A Proof of Theorem 4.2

Additional notations. We use the following notations. A Bernoulli random variable B with success probability $p \in [0,1]$ is denoted as $B \sim \text{Bern}(p)$. We denote $p_a := P(A = a)$ and $p_{Y|a} := P(Y = 1|A = a)$; Further, we denote by P_X and $P_{X|a}$ the distribution of X and the conditional distribution of X given $A = a$, respectively.

The proof of Theorem 4.2 relies on the following two technical lemmas. For any $\tau \in [0,1]$, consider a Bernoulli random variable $B \sim \text{Bern}(\tau)$, independent of other sources of randomness considered. For all $a \in \mathcal{A}$, $t \in [0,1]$ and $\tau \in [0,1]$, define the random variable $\hat{Y}_{a,t,\tau}$ by

$$\hat{Y}_{a,t,\tau} = I(\eta_a(X) > t) + B \cdot I(\eta_a(X) = t).$$

For all $a \in \mathcal{A}$, define the set S

$$S = \{(t,\tau) \in [0,1]^2 : \; P(\eta_A(X) > t|A = a) + \tau \cdot P(\eta_A(X) = t|A = a) > 0\}.$$

For all $(t,\tau) \in S$, denote

$$g_a(t,\tau) = P(Y = 1|\hat{Y}_{a,t,\tau} = 1, A = a). \tag{7}$$

This is well-defined due to the definition of S.

Lemma A.1. *For all $a \in \mathcal{A}$, $t \in [0,1]$ and $0 \leq \tau_1 \leq \tau_2 \leq 1$, such that $(t,\tau_1) \in S$, we have*

$$g_a(t,\tau_1) \geq g_a(t,\tau_2) \geq t. \tag{8}$$

Furthermore, for all $a \in \mathcal{A}$, $\tau_1, \tau_2 \in [0,1]$ and $0 \leq t_1 \leq t_2 \leq 1$ such that $(t_1,\tau_1) \in S$, $(t_1,\tau_2) \in S$,

$$g_a(t_1,\tau_1) \leq g_a(t_2,\tau_2). \tag{9}$$

Proof. For all $a \in \mathcal{A}$, $y \in \mathcal{Y}$, and $t \in [0,1]$, denote

$$\begin{aligned} w_{ay}(t) &= P(\eta_A(X) > t|A = a, Y = y), \qquad v_{ay}(t) = P(\eta_A(X) = t|A = a, Y = y), \\ w_a(t) &= P(\eta_A(X) > t|A = a), \qquad v_a(t) = P(\eta_A(X) = t|A = a). \end{aligned}$$

Let $0 \leq t_1 \leq t_2 \leq 1$. Recalling the conditional density $dP_{X|a,y}$ of X given $A = a$ and $Y = y$, we have that $\eta_a(x) = \frac{p_{Y,a} dP_{X|A=a,Y=1}(x)}{dP_{X|a}(x)}$. We thus have for all $t \in [0,1]$ for which $w_a(t) > 0$ that

$$\frac{p_{Y|a} w_{a1}(t)}{w_a(t)} = \frac{p_{Y|a} \int_{\eta_a(x)>t} dP_{X|A=a,Y=1}(x)}{\int_{\eta_a(x)>t} dP_{X|a}(x)} = \frac{\int_{\eta_a(x)>t} \eta_a(x) dP_{X|a}(x)}{\int_{\eta_a(x)>t} dP_{X|a}(x)} > t.$$

Further, when $v_a(t) > 0$,

$$\frac{p_{Y|a} v_{a1}(t)}{v_a(t)} = \frac{p_{Y|a} \int_{\eta_a(x)=t} dP_{X|A=a,Y=1}(x)}{\int_{\eta_a(x)=t} dP_{X|a}(x)} = \frac{\int_{\eta_a(x)=t} \eta_a(x) dP_{X|a}(x)}{\int_{\eta_a(x)=t} dP_{X|a}(x)} = t.$$

It follows that, for $t \in [0,1]$ and $0 \leq \tau_1 \leq \tau_2 \leq 1$, such that $(t,\tau_1) \in S$,

$$t \leq \frac{w_{a1}(t) + v_{a1}(t)}{w_a(t) + v_a(t)} \leq \frac{w_{a1}(t) + \tau_2 v_{a1}(t)}{w_a(t) + \tau_2 v_a(t)} \leq \frac{w_{a1}(t) + \tau_1 v_{a1}(t)}{w_a(t) + \tau_1 v_a(t)}.$$

Eq. (8) follows since for all $t, \tau \in [0,1]$ such that $(t,\tau) \in S$,

$$g_a(t,\tau) = \frac{p_{Y,a}[w_{a1}(t) + \tau v_{a1}(t)]}{w_a(t) + \tau v_a(t)}.$$

For Eq. (9), we have that, when $0 \leq t_1 \leq t_2 \leq 1$ and $P(\eta_A(X) > t_2|A = a) > 0$,

$$\begin{aligned} g_a(t_1,\tau_1) - g_a(t_2,\tau_2) &= \frac{p_{Y|a}[w_{a1}(t_1) + \tau_1 v_{a1}(t_1)]}{w_a(t_1) + \tau_1 v_a(t_1)} - \frac{p_{Y,a}[w_{a1}(t_2) + \tau_2 v_{a1}(t_2)]}{w_a(t_2) + \tau_2 v_a(t_2)} \\ &\leq \frac{p_{Y,a} w_{a1}(t_1)}{w_a(t_1)} - \frac{p_{Y,a}[w_{a1}(t_2) + v_{a1}(t_2)]}{w_a(t_2) + v_a(t_2)}. \end{aligned}$$

This further equals

$$\begin{aligned}
&\frac{\int_{\eta_a(x)>t_1}\eta_a(x)dP_{X|a}(x)}{\int_{\eta_a(x)>t_1}dP_{X|a}(x)}-\frac{\int_{\eta_a(x)\geq t_2}\eta_a(x)dP_{X|a}(x)}{\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)}\\
=\quad&\frac{\int_{t_1<\eta_a(x)<t_2}\eta_a(x)dP_{X|a}(x)+\int_{\eta_a(x)\geq t_2}\eta_a(x)dP_{X|a}(x)}{\int_{t_1<\eta_a(x)<t_2}dP_{X|a}(x)+\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)}-\frac{\int_{\eta_a(x)\geq t_2}\eta_a(x)dP_{X|a}(x)}{\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)}.
\end{aligned}$$

