# Fair Policy Targeting: Supplementary Material

Davide Viviano[*] Jelena Bradic[†]

October 28, 2022

# A Main Proofs

Throughout the rest of our discussion we define

$$\Pi_{o,n} = \Big\{\pi_\alpha \in \Pi : \pi_\alpha \in \arg\sup_{\pi\in\Pi} \Big\{\alpha \hat{W}_0(\pi) + (1-\alpha)\hat{W}_1(\pi)\Big\}, \text{ s.t. } \alpha \in (0,1)\Big\}, \tag{A.1}$$

and let $N = \sqrt{n}$ as discussed in the main text. We denote $\alpha_1 - \alpha_2 = \varepsilon$, where, recall, the grid of $(\alpha_i)_{i=1}^N$ contains elements equally spaced. We say that $x \lesssim y$ if $y \leq c_0 x$ for a finite constant $c_0$ independent of $n$.

## A.1 Auxiliary Lemmas

**Lemma A.1.** *Under Assumption 2.1, 4.2 for any sensitive attribute* $s \in \{0,1\}$

$$W_s(\pi) = \mathbb{E}\Big[\frac{1\{S_i = s\}}{p_s}\Big(\frac{Y_i D_i}{e(X_i, s)} - \frac{Y_i(1-D_i)}{1-e(X_i,s)}\Big)\pi(X_i, s)\Big]. \tag{A.2}$$

[*] Stanford Graduate School of Business and Department of Economics, Harvard University. Email: dviviano@fas.harvard.edu. This work was mostly conducted while at the Department of Economics, University of California at San Diego, La Jolla, CA, 92093.

[†] Department of Mathematics and Halicioğlu Data Science Institute, University of California at San Diego, La Jolla, CA, 92093. Email: jbradic@ucsd.edu.

*Proof of Lemma A.1.* Assumption 4.2 guarantees existence of the expectation. By definition of the conditional expectation

$$(\text{A.2}) = \mathbb{E}\Big[\Big(\frac{Y_i D_i}{e(X_i, s)} - \frac{Y_i(1-D_i)}{1-e(X_i, s)}\Big)\pi(X_i, s)\Big| S_i = s\Big].$$

Using the law of iterated expectations and Assumption 2.1 the result directly follows. □

**Lemma A.2.** *Let $W_{s,n} = \frac{1}{n}\sum_{i=1}^n (\Gamma_{1,s,i} - \Gamma_{0,s,i})\pi(X_i, s)$, where $\Gamma_{d,s,i}$ is defined as in Equation (10). Let Assumptions 2.1, 4.1, and 4.2 hold. Then with probability at least $1-\gamma$,*

$$\sup_{\alpha \in (0,1)} \sup_{\pi \in \Pi} \Big|\alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \alpha W_{0,n}(\pi) + (1-\alpha) W_{1,n}(\pi)\Big| \le \bar{C}\frac{M}{\delta^2}\sqrt{v/n} + \frac{\bar{C}M}{\delta^2}\sqrt{\log(2/\gamma)/n} \tag{A.3}$$

*for a universal constant $\bar{C} < \infty$. In addition,*

$$\mathbb{E}\Big[\sup_{\alpha \in (0,1)} \sup_{\pi \in \Pi} \Big|\alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \alpha W_{0,n}(\pi) + (1-\alpha) W_{1,n}(\pi)\Big|\Big] \le \bar{C}\frac{M}{\delta^2}\sqrt{v/n}. \tag{A.4}$$

*Proof of Lemma A.2.* Throughout the proof we refer to $\bar{C} < \infty$ as a universal constant. Observe first that under Assumption 4.2 and Assumption 4.1, we have

$$\sup_{\alpha \in (0,1)} \sup_{\pi \in \Pi} \Big|\alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \alpha W_{0,n}(\pi) + (1-\alpha) W_{1,n}(\pi)\Big|, \tag{A.5}$$

satisfies the bounded difference assumption (Boucheron et al., 2013) with constant $\frac{2M}{\delta^2 n}$.[1] See for instance Boucheron et al. (2005). By the bounded difference inequality, with probability

