left(\frac{1}{\hat{r}_{1,1} + \Delta U} + \frac{1}{\hat{r}_{0,1} - \Delta U}\right). \tag{69}$$

Now we prove that such a classifier has minimal bounds. Theorem 1 tells us that for $f \in \mathcal{F}$

$$|\Delta_{\text{TPR}}(f)| \leq B_{\text{TPR}}(f) \triangleq \max\{|B_1 + C_{0,1}|, |B_1 - C_{1,1}|\}$$

Note that B_1 is linear in $\widehat{\alpha}_{1,1}$ and $\widehat{\alpha}_{0,1}$ and that $C_{0,1}$ and $C_{1,1}$ are constants such that $B_1 + C_{0,1} \geq B_1 - C_{1,1}$ simply because $B_1 + C_{0,1}$ is the upper bound for Δ_{TPR} and $B_1 - C_{0,1}$ is the lower bound. Since these bounds are affine functions shifted by a constant, then $\min\max\{|B_1 + C_{0,1}|, |B_1 - C_{1,1}|\}$ necessarily occurs when

$$B_1 + C_{0,1} = -B_1 - C_{1,1} \tag{70}$$

meaning

$$2B_1 = -(C_{1,1} + C_{0,1}) \tag{71}$$

have minimal upper bounds on $|\Delta_{\text{TPR}}|$. After rearranging terms, this condition is precisely

$$\frac{\hat{r}_{0,1}}{\hat{r}_{0,1} - \Delta U}\widehat{\alpha}_{0,1} - \frac{\hat{r}_{1,1}}{\hat{r}_{1,1} + \Delta U}\widehat{\alpha}_{1,1} = \frac{\Delta U}{2}\left(\frac{1}{\hat{r}_{1,1} + \Delta U} + \frac{1}{\hat{r}_{0,1} - \Delta U}\right). \tag{72}$$