This can also be written as

$$\begin{aligned}
&\frac{\int_{t_1<\eta_a(x)<t_2}\eta_a(x)dP_{X|a}(x)\cdot\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)}{[\int_{t_1<\eta_a(x)<t_2}dP_{X|a}(x)+\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)]\cdot\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)}\\
&-\frac{\int_{t_1<\eta_a(x)<t_2}dP_{X|a}(x)\cdot\int_{\eta_a(x)\geq t_2}\eta_a(x)dP_{X|a}(x)}{[\int_{t_1<\eta_a(x)<t_2}dP_{X|a}(x)+\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)]\cdot\int_{\eta_a(x)\geq t_2}dP_{X|a}(x)}\leq 0.
\end{aligned}$$

This finishes the proof. □

Lemma A.2. *For any $a\in\mathcal{A}$ and $s\in[p_{Y|a},1]$, there exists $(t_s,\tau_s)\in[0,1]^2$ such that, with g_a from* (7)*,*

$$g_a(t_s,\tau_s)=s.$$

Proof. For all $a\in\mathcal{A}$, define the set T on which $g_a(t,0)$ and $g_a(t,1)$, respectively, are well-defined:

$$T=\{t\in[0,1]:\ P(\eta_A(X)>t|A=a)>0\}.$$

As a function of $t\in T$, $t\mapsto g_a(t,1)$ is left-continuous. Letting $t^*=\sup T\in[0,1]$. Since $g_a(0,1)=p_{Y|a}\leq s$, $t_s=\sup\{t\in T: g_a(t,1)\leq s\}$ is well-defined. From Lemma A.1, the definition of t_s, and the left-continuity of $t\mapsto g_a(t,1)$ on T, it follows that

$$g_a(t_s,1)\leq s\leq g_a(t_s,0).$$

(1) When $P(\eta_a(X)=t_s|A=1)=0$, for all $\tau\in[0,1]$ we have

$$g_a(t_s,0)=g_a(t_s,\tau)=g_a(t_s,1)=s.$$

In this case, we can set $\tau_s\in[0,1]$.

(2) When $P(\eta_a(X)>t_s|A=1)=0$ for $a\in\mathcal{A}$, we have $s=t_s$ and we can set $\tau_s\in[0,1]$.

(3) When $P(\eta_a(X)=t_s|A=1)\neq 0$, we have $g_a(t_s,\tau_s)=s$ for

$$\tau_s=\frac{p_{Y|a}\cdot P(\eta_a(X)>t_s|A=a,Y=1)-s\cdot P(\eta_a(X)>t_s|A=a)}{p_{Y|a}\cdot P(\eta_a(X)=t_s|A=a)-s\cdot P(\eta_a(X)=t_s|A=a,Y=1)}.$$

□

Lemma A.3. *Let f be any classifier and $f_G=I(\eta_a(x)>t_a)+\tau_a(x)I(\eta_a(x)=t_a)$ be a GWTR satisfies*

$$I(\tau_a(x)\equiv 1)+I\left(\int f_G(x,a)\eta_a(x)dP_{X|a}(x)>t_a\int f_G(x,a)dP_{X|a}(x)\right)\geq 1. \tag{10}$$

Suppose that, for all $a\in\mathcal{A}$,

$$\int f(x,a)dP_{X|a}(x)=\int f_G(x,a)dP_{X|a}(x) \tag{11}$$

and

$$\frac{\int f(x,a)\eta_a(x)dP_{X|a}(x)}{\int f(x,a)dP_{X|a}(x)}=\frac{\int f_G(x,a)\eta_a(x)dP_{X|a}(x)}{\int f_G(x,a)dP_{X|a}(x)}. \tag{12}$$

Then, f is also a GWTR. Conversely, if f is not a GWTR and (12) *holds for all $a\in\mathcal{A}$, we have*

$$\sum_{a=1}^{|\mathcal{A}|}p_a\int[f_G(x,a)-f(x,a)]dP_{X|a}(x)>0.$$

Proof. We assume f_G takes the following form: for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$f_G(x,a) = I(\eta_a(x) > t_a) + \tau_a(x,a) I(\eta_a(x) = t_a).$$

From (11) and (12), we have

$$\int (f(x,a) - f_G(x,a)) dP_{X|a}(x) = \int_{\eta(x) > t_a} (f(x,a) - 1) dP_{X|a}(x) + \int_{\eta(x) < t_a} f(x,a) dP_{X|a}(x) + \int_{\eta(x) = t_a} (f(x,a) - \tau_a(x)) dP_{X|a}(x) = 0, \quad (13)$$

and

$$\int (f(x,a) - f_G(x,a)) \eta_a(x) dP_{X|a}(x) = \int_{\eta(x) > t_a} (f(x,a) - 1) \eta_a(x) dP_{X|a}(x) + \int_{\eta(x) < t_a} f(x,a) dP_{X|a}(x) + t_a \int_{\eta(x) = t_a} (f(x,a) - \tau_a(x)) dP_{X|a}(x) = 0. \quad (14)$$

Combining (13) and (14) gives us, for all $a \in \mathcal{A}$,

$$\int_{\eta_a(x) > t_a} (f(x,a) - 1)(\eta_a(x) - t_a) dP_{X|a}(x) + \int_{\eta_a(x) < t_a} f(x,a)(\eta_a(x) - t_a) dP_{X|a}(x) = 0.$$

Noting that $I(\eta_a(x) > t_a)(f(x,a) - 1)(\eta_a(x) - t_a) \leq 0$ and $I(\eta_a(x) < t_a) f(x,a)(\eta_a(x) - t_a) \leq 0$, we have

$$\int_{\eta_a(x) > t_a} (f(x,a) - 1)(\eta_a(x) - t_a) dP_{X|a}(x) + \int_{\eta_a(x) < t_a} f(x,a)(\eta_a(x) - t_a) dP_{X|a}(x) \leq 0. \quad (15)$$

The equality holds if and only if, for all $a \in \mathcal{A}$, $f(x,a) = f_G(x,a)$ almost surely on the set $\{\eta_a(x) > t_a\} \cup \{\eta_a(x) > t_a\}$. In other words, f is also a GWTR.