[1] This follows from the triangular inequality and the fact that under Assumption 4.2 the inverse probability weight is uniformly bounded by $1/\delta^2$ and under Assumption 4.1 the conditional mean function is bounded by $M$.

at least $1-\gamma$,

$$\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_0(\pi)+(1-\alpha)W_1(\pi)-\alpha W_{0,n}(\pi)+(1-\alpha)W_{1,n}(\pi)\Big|$$
$$\leq \mathbb{E}\Big[\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_0(\pi)+(1-\alpha)W_1(\pi)-\alpha W_{0,n}(\pi)+(1-\alpha)W_{1,n}(\pi)\Big|\Big]+\bar{C}\frac{M}{\delta^2}\sqrt{\log(2/\gamma)/n}. \tag{A.6}$$

We now move to bound the expectation in the right-hand side of Equation (A.6). Under Assumption 2.1, we obtain by Lemma A.1 and trivial rearrangments, that

$$\mathbb{E}\Big[\alpha W_0(\pi)+(1-\alpha)W_1(\pi)-\alpha W_{0,n}(\pi)+(1-\alpha)W_{1,n}(\pi)\Big]=0. \tag{A.7}$$

Using the symmetrization argument (Van Der Vaart and Wellner, 1996), we can now bound the above supremum with the Radamacher complexity of the function class of interest, which combined with the triangle inequality reads as follows:

$$\mathbb{E}\Big[\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_0(\pi)+(1-\alpha)W_1(\pi)-\alpha W_{0,n}(\pi)+(1-\alpha)W_{1,n}(\pi)\Big|\Big]$$
$$\leq \mathbb{E}\Big[\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_0(\pi)-\alpha W_{0,n}(\pi)\Big|\Big]+\mathbb{E}\Big[\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|(1-\alpha)W_1(\pi)-(1-\alpha)W_{1,n}(\pi)\Big|\Big]$$
$$\leq \mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|W_0(\pi)-W_{0,n}(\pi)\Big|\Big]+\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|W_1(\pi)-W_{1,n}(\pi)\Big|\Big]$$
$$\leq \mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i\pi(X_i,1)\Gamma_{1,1,i}\Big|\Big]+\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i\pi(X_i,1)\Gamma_{0,1,i}\Big|\Big]$$
$$+\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i\pi(X_i,0)\Gamma_{1,0,i}\Big|\Big]+\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i\pi(X_i,0)\Gamma_{0,0,i}\Big|\Big], \tag{A.8}$$

where here $\sigma_i$ are independent Radamacher random variables. We can study each component of the above expression separately. By the Dudley's entropy integral bound, since the VC-dimension of the function class $\Pi$ is bounded by Assumption 4.1, and since each $\Gamma_{d,s,}$ is bounded, we obtain (see for instance Wainwright (2019)), under Assumption 4.1 (A) and

(B), with trivial rearrangement

$$\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\pi(X_i,s)\Gamma_{d,s,i}\Big|\Big] \leq \frac{M\bar{C}}{\delta^2}\sqrt{v/n}. \tag{A.9}$$

for each $d, s$. The remaining terms follow similarly. The proof is complete. □

**Lemma A.3.** *Let Assumptions 2.1, 4.1-4.3 hold. Then with probability at least $1-\gamma$,*

$$\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_0(\pi)+(1-\alpha)W_1(\pi)-\alpha\hat{W}_0(\pi)+(1-\alpha)\hat{W}_1(\pi)\Big| \leq \bar{C}\frac{M}{\delta^2}\sqrt{v/n}+\frac{\bar{C}M}{\delta^2}\sqrt{\log(2/\gamma)/n} \tag{A.10}$$

*for a universal constant $\bar{C}<\infty$.*

*Proof of Lemma A.3.* First observe that we can bound the above expression as

$$\begin{aligned}&\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_0(\pi)+(1-\alpha)W_1(\pi)-\alpha\hat{W}_0(\pi)+(1-\alpha)\hat{W}_1(\pi)\Big| \leq \\ &\underbrace{\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_0(\pi)+(1-\alpha)W_1(\pi)-\alpha W_{0,n}(\pi)+(1-\alpha)W_{1,n}(\pi)\Big|}_{(I)} \\ &+\underbrace{\sup_{\alpha\in(0,1)}\sup_{\pi\in\Pi}\Big|\alpha W_{0,n}(\pi)+(1-\alpha)W_{1,n}(\pi)-\alpha\hat{W}_0(\pi)+(1-\alpha)\hat{W}_1(\pi)\Big|}_{(II)}.\end{aligned} \tag{A.11}$$

Here $W_{s,n}$ is as defined in Lemma A.2. The term (I) is bounded as in Lemma A.2. Therefore, we are only left to discuss (II).