When f is not a GWTR, let

$$\frac{\int f(x,a) \eta_a(x) dP_{X|a}(x)}{\int f(x,a) dP_{X|a}(x)} = \frac{\int f_G(x,a) \eta_a(x) dP_{X|a}(x)}{\int f_G(x,a) dP_{X|a}(x)} = s_G.$$

We have $0 \leq t_a \leq s_G$ by Lemma A.2. Suppose there exists a $a \in \mathcal{A}$ such that

$$\int f(x,a) dP_{X|a}(x) > \int f_G(x,a) dP_{X|a}(x). \quad (16)$$

(1) When $t_a < s_G$, we have,

$$\begin{aligned} & \int [f(x,a) - f_G(x,a)] \eta_a(x) dP_{X|a}(x) - t_a \int [f(x,a) - f_G(x,a)] dP_{X|a}(x) \\ = & \int_{\eta_a(x) > t_a} (f(x,a) - 1)(\eta_a(x) - t_a) dP_{X|a}(x) + \int_{\eta_a(x) < t_a} f(x,a)(\eta_a(x) - t_a) dP_{X|a}(x) > 0. \end{aligned}$$

This contradicts (15).

(2) When $t_a = s_G$, we have $f(x,a) = I(\eta(x,a) \geq t_a)$. Then,

$$\int_{\eta_a(x) > t_a} (f(x,a) - 1)(\eta_a(x) - t_a) dP_{X|a}(x) + \int_{\eta_a(x) < t_a} f(x,a)(\eta_a(x) - t_a) dP_{X|a}(x) = 0.$$

This equation holds if and only if $f(x,a) = f_G(x,a)$ almost surely on the set $\{\eta_a(x) > t_a\} \cup \{\eta_a(x) > t_a\}$. Then,

$$\int f(x,a) dP_{X|a}(x) - \int f_G(x,a) dP_{X|a}(x) = \int_{\eta(x,a) = t_a} (f(x,a) - 1) dP_{X|a}(x) \leq 0.$$

Again, we have a contradiction since $\int f(x,a) dP_{X|a}(x) - \int f_G(x,a) dP_{X|a}(x) > 0$.

As a result, we can conclude that, for all $a \in \mathcal{A}$,

$$\int f(x,a) dP_{X|a}(x) \leq \int f_G(x,a) dP_{X|a}(x).$$

Moreover, there exists at least one $a \in \mathcal{A}$ such that

$$\int f(x,a) dP_{X|a}(x) < \int f_G(x,a) dP_{X|a}(x).$$

Otherwise, f is also a GWTR. This finishes the proof. □

We adopt the following strategy to prove Theorem 4.2. Consider any classifier f that satisfies predictive parity, which is not a GWTR. We will show that there exist a GWTR satisfying predictive parity with a smaller risk. Thus, at least one of the fair Bayes-optimal classifier under predictive parity is a GWTR.

Recall that $\hat{Y}_f$ is the prediction of f at (x, a). As f satisfies predictive parity, there exists $s_f \in [0, 1]$ such that

$$P(Y = 1|A = a, \hat{Y}_f = 1) = s_f \leq 1 \quad \text{for } a \in \mathcal{A}.$$

We set

$$s^\dagger = \begin{cases} \max(s_f, \max_a p_{Y|a}), & \max(s_f, \max_a p_{Y|a}) > c; \\ c + \varepsilon, & \max(s_f, \max_a p_{Y|a}) \leq c. \end{cases} \tag{17}$$

Here, $\varepsilon < 1 - c$ is a small constant such that there exists a $a \in \mathcal{A}$ with $P(\eta_a(X) > c + \varepsilon | A = a) > 0$. By our construction, we have $s^\dagger \in [\max_a p_{Y|a}, 1]$ and, according to Lemma A.2, there exist combinations $(t_a^\dagger, \tau_a^\dagger)_{a=1}^{|\mathcal{A}|}$ such that, for g_a from (7),

$$g_a(t_a^\dagger, \tau_a^\dagger) = s^\dagger, \quad a \in \mathcal{A}. \tag{18}$$

Now, we consider the GWTR $f^\dagger$ defined for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$ by

$$f^\dagger(x, a) = I(\eta_a(x) > t_a^\dagger) + \tau_a^\dagger I(\eta_a(x) = t_a^\dagger). \tag{19}$$

Here, we follow the construction in Lemma A.2 to set $t_a^\dagger = \sup\{t : g_a(t) < s^\dagger\}$, and let $\tau_a(x) \equiv \tau_a^\dagger$ be a constant function. Moreover, we set $\tau_a^\dagger = 1$ whenever $P(\eta_a(X) > t_a^\dagger | A = a) = 0$ or $P(\eta_a(X) = t_a^\dagger | A = a) = 0$. Clearly, $f^\dagger$ satisfies predictive parity, and thus it is enough to show that $f^\dagger$ has a smaller risk than f, i.e., $R_c(f^\dagger) - R_c(f) < 0$. Now, we can write

$$\begin{aligned} R_c(f) &= \sum_{a \in \mathcal{A}} \Big[(1 - c)P(\hat{Y}_f = 0, Y = 1, A = a) + c \cdot P(\hat{Y}_f = 1, Y = 0, A = a)\Big] \\ &= (1 - c)P(Y = 1) - \sum_{a \in \mathcal{A}} p_a (1 - c) \int f(x, a)\eta_a(x) dP_{X|a}(x) \\ &\quad + \sum_{a \in \mathcal{A}} p_a c \int f(x, a)(1 - \eta_a(x)) dP_{X|a}(x). \end{aligned}$$

Next, for any classifier f satisfying predictive parity with positive predictive value s_f, we have that

$$\begin{aligned} s_f &= P(Y = 1|\hat{Y}_f = 1, A = a) = \tfrac{p_{Y|a} P(\hat{Y}_f = 1|Y = 1, A = a)}{P(\hat{Y}_f = 1|A = a)} \\ &= \tfrac{p_{Y|a} \int f(x,a) dP_{X|A=a,Y=1}(x)}{\int f(x,a) dP_{X|a}(x)} = \tfrac{\int f(x,a)\eta_a(x) dP_{X|a}(x)}{\int f(x,a) dP_{X|a}(x)}. \end{aligned}$$

It follows that $R_c(f)$ further equals

$$\begin{aligned} & \sum_{a \in \mathcal{A}} p_a \int f(x, a)(c - \eta_a(x)) dP_{X|a}(x) + (1 - c)P(Y = 1) \\ = & \sum_{a \in \mathcal{A}} p_a (c - s_f) \int f(x, a) dP_{X|a}(x) + (1 - c)P(Y = 1). \end{aligned}$$

As a result, $R_c(f^\dagger) - R_c(f)$ equals

$$\sum_{a \in \mathcal{A}} p_a (c - s^\dagger) \int f^\dagger(x, a) dP_{X|a}(x) - \sum_{a \in \mathcal{A}} p_a (c - s_f) \int f(x, a) dP_{X|a}(x).$$

We consider the following three cases in order: (1) $s_f \leq \min(c, \max_a p_{Y|a})$, (2) $s_f > \max(c, \max_a p_{Y|a})$, and (3) $\min(c, \max_a p_{Y|a}) < s_f \leq \max(c, \max_a p_{Y|a})$.