Using the triangular inequality, we only need to bound

$$\sup_{\pi\in\Pi}\Big|W_{0,n}(\pi)-\hat{W}_{0,n}(\pi)\Big|+\sup_{\pi\in\Pi}\Big|W_{1,n}(\pi)-\hat{W}_{1,n}(\pi)\Big|. \tag{A.12}$$

We bound the first term while the second term follows similarly. We write

$$\begin{aligned}
&\sup_{\pi\in\Pi}\Big|W_{s,n}(\pi)-\hat{W}_s(\pi)\Big| \\
&\leq \Big|\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i=s\}}{p_s}\frac{D_i(Y_i-m_{1,s}(X_i))}{e(X_i,s)}\pi(X_i,s)+\frac{1\{S_i=s\}}{p_s}m_{1,s}(X_i)\pi(X_i,s) \\
&-\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i=s\}}{\hat{p}_s}\frac{D_i(Y_i-\hat{m}_{1,s}(X_i))}{\hat{e}(X_i,s)}\pi(X_i,s)-\frac{1\{S_i=s\}}{\hat{p}_s}\hat{m}_{1,s}(X_i)\pi(X_i,s)\Big| \\
&+\Big|\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i=s\}}{p_s}\frac{(1-D_i)(Y_i-m_{0,s}(X_i))}{1-e(X_i,s)}\pi(X_i,s)+\frac{1\{S_i=s\}}{p_s}m_{0,s}(X_i)\pi(X_i,s) \\
&-\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i=s\}}{\hat{p}_s}\frac{(1-D_i)(Y_i-\hat{m}_{s,0}(X_i))}{1-\hat{e}(X_i,s)}\pi(X_i,s)-\frac{1\{S_i=s\}}{\hat{p}_s}\hat{m}_{0,s}(X_i)\pi(X_i,s)\Big|.
\end{aligned} \tag{A.13}$$

We discuss the first component while the second follows similarly.

With trivial re-arrengment, using the triangular inequality, we obtain that the following holds

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i=s\}}{p_s}\frac{D_i(Y_i-m_{1,s}(X_i))}{e(X_i,s)}\pi(X_i,s)+\frac{1\{S_i=s\}}{p_s}m_{1,s}(X_i)\pi(X_i,s) \\
&-\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i=s\}}{\hat{p}_s}\frac{D_i(Y_i-\hat{m}_{1,s}(X_i))}{\hat{e}(X_i,s)}\pi(X_i,s)-\frac{1\{S_i=s\}}{\hat{p}_s}\hat{m}_{1,s}(X_i)\pi(X_i,s)\Big| \\
&\leq \underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n 1\{S_i=s\}D_i(Y_i-m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)}-\frac{1}{\hat{p}_s\hat{e}(X_i,s)}\Big)\pi(X_i,s)\Big|}_{(i)} \qquad \text{(A.14)} \\
&+\underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n \Big(\frac{1\{S_i=s\}D_i}{\hat{e}(X_i,s)\hat{p}_s}-\frac{1\{S_i=s\}}{p_s}\Big)(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))\pi(X_i,s)\Big|}_{(ii)}.
\end{aligned}$$

We study $(i)$ and $(ii)$ separately. We start from $(i)$. Recall, that by cross fitting $\hat{e}(X_i,s)=\hat{e}^{-k(i)}(X_i,s)$, where $k(i)$ is the fold containing unit $i$. Therefore, observe that given the $K$

folds for cross-fitting, we have

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n} 1\{S_i = s\}D_i(Y_i - m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)} - \frac{1}{\hat{p}_s\hat{e}(X_i,s)}\Big)\pi(X_i,s)\Big| \\
&\leq \sum_{k\in\{1,\ldots,K\}}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k} 1\{S_i = s\}D_i(Y_i - m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)} - \frac{1}{\hat{p}_s^{(-k(i))}\hat{e}^{(-k(i))}(X_i,s)}\Big)\pi(X_i,s)\Big|.
\end{aligned} \tag{A.15}$$

In addition, we have that

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i\in\mathcal{I}_k} 1\{S_i = s\}D_i(Y_i - m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)} - \frac{1}{\hat{p}_s^{(-k(i))}\hat{e}^{(-k(i))}(X_i,s)}\Big)\pi(X_i,s)\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\sum_{i\in\mathcal{I}_k} 1\{S_i = s\}D_i(Y_i - m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)} - \frac{1}{\hat{p}_s^{(-k(i))}\hat{e}^{(-k(i))}(X_i,s)}\Big)\pi(X_i,s)\Big|\hat{p}^{(-k(i))},\hat{e}^{(-k(i))}\Big]\Big] \\
&= 0,
\end{aligned} \tag{A.16}$$

by cross-fitting. By Assumption 4.3, we know that

$$\Big|\frac{1}{p_S e(X,S)} - \frac{1}{\hat{p}_S^{(-k(i))}\hat{e}^{(-k(i))}(X,S)}\Big| \leq 2/\delta^2 \tag{A.17}$$

almost surely and therefore each summand in Equation (A.15) is bounded by a finite constant $2M\bar{C}/\delta^2$, for a universal constant $\bar{C}$. We now obtain, using the symmetrization argument (Van Der Vaart and Wellner, 1996), and the Dudley's entropy integral (Wainwright, 2019)

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k} 1\{S_i = s\}D_i(Y_i - m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)} - \frac{1}{\hat{p}_s^{(-k(i))}\hat{e}^{(-k(i))}(X_i,s)}\Big)\pi(X_i,s)\Big|\Big|\hat{p}^{(-k(i))},\hat{e}^{(-k(i))}\Big] \\
&\lesssim \frac{M}{\delta^2}\sqrt{v/n}.
\end{aligned} \tag{A.18}$$

In addition, by the bounded difference inequality (Boucheron et al., 2005), with probability at least $1-\gamma$, for a universal constant $c<\infty$

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k}1\{S_i=s\}D_i(Y_i-m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)}-\frac{1}{\hat{p}_s^{(-k(i))}\hat{e}^{(-k(i))}(X_i,s)}\Big)\pi(X_i,s)\Big| \leq \\
&\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k}1\{S_i=s\}D_i(Y_i-m_{1,s}(X_i))\Big(\frac{1}{p_s e(X_i,s)}-\frac{1}{\hat{p}_s^{(-k(i))}\hat{e}^{(-k(i))}(X_i,s)}\Big)\pi(X_i,s)\Big|\Big|\hat{p}^{(-k(i))},\hat{e}^{(-k(i))}\Big] \\
&+c\frac{M}{\delta^2}\sqrt{\frac{\log(2/\gamma)}{n}}.
\end{aligned}
\tag{A.19}
$$

We now consider the term $(ii)$. Observe that we can write

$$
\begin{aligned}
(ii)\leq&\underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{D_i1\{S_i=s\}}{\hat{p}_s\hat{e}(X_i,s)}-\frac{D_i1\{S_i=s\}}{p_s e(X_i,s)}\Big)(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))\pi(X_i,s)\Big|}_{(j)} \\
&+\underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{D_i1\{S_i=s\}}{p_s e(X_i,s)}-\frac{1\{S_i=s\}}{p_s}\Big)(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))\pi(X_i,s)\Big|}_{(jj)}.
\end{aligned}
\tag{A.20}
$$

We consider each term seperately. Consider $(jj)$ first. Using the cross-fitting argument we obtain