(1) Case 1: $s_f \leq \min(c, \max_a p_{Y|a})$.

It is clear that $R_c(f^\dagger) - R_c(f) < 0$ since $c - s^\dagger < 0$ and $c - s_f \geq 0$.

(2) Case 2: $s_f > \max(c, \max_a p_{Y|a})$.

We have from the definition of $s^\dagger$, (18) and (8) that for all $a \in \mathcal{A}$, $s^\dagger = s_f \geq t_a^\dagger$. Further, we can write

$$R_c(f^\dagger) - R_c(f) = \sum_{a\in\mathcal{A}} p_a(c - s_f) \int [f^\dagger(x,a) - f(x,a)] dP_{X|a}(x).$$

Suppose that $s_f = t_a$. Specifically, $s_f = s^\dagger = t_a^\dagger$ equals

$$t_a^\dagger = \frac{\int f^\dagger(x,a)\eta_a(x)dP_{X|a}(x)}{\int f^\dagger(x,a)dP_{X|a}(x)} \quad = \quad \frac{\int_{\eta_a(x)>t_a^\dagger} \eta_a(x)dP_{X|a}(x) + \tau_a^\dagger \int_{\eta_a(x)=t_a^\dagger} \eta_a(x)dP_{X|a}(x)}{\int_{\eta_a(x)>t_a^\dagger} dP_{X|a}(x) + \tau_a^\dagger \int_{\eta_a(x)=t_a^\dagger} dP_{X|a}(x)}.$$

This implies $P(\eta_a(X) > t_a^\dagger | A = a) = 0$ and, by our construction, $\tau_a(x) \equiv \tau_a^\dagger = 1$. Thus, $f^\dagger$ satisfies the condition (10) in Lemma A.3. As a result, we have,

$$\sum_{a=1}^{|\mathcal{A}|} p_a \int [f^\dagger(x,a) - f(x,a)] dP_{X|a}(x) > 0.$$

This implies $R_c(f^\dagger) - R_c(f) < 0$ since $c - s_f < 0$.

(3) Case 3: $\min(c, \max_a p_{Y|a}) < s_f \leq \max(c, \max_a p_{Y|a})$.

In fact, the case 3 can be further divided into two possible sub-cases, depending on the relations between c and $\max_a p_{Y|a}$: (3.i) $\max_a p_{Y|a} < s_f \leq c$, and (3.ii) $c < s_f \leq \max_a p_{Y|a}$.

Sub-case (3.i): In this case, we have $s^\dagger = c + \varepsilon$ and $s_f \leq c$. Then,

$$\begin{aligned} R_c(f^\dagger) - R_c(f) &= \sum_{a\in\mathcal{A}} p_a(c - s^\dagger) \int f^\dagger(x,a)dP_{X|a}(x) - \sum_{a\in\mathcal{A}} p_a(c - s_f) \int f(x,a)dP_{X|a}(x) \\ &\leq -\varepsilon \sum_{a\in\mathcal{A}} p_a P(\eta_a(X) > c + \varepsilon | A = a) < 0 \end{aligned}$$

Sub-case (3.ii): In this case, we partition $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$ into the sets $\mathcal{A}_1 = \{a : p_{Y|a} \leq s_f\}$ and $\mathcal{A}_2 = \{a : p_{Y|a} > s_f\}$. Denoting $s_a^\flat = \max(s_f, p_{Y|a})$, it is clear that $s_f \leq s_a^\flat \leq s^\dagger$. According to Lemma A.2, there exist combinations $(t_a^\flat, \tau_a^\flat)_{a=1}^{|\mathcal{A}|}$ such that

$$g_a(t_a^\flat, \tau_a^\flat) = s_a^\flat, \quad a \in \mathcal{A}.$$

We now consider the classifier $f^\flat$ defined for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$ by

$$f^\flat(x,a) = I(\eta_a(x) > t_a^\flat) + \tau_a^\flat I(\eta_a(x) = t_a^\flat).$$

Again, we follow the construction in Lemma A.2 to set $t_a^\flat = \sup\{t : g_a(t) < s_a^\flat\}$, and let $\tau_a(x) \equiv \tau_a^\flat$ be a constant function. Moreover, we set $\tau_a^\flat = 1$ whenever $P(\eta_a(X) > t_a^\flat | A = a) = 0$ or $P(\eta_a(X) = t_a^\flat | A = a) = 0$.

Note that $s_f = s_a^\flat > c$ for $a \in \mathcal{A}_1$. Following the same argument as in case (2), we have,

$$\sum_{a\in\mathcal{A}_1} (c - s_a^\flat) \int f^\flat(x,a)dP_{X|a}(x) - \sum_{a\in\mathcal{A}_1} (c - s_f) f(X,A) dP_{X|a}(x) < 0. \tag{20}$$

For $a \in \mathcal{A}_2$, we have $s_a^\flat = p_{Y|a}$, which implies that $(t_a^\flat, \tau_a^\flat) = (0, 1)$. As a consequence, for $a \in \mathcal{A}_2$,

$$\begin{aligned} &(c - s_a^\flat) \int f^\flat(x,a)dP_{X|a}(x) - (c - s_f) \int f(x,a)dP_{X|a}(x) \\ = \quad &(c - p_{Y|a}) - (c - s_f) \int f(x,a)dP_{X|a}(x) < 0 \end{aligned} \tag{21}$$

since $\int f(x,a)dP_{X|a}(x) \leq 1$ and $c - p_{Y|a} < c - s_f \leq 0$. Combining (20) and (21) shows that $R_c(f^\flat) - R_c(f)$ equals

$$\sum_{a\in\mathcal{A}} p_a(c - s^\flat) \int f^\flat(x,a)dP_{X|a}(x) - \sum_{a\in\mathcal{A}} p_a(c - s_f) \int f(x,a)dP_{X|a}(x) < 0.$$

Now, under the Condition 4.1, we have, for $a \in \mathcal{A}$,

$$g_a(c,1) \geq \max_a p_{Y|a} = s^\dagger = g_a(t_a^\dagger, \tau_a^\dagger) \geq s_a^\flat = g_a(t_a^\flat, \tau_a^\flat).$$

From (8), we have $t_a^\flat \leq t_a^\dagger \leq c$. Thus, $R_c(f^\dagger) - R_c(f^\flat)$ equals

$$\begin{aligned} & \sum_{a\in\mathcal{A}} p_a \int (c - \eta_a(x))[f^\dagger(x,a) - f^\flat(x,a)]dP_{X|a}(x) \\ = & \sum_{a\in\mathcal{A}} p_a \int (\eta_a(x) - c)[f^\flat(x,a) - f^\dagger(x,a)]dP_{X|a}(x) \\ \leq & \sum_{a\in\mathcal{A}} p_a \int (\eta_a(x) - c)I(t_a^\flat \leq \eta_a(x) \leq t_a^\dagger)dP_{X|a}(x) \leq 0. \end{aligned}$$

As a result,

$$R_c(f^\dagger) - R_c(f) = R_c(f^\dagger) - R_c(f^\flat) + R_c(f^\flat) - R_c(f) < 0.$$

This finishes the proof.

B Proof of Theorem 4.3

Let f_G be any GWTR, say of the form

$$f_G(x,a) = I(\eta_a(x) > t_{G,a}) + \tau_{G,a}(x)I(\eta_a(x) = t_{G,a}),$$

satisfying predictive parity with

$$P(Y = 1|\hat{Y}_{f_G(x,a)}) = 1, A = a) = s_G, \quad \text{for } a \in \mathcal{A}.$$

According to Lemma A.1, we have $p_{Y|A=0} = P(Y = 1|\eta_0(x) \geq 0, A = 0) \leq s_{f_G}$. By the definition of t_1, we have $c < t_1 \leq t_{G_1}$. Thus, $P(c < \eta_1(X) < t_{G_1}|A = 1) > P(c < \eta_1(X) < t_1|A = 1) > 0$.

Denote $s_{NG} = P(Y = 1|\eta_1(X) \geq c, A = 1)$. We have $s_G \geq s_{NG}$, since $t_{G,1} > c$. Further, following the same argument as in Lemma A.2, there exist (t_0, τ_0) such that

$$\frac{P_{Y|a}[P(\eta_0(X) < t_0|A = 0, Y = 1) + \tau_0 P(\eta_0(X) = t_0|A = 0, Y = 1)]}{P(\eta_0(X) < t_0|A = 0) + \tau_0 P(\eta_0(X) = t_0|A = 0)} = s_{NG}.$$

We consider the following classifier f_{NG}, which is not a GWTR:

$$f_{NG}(x,a) = \begin{cases} I(\eta_a(x) \geq c), & a = 1; \\ I(\eta_a(x) < t_0) + \tau_0 I(\eta_a(x) = t_0), & a = 0. \end{cases} \tag{22}$$

By construction, f_{NG} satisfies predictive parity. Moreover, when $p_1 > \frac{2}{2+\delta_1\delta_2}$, we have

$$\begin{aligned} & R_c(f_G) - R_c(f_{NG}) = p_1 \int (c - \eta_1(x))[f_G(x,1) - f_{NG}(x,1)]dP_{X|1}(x) \\ & + p_0 \int (c - \eta_0(x))[f_G(x,1) - f_{NG}(x,0)]dP_{X|0}(x) \\ & \geq p_1 \int_{c<\eta_1(x)<t_{G_1}} (c - \eta_1(x))dP_{X|1}(x) - 2p_0 \geq p_1 \int_{c+\delta_2<\eta_1(x)<t_{G_1}} (c - \eta_1(x))dP_{X|1}(x) - 2p_0 \\ & \geq \delta_1\delta_2 p_1 - 2(1 - p_1) > 0. \end{aligned}$$

Thus, we have constructed a classifier that is not a GWTR satisfying predictive parity and achieving a smaller cost-sensitive risk than any fair GWTR. We can conclude that no fair Bayes-optimal classifier under predictive parity is a GWTR.

C Fair and Unconstrained Bayes-optimal Classifiers of the Synthetic Model

In this section, we derive the unconstrained and fair Bayes-optimal classifiers for our synthetic model used in Section 6.1. Consider the following data distribution for (X, A, Y) where $A \in \{0, 1\}$, $Y \in \{0, 1\}$ with

- For $a \in \{0, 1\}$, $P(A = a) = p_a$ and $P(Y = 1|A = a) = 1 - P(Y = 0|A = a) = p_{Y|a}$;
- For $(a, y) \in \{0, 1\}^2$, $X|A = a, Y = y \sim \mathcal{N}(\mu_{a,y}, \sigma^2 I_2)$ with $\mu_{a,y} = (2a - 1, 2y - 1)^\top$.

Denote by $g_{a,y}(x) = \frac{1}{2\pi\sigma^2}\exp(-\frac{1}{2\sigma^2}\|x - \mu_{a,y}\|^2)$ the conditional density function of X given $A = a$ and $Y = y$. we have

$$\begin{aligned}\eta_a(x) &= P(Y = 1|X = x, A = a) = \frac{p_{Y|a}g_{a,1}(x)}{p_{Y|a}g_{a,1}(x) + (1 - p_{Y|a})g_{a,0}(x)}\\ &= \frac{p_{Y|a}\exp(-\frac{1}{2\sigma^2}\|x - \mu_{a,1}\|^2)}{p_{Y|a}\exp(-\frac{1}{2\sigma^2}\|x - \mu_{a,1}\|^2) + (1 - p_{Y|a})\exp(-\frac{1}{2\sigma^2}\|x - \mu_{a,0}\|^2)}.\end{aligned}$$

Then, the unconstrained deterministic Bayes-optimal classifier $f^\star$ is

$$\begin{aligned}f^\star(x, a) &= I(\eta_a(x) > c)\\ &= I\left((1 - c)p_{Y|a}\exp(-\frac{1}{2\sigma^2}\|x - \mu_{a,1}\|^2) > c(1 - p_{Y|a})\exp(-\frac{1}{2\sigma^2}\|x - \mu_{a,0}\|^2)\right)\\ &= I\left(x^\top(\mu_{a,0} - \mu_{a,1}) < \log\frac{(1 - c)p_{Y|a}}{c(1 - p_{Y|a})}\right).\end{aligned}$$

For given $p_{Y|A=0}$, Condition 4.1 is equivalent to

$$\begin{aligned}p_{Y|A=1} &\leq P(Y = 1|\eta_A(X) > c, A = 0)\frac{p_{Y|A=0}P(\eta_A(X) > c|A = 0, Y = 1)}{P(\eta_A(X) > c|A = 0)}\\ &= \frac{p_{Y|A=0}P\left(X^\top(\mu_{0,0} - \mu_{0,1}) < \log\frac{(1-c)p_{Y|A=0}}{c(1-p_{Y|A=0})}|A = 0, Y = 1\right)}{P\left(X^\top(\mu_{0,0} - \mu_{0,1}) < \log\frac{(1-c)p_{Y|A=0}}{c(1-p_{Y|A=0})}|A = 0\right)}\\ &= \frac{p_{Y|A=0}\bar{\Phi}\left(\frac{\sigma\log(q_0(c))}{2} - \frac{1}{\sigma}\right)}{p_{Y|A=0}\bar{\Phi}\left(\frac{\sigma\log(q_0(c))}{2} - \frac{1}{\sigma}\right) + (1 - p_{Y|A=0})\bar{\Phi}\left(\frac{\sigma\log(q_0(c))}{2} + \frac{1}{\sigma}\right)},\end{aligned} \tag{23}$$

where $q_a(c) = \frac{c(1-p_{Y|a})}{(1-c)p_{Y|a}}$ and $\bar{\Phi}(t) = 1 - \Phi(t)$ with $\Phi(t)$ the cumulative distribution function of the standard normal distribution.