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{D_i1\{S_i=s\}}{p_s e(X_i,s)}-\frac{1\{S_i=s\}}{p_s}\Big)(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))\pi(X_i,s)\Big| \\
&\leq\sum_{k\in\{1,\dots,K\}}\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k}\Big(\frac{D_i1\{S_i=s\}}{p_s e(X_i,s)}-\frac{1\{S_i=s\}}{p_s}\Big)(m_{1,s}(X_i)-\hat{m}_{1,s}^{(-k(i))}(X_i))\pi(X_i,s)\Big|.
\end{aligned}
\tag{A.21}
$$

Observe now that

$$
\mathbb{E}\Big[\Big(\frac{D_i1\{S_i=s\}}{p_s e(X_i,s)}-\frac{1\{S_i=s\}}{p_s}\Big)(m_{1,s}(X_i)-\hat{m}_{1,s}^{(-k(i))}(X_i))\pi(X_i,s)\Big|\hat{m}_{1,s}^{(-k(i))}\Big]=0, \tag{A.22}
$$

since by cross-fitting, $\hat{m}_{1,s}^{(-k(i))}$ is independent of $(D_i,S_i,X_i)$ and, as a result, the conditional expectation of the left-hand side in Equation (A.22), also conditional on $X_i$ equals zero. Therefore, following the same argument used for $(i)$ in Equation (A.14), we obtain that with probability at least $1-\gamma$

$$\begin{aligned}&\sum_{k\in\{1,\cdots,K\}}\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k}\Big(\frac{D_i 1\{S_i=s\}}{p_s e(X_i,s)}-\frac{1\{S_i=s\}}{p_s}\Big)(m_{1,s}(X_i)-\hat{m}_{1,s}^{(-k(i))}(X_i))\pi(X_i,s)\Big|\\&\lesssim\frac{KM}{\delta^2}\sqrt{\frac{vK}{n}}+\frac{MK}{\delta^2}\sqrt{\frac{\log(2K/\gamma)}{n}},\end{aligned}\tag{A.23}$$

where the number of folds $K$ is a constant. We are now left to bound $(j)$ in Equation (A.20). We obtain that

$$(j)\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{\hat{p}_s\hat{e}(X_i,s)}-\frac{1}{p_s e(X_i,s)}\Big)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))^2}.\tag{A.24}$$

Such a bound does not depend on $\pi$. Observe now that we can write by Assumption 4.3

$$\begin{aligned}&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{\hat{p}_s\hat{e}(X_i,s)}-\frac{1}{p_s e(X_i,s)}\Big)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))^2}\\&\leq\frac{1}{\delta}\sqrt{\sum_{k\in\{1,\dots,K\}}\frac{1}{n}\sum_{i\in\mathcal{I}_k}\Big(\frac{1}{\hat{e}^{(-k(i))}(X_i,s)\hat{p}^{-k(i)}}-\frac{1}{e(X_i,s)p_s}\Big)^2}\sqrt{\sum_{k\in\{1,\dots,K\}}\frac{1}{n}\sum_{i\in\mathcal{I}_k}(m_{1,s}(X_i)-\hat{m}_{1,s}^{(-k(i))}(X_i))^2}.\end{aligned}\tag{A.25}$$

By the bounded difference inequality, and the union bound we obtain that the following holds:

$$\begin{aligned}&\sqrt{\sum_{k\in\{1,\dots,K\}}\frac{1}{n}\sum_{i\in\mathcal{I}_k}\Big(\frac{1}{\hat{e}^{(-k(i))}(X_i,s)\hat{p}^{-k(i)}}-\frac{1}{e(X_i,s)p_s}\Big)^2}\sqrt{\sum_{k\in\{1,\dots,K\}}\frac{1}{n}\sum_{i\in\mathcal{I}_k}(m_{1,s}(X_i)-\hat{m}_{1,s}^{(-k(i))}(X_i))^2}\\&\leq K\sqrt{\mathbb{E}\Big[\Big(\frac{1}{\hat{e}(X_i,s)\hat{p}}-\frac{1}{e(X_i,s)p_s}\Big)^2\Big]}\sqrt{\mathbb{E}\Big[(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))^2\Big]}\\&+2\sqrt[4]{\log(2K/\gamma)/n}\sqrt{\mathbb{E}\Big[\Big(\frac{1}{\hat{e}(X_i,s)\hat{p}}-\frac{1}{e(X_i,s)p_s}\Big)^2\Big]}+2\sqrt[4]{\log(2K/\gamma)/n}\sqrt{\mathbb{E}\Big[(m_{1,s}(X_i)-\hat{m}_{1,s}(X_i))^2\Big]}\\&+2\sqrt{\log(2K/\gamma)/n},\end{aligned}\tag{A.26}$$

with probability at least $1-\gamma$. Under Assumption 4.3 and the union bound, the result completes since $K$ is a finite number. □

**Lemma A.4.** *Let*

$$G(\alpha) = \sup_{\pi \in \Pi} \Big\{ \alpha W_0(\pi) + (1-\alpha) W_1(\pi) \Big\} - \sup_{\pi \in \hat{\Pi}_o} \Big\{ \alpha W_0(\pi) + (1-\alpha) W_1(\pi) \Big\}. \tag{A.27}$$

*Define*

$$\mathcal{G} = \{G(\alpha), \alpha \in (0,1)\}.$$

*Under Assumption 4.1, for any $\varepsilon > 0$, there exist a set $\{\alpha_1, ..., \alpha_{N(\varepsilon)}\}$, such that for all $\alpha \in (0,1)$,*

$$|G(\alpha) - \max_{j \in \{1,...,N(\varepsilon)\}} G(\alpha_j)| \leq 4\varepsilon M, \tag{A.28}$$

*and $N(\varepsilon) \leq 1 + 1/\varepsilon$.*

*Proof of Lemma A.4.* We denote $\{\alpha_1, ..., \alpha_{N(\varepsilon)}\}$ an $\varepsilon$-cover of the interval $(0,1)$ with respect to the L1 norm. Namely, $\{\alpha_1, ..., \alpha_{N(\varepsilon)}\}$ are equally spaced numbers between $(0,1)$. Clearly, we have that the covering number $N(\varepsilon) \leq 1 + 1/\varepsilon$. We denote

$$G(\alpha) = \sup_{\pi \in \Pi} \alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \sup_{\pi \in \hat{\Pi}_{\mathrm{o}}} \Big\{ \alpha W_0(\pi) + (1-\alpha) W_1(\pi) \Big\}. \tag{A.29}$$

To characterize the corresponding cover of the function class

$$\mathcal{G} = \{G(\alpha), \alpha \in (0,1)\},$$

we claim that for any $\alpha \in (0,1)$, there exist an $\alpha_j$ in the $\varepsilon$ cover such that

$$|G(\alpha) - G(\alpha_j)| \leq 4\varepsilon M. \tag{A.30}$$

Such a result follows by the argument outlined in the following lines.

*Take $\alpha_j$ closest to $\alpha$.* Consider

$$\begin{aligned}
&|G(\alpha) - G(\alpha_j)| \\
&= \Big| \sup_{\pi \in \Pi} \Big\{ \alpha W_0(\pi) + (1-\alpha) W_1(\pi) \Big\} - \sup_{\pi \in \hat{\Pi}_o} \Big\{ \alpha W_0(\pi) + (1-\alpha) W_1(\pi) \Big\} \\
&- \sup_{\pi \in \Pi} \Big\{ \alpha_j W_0(\pi) + (1-\alpha_j) W_1(\pi) \Big\} + \sup_{\pi \in \hat{\Pi}_o} \Big\{ \alpha_j W_0(\pi) + (1-\alpha_j) W_1(\pi) \Big\} \Big| \\
&\leq \underbrace{\Big| \sup_{\pi \in \Pi} \Big\{ \alpha W_0(\pi) + (1-\alpha) W_1(\pi) \Big\} - \sup_{\pi \in \Pi} \Big\{ \alpha_j W_0(\pi) + (1-\alpha_j) W_1(\pi) \Big\} \Big|}_{(i)} \\
&+ \underbrace{\Big| \sup_{\pi \in \hat{\Pi}_o} \Big\{ \alpha W_0(\pi) + (1-\alpha) W_1(\pi) \Big\} - \sup_{\pi \in \hat{\Pi}_o} \Big\{ \alpha_j W_0(\pi) + (1-\alpha_j) W_1(\pi) \Big\} \Big|}_{(ii)}.
\end{aligned} \tag{A.31}$$