Now we consider fair Bayes optimal classifiers under (23). We consider the GWTR f_{t_1,t_0} such that for $a \in \{0, 1\}$ and all $x \in \mathcal{X}$, $f_{t_1,t_0}(x, a) = I(\eta_a(x) > t_a)$. Following the same argument as in (23), we have

$$P(Y = 1|\eta_a(x) > t_a, A = a) = \frac{p_{Y|a}\bar{\Phi}\left(\frac{\sigma\log(q_a(t_a))}{2} - \frac{1}{\sigma}\right)}{p_{Y|a}\bar{\Phi}\left(\frac{\sigma\log(q_a(t_a))}{2} - \frac{1}{\sigma}\right) + (1 - p_{Y|a})\bar{\Phi}\left(\frac{\sigma\log(q_a(t_a))}{2} + \frac{1}{\sigma}\right)}.$$

Then, f_{t_1,t_0} satisfies predictive parity if

$$\frac{p_{Y|A=1}\bar{\Phi}\left(\frac{\sigma\log(q_1(t_1))}{2} - \frac{1}{\sigma}\right)}{(1 - p_{Y|A=1})\bar{\Phi}\left(\frac{\sigma\log(q_1(t_1))}{2} + \frac{1}{\sigma}\right)} = \frac{p_{Y|A=0}\bar{\Phi}\left(\frac{\sigma\log(q_0(t_0))}{2} - \frac{1}{\sigma}\right)}{(1 - p_{Y|A=0})\bar{\Phi}\left(\frac{\sigma\log(q_0(t_0))}{2} + \frac{1}{\sigma}\right)}.$$

Note that, as a function of t_a, $t_a \mapsto P(Y = 1|\eta_a(x) > t_a, A = a)$ is strictly monotone increasing. Thus, for $T_1(t) = t$, there exists a function $t \mapsto T_0(t)$ such that

$$\frac{p_{Y|A=1}\bar{\Phi}\left(\frac{\sigma\log(q_1(t))}{2} - \frac{1}{\sigma}\right)}{(1 - p_{Y|A=1})\bar{\Phi}\left(\frac{\sigma\log(q_1(t))}{2} + \frac{1}{\sigma}\right)} = \frac{p_{Y|A=0}\bar{\Phi}\left(\frac{\sigma\log(q_0(T_0(t)))}{2} - \frac{1}{\sigma}\right)}{(1 - p_{Y|A=0})\bar{\Phi}\left(\frac{\sigma\log(q_0(T_0(t)))}{2} + \frac{1}{\sigma}\right)}.$$

Then $f_{T_1(t),T_0(t)}$ satisfies predictive parity and its cost-sensitive risk $R_c(f_{T_1(t),T_0(t)})$ is

$$\begin{aligned} &\sum_{a\in\{0,1\}} cp_a(1-p_{Y|a})P(\eta_a(X)\geq t_a|A=a,Y=0) \\ &\qquad\qquad + \sum_{a\in\{0,1\}} (1-c)p_a p_{Y|a} P(\eta_a(X) < t_a|A=a,Y=1) \\ =& \sum_{a\in\{0,1\}} (1-c)p_a p_{Y|a}\Phi\left(\frac{\sigma\log(q_a(T_a(t)))}{2}-\frac{1}{\sigma}\right) \\ &\qquad\qquad + \sum_{a\in\{0,1\}} cp_a(1-p_{Y|a})\bar{\Phi}\left(\frac{\sigma\log(q_a(T_a(t)))}{2}+\frac{1}{\sigma}\right). \end{aligned}$$

Let $t^\star$ be defined as

$$t^\star = \operatorname*{argmin}_{t\in[0,1]} R_c(f_{T_1(t),T_0(t)}).$$

Thus, under (23), the fair Bayes-optimal classifier under predictive parity is given by $f_{t^\star,T_0(t^\star)}$. This classifier can be computed numerically as both $T_0(t)$ and $t^\star$ can be found numerically.

D Experimental Settings and More Simulation Results

Training details. Our experiments are conducted on a personal computer with an Intel(R) Core(TM) i9-9920X CPU @ 3.50Ghz and an NVIDIA GeForce RTX 2080 Ti GPU. For the Adult and COMPAS datasets, we employ the same training settings as in [5]. We train the conditional probability predictor using a three-layer fully connected net with 32 neurons in the hidden layers. For the CelebA dateset, we adopt the same settings in [51] to train the conditional probability predictor with ResNet-50, pre-trained on the ImageNet dataset. We also apply the dropout technique with $p = 0.5$ to improve the model performance. In all the simulations, we use the Adam optimizer with the default parameters. The details are summarized in Table 3.

Table 3: Training details for three datasets

DATASET	ADULT CENSUS	COMPAS	CELEBA
BATCH SIZE	512	2048	32
TRAINING EPOCHS	200	500	50
OPTIMIZER	ADAM	ADAM	ADAM
LEARNING RATE	1E-1	5E-4	1E-4
PRE-TRAINING	N/A	N/A	IMAGENET
DROPOUT	N/A	N/A	0.5

D.1 Synthetic Data

We conduct more experiments to evaluate the performance of our FairBayes-DPP algorithm under different model and training settings. We consider the same synthetic model as in Section 6.1 with different settings on sample size, proportion $P(A = 0)$ of the minority group and cost parameters. We also extend the synthetic model to a multi-class protected attribute. In all scenarios, we repeat the experiments 100 times[12].