We study $(i)$ and $(ii)$ separately. Consider first $(i)$. We observe the following fact: whenever

$$\sup_{\pi \in \Pi} \alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \sup_{\pi \in \Pi} \alpha_j W_0(\pi) + (1-\alpha_j) W_1(\pi) > 0 \tag{A.32}$$

then we can bound

$$(i) \leq \Big| \alpha W_0(\pi^*) + (1-\alpha) W_1(\pi^*) - \alpha_j W_0(\pi^*) + (1-\alpha_j) W_1(\pi^*) \Big|. \tag{A.33}$$

Here $\pi^* \in \arg\sup_{\pi \in \Pi} \alpha W_0(\pi) + (1-\alpha) W_1(\pi)$. When instead

$$\sup_{\pi \in \Pi} \alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \sup_{\pi \in \Pi} \alpha_j W_0(\pi) + (1-\alpha_j) W_1(\pi) \leq 0 \tag{A.34}$$

we can use the same argument by switching sign, which, with trivial rearrengment reads as

$$(i) \leq \Big| \alpha W_0(\pi^{**}) + (1-\alpha) W_1(\pi^{**}) - \alpha_j W_0(\pi^{**}) + (1-\alpha_j) W_1(\pi^{**}) \Big|. \tag{A.35}$$

Here $\pi^{**} \in \arg\sup_{\pi\in\Pi} \alpha_j W_0(\pi) + (1-\alpha_j)W_1(\pi)$. Therefore we obtain,

$$(i) \leq \sup_{\pi\in\Pi} \Big|\alpha W_0(\pi) + (1-\alpha)W_1(\pi) - \alpha_j W_0(\pi) + (1-\alpha_j)W_1(\pi)\Big| \leq 2|\alpha - \alpha_j|M \quad (A.36)$$

where the last inequality follows by Assumption 4.1 and the triangle inequality. Similar reasoning also applies to $(ii)$. Since $\alpha_j$ was chosen to be the closest to $\alpha$, we have $|\alpha_j - \alpha| \leq \varepsilon$. □

## A.2 Proof of Lemma 2.1

The proof follows similarly to standard microeconomic textbook (Mas-Colell et al., 1995). Let

$$\tilde{\Pi} = \{\pi_\alpha : \pi_\alpha \in \arg\sup_{\pi\in\Pi} \alpha_1 W_0(\pi) + \alpha_2 W_1(\pi), \quad \alpha \in \mathbb{R}^2_+, \alpha_1 + \alpha_2 > 0\}. \quad (A.37)$$

Then we want to show that $\Pi_\circ = \tilde{\Pi}$. Trivially $\tilde{\Pi} \subseteq \Pi_\circ$, since otherwise the definition of Pareto optimality would be violated. Consider now some $\pi^* \in \Pi_\circ$. Then we show that there exist a vector $\alpha \in \mathbb{R}^2_+$, such that $\pi^*$ maximizes the expression

$$\sup_{\pi\in\Pi} \alpha_1 W_0(\pi) + \alpha_2 W_1(\pi). \quad (A.38)$$

Denote the set

$$\mathcal{F} = \{(\tilde{W}_0, \tilde{W}_1) \in \mathbb{R}^2 : \exists \pi \in \Pi : \tilde{W}_0 \leq W_0(\pi) \text{ and } \tilde{W}_1 \leq W_1(\pi)\}. \quad (A.39)$$

Since $(0,0) \in \mathcal{F}$, such a set is non-empty. Notice now that $W_s(\pi)$ is linear is $\pi$ for $s \in \{0,1\}$. Therefore, we obtain that the set $\mathcal{F}$ is a convex set, since it denotes the subgraph of a concave functional. We denote $\bar{W} = (W_0(\pi^*), W_1(\pi^*))$ and $\mathcal{G} = \mathbb{R}^2_{++} + \bar{W}$ the set of welfares that strictly dominates $\pi^*$. Then $\mathcal{G}$ is non-empty and convex. Since $\pi^* \in \Pi_\circ$, we must have that $\mathcal{F} \cap \mathcal{G} = \emptyset$. Therefore, by the separating hyperplane theorem, there exist an $\alpha \in \mathbb{R}^2$, with $\alpha \neq 0$, such that $\alpha^\top F \leq \alpha^\top(\bar{W} + d)$ for any $F \in \mathcal{F}$, $d \in \mathbb{R}^2_{++}$. Let