D.1.1 Sample Size

We first evaluate FairBayes-DPP with different sample sizes. In the experiment, we fix $c = 0.5$, $p(A = 1) = 0.3$, $p(Y = 1|A = 1) = 0.6$ and $p(Y = 1|A = 1) = 0.2$. We further fix the number of test data points to be 5000, and change the number of training data points from 5000 to 25000. The simulation results are presented in Figure 2. It can be seen that FairBayes-DPP has a smaller disparity than the unconstrained classifier. As the sample size grows, the performance of FairBayes-DPP improves, since the estimation error reduces with more training data points.

---

[12] The randomness of the experiment comes from the random generation of the training and test data sets.

D.1.2 Proportion of Minority Group

Next, we evaluate the effect of the proportion $P(A = 0)$ of the minority group on the performance of FairBayes-DPP. We fix $c = 0.5$, $p(Y = 1|A = 1) = 0.6$, $p(Y = 1|A = 0) = 0.2$, and vary $P(A = 0)$ from 0.5 to 0.9. Moreover, we set the training data size and test data size to be 25000 and 5000, respectively. Figure 3 presents the simulation results.

We observe that, for both FairBayes-DPP and unconstrained learning, the test accuracy increases with $P(A = 0)$. The sample complexity of learning the unconstrained classifier should intuitively depend on the sample size of the smallest group. When $P(A = 0)$ is very small, the estimator of η_0 has large variability and results in a small test accuracy.

We also observe that the performance of FairBayes-DPP is unstable when $P(A = 0)$ is very small. This limitation is caused by the unstable estimation of η_0, which is used by FairBayes-DPP to adjusts the per-class thresholds. As we can see, the performance of FairBayes-DPP improves rapidly when $P(A = 0)$ grows. We emphasize that the success of FairBayes-DPP relies on the consistent estimation of the per-group feature-conditional probabilities of the labels.

D.1.3 Cost Parameter

We then evaluate the effect of cost parameter c. We fix $P(Y = 1) = 0.3$, $p(Y = 1|A = 1) = 0.5$, $p(Y = 1|A = 0) = 0.2$, and vary c from 0.4 to 0.8. Again, we set the training and test data sizes to be 25000 and 5000, respectively. We present the simulation results in Figure 3. We observe that FairBayes-DPP successfully mitigates disparity with a wide range of cost parameters.

D.1.4 Multi-class Protected Attribute

Finally, we study a multi-class protected attribute. We generate data $a \in \mathcal{A} = \{1, 2, ..., |\mathcal{A}|\}$ and $y \in \{0, 1\}$ by setting $\mu_{ay} = (2y - 1)e_a$, where $e_a \in \mathbb{R}^{|\mathcal{A}|}$ the unit vector with the a-th element equal to unity. Conditional on $A = a$ and $Y = y$, X is generated from a multivariate Gaussian distribution $N(\mu_{ay}, 2^2 I_{|\mathcal{A}|})$.

We consider two cases, $|\mathcal{A}| = 3$ and $|\mathcal{A}| = 5$, with the model parameters presented in Table 4. For both cases, we set $c = 0.5$, the training data sample size as 50000 and the test data sample size as 5000. We present the simulation results in Figure 5. Again, FairBayes-DPP achieves superior performance in preserving accuracy and mitigating bias.

D.2 CelebA Dataset

In the main text, we have presented the simulation results for the first six attributes of the CelebA dataset. Here, we show the simulation results for the remaining 20 attributes in Table 5. Again, we

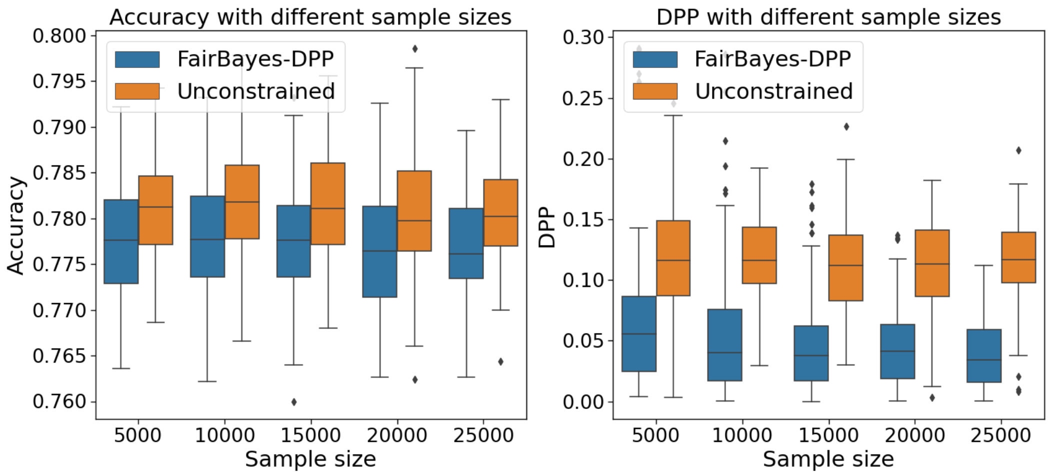

Figure 2: Accuracy and DPP as a function of sample size.

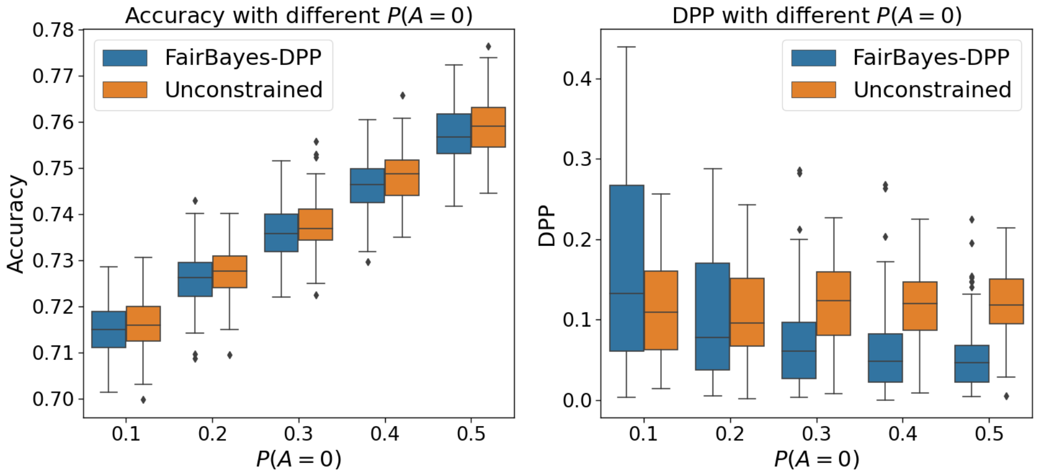

Figure 3: Accuracy and DPP as a function of $P(A=0)$.

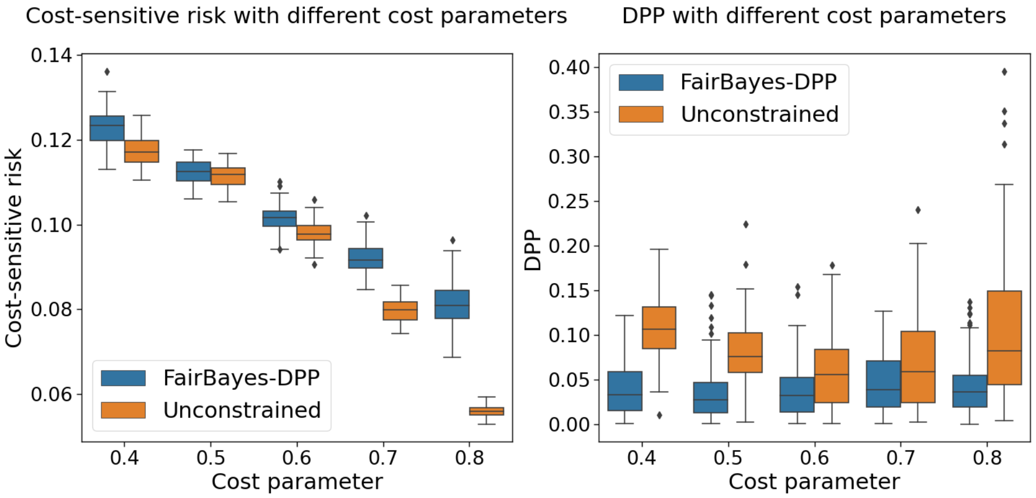

Figure 4: Cost-sensitive risk and DPP as a function of the cost parameter.

observe that FairBayes-DPP mitigates the gender bias effectively in most cases, and preserves model accuracy.

Table 4: Parameters of synthetic model for multi-clase protected attribute.

	$\|\mathcal{A}=3\|$				
a	1	2	3		
p_a	0.3	0.3	0.4		
$p_{Y\|a}$	0.2	0.6	0.3		
	$\|\mathcal{A}=5\|$				
a	1	2	3	4	5
p_a	0.2	0.3	0.2	0.15	0.15
$p_{Y\|a}$	0.2	0.6	0.3	0.4	0.2

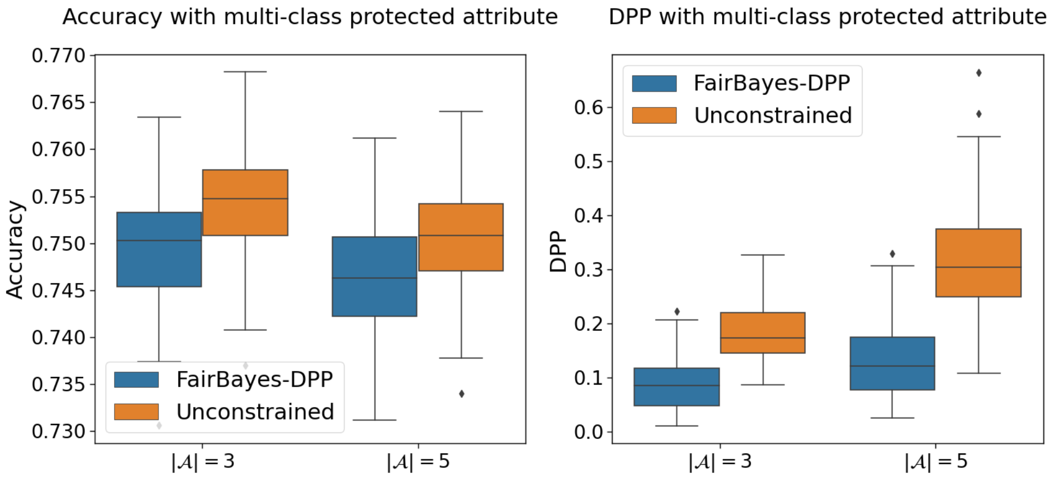

Figure 5: Accuracy and DPP with multi-class protected attribute.

Table 5: Per-attribute accuracy and DPP of the remaining 20 attributes from the CelebA dataset.

Attributes	Per-attribute Accuracy		Per-attribute DPP	
	FairBayes-DPP	Unconstrained	FairBayes-DPP	Unconstrained
Black Hair	0.895(0.004)	0.899(0.003)	0.023(0.009)	0.033(0.013)
Blond Hair	0.958(0.001)	0.959(0.001)	0.028(0.014)	0.119(0.042)
Blurry	0.963(0.001)	0.963(0.001)	0.023(0.017)	0.047(0.017)
Brown Hair	0.886(0.003)	0.889(0.004)	0.029(0.009)	0.078(0.028)
Bushy Eyebrows	0.928(0.001)	0.926(0.001)	0.055(0.030)	0.166(0.038)
Chubby	0.957(0.002)	0.957(0.002)	0.032(0.012)	0.043(0.026)
Eyeglasses	0.996(0.000)	0.997(0.000)	0.010(0.005)	0.004(0.003)
High Cheekbones	0.875(0.002)	0.876(0.002)	0.044(0.008)	0.143(0.016)
Mouth Slightly Open	0.940(0.001)	0.940(0.001)	0.011(0.003)	0.017(0.008)
Narrow Eyes	0.873(0.002)	0.875(0.003)	0.110(0.025)	0.063(0.026)
Oval Face	0.756(0.002)	0.756(0.003)	0.033(0.016)	0.108(0.031)
Pale Skin	0.970(0.001)	0.970(0.001)	0.059(0.040)	0.111(0.034)
Pointy Nose	0.775(0.003)	0.774(0.003)	0.032(0.018)	0.063(0.022)
Receding Hairline	0.939(0.001)	0.938(0.001)	0.067(0.019)	0.036(0.034)
Smiling	0.928(0.001)	0.928(0.002)	0.021(0.005)	0.046(0.014)
Straight Hair	0.842(0.002)	0.842(0.003)	0.056(0.007)	0.020(0.013)
Wavy Hair	0.844(0.003)	0.847(0.003)	0.019(0.014)	0.087(0.021)
Wearing Earrings	0.889(0.027)	0.908(0.001)	0.075(0.050)	0.207(0.037)
Wearing Hat	0.991(0.000)	0.991(0.000)	0.012(0.013)	0.047(0.018)
Wearing Necklace	0.868(0.002)	0.868(0.001)	0.077(0.047)	0.069(0.052)