$d_1 \to \infty$, it must be that $\alpha_1 \in \mathbb{R}_+$, and similarly for $\alpha_2$. So $\alpha \in \mathbb{R}^2_+$. By letting $d \to 0$, we have that $\alpha^\top F \leq \alpha^\top \bar{W}$. This implies that

$$\alpha_1 W_0(\pi) + \alpha_2 W_1(\pi) \leq \alpha_1 W_0(\pi^*) + \alpha_2 W_1(\pi^*) \tag{A.40}$$

for any $\pi \in \Pi$ (since it is true for any $F \in \mathcal{F}$). Hence $\pi^*$ maximizes welfare over all possible feasible allocations once reweighted by $(\alpha_1, \alpha_2)$. Since the maximizer is invariant to multiplication of the objective function by constants, the result follows after dividing the objective function by the sums of the coefficients, which is non-zero by the separating hyperplane theorem. This completes the proof.

## A.3 Proof of Proposition 2.2

First, observe that by rationality, preferences are complete and transitive. Observe also that the preference function equivalently correspond to lexico-graphic with $\pi \succ \pi'$ if $\pi$ Pareto dominates $\pi'$. If instead neither $\pi, \pi'$, Pareto dominates the other, then $\pi \succ \pi'$ is UnFairness $(\pi) <$ UnFairness $(\pi')$. Therefore, it must be that $\mathcal{C}(\Pi) \subseteq \Pi_{\text{o}}$, with $\pi^\star \in \mathcal{C}(\Pi)$ if and only if

$$\pi^\star \in \arg\min_{\pi \in \Pi_{\text{o}}} \text{UnFairness}(\pi).$$

By Lemma 2.1 the result directly follows.

## A.4 Proof of Corollary 1

Define $\widetilde{\Pi} \subseteq \Pi$ the set of policies that satisfy the constraint in Equation (7) (i.e., feasible allocations). By Proposition 2.2 $\widetilde{\Pi} = \Pi_o$. Observe now that $\pi_\omega$ is a feasible allocation under the constraint in Equation (7). This directly implies the conclusion for $\pi_\omega$.

Consider now $\tilde{\pi}$, and fairness constraints not being binding. If $\widetilde{\pi}$ is Pareto optimal, then it represents a feasible allocation (i.e. it satisfies the constraint in Equation (7)). If it is not, then any other allocation that *is* Pareto optimal and Pareto dominates $\widetilde{\pi}$ is feasible under the constraint in Equation (7) completing the proof. Finally, whenever fairness constraints are binding, the estimated policy contains as one possible solution the policy which maximizes the utilitarian welfare under fairness constraints. This follows from the fact that in such case

$$\widetilde{\pi} \in \Big\{ \arg\max_{\pi \in \Pi} p_1 W_1(\pi) + (1-p_1) W_0(\pi) \Big\} \subseteq \Pi_o,$$

since $\Pi = \Pi(\kappa)$.

## A.5 Proof of Theorem 4.1

Throughout the proof we refer to $\bar{C} < \infty$ as a universal constant. We write

$$\begin{aligned} &\sup_{\alpha \in (0,1)} \sup_{\pi \in \Pi} \Big| \alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \max_{\alpha_j \in \{\alpha_1, \dots, \alpha_N\}} \alpha_j \hat{W}_0(\pi) - (1-\alpha_j) \hat{W}_1(\pi) - \lambda/\sqrt{n} \Big| \\ &\leq \underbrace{\sup_{\alpha \in (0,1)} \sup_{\pi \in \Pi} \Big| \alpha W_0(\pi) + (1-\alpha) W_1(\pi) - \alpha \hat{W}_0(\pi) - (1-\alpha) \hat{W}_1(\pi) \Big|}_{(I)} + \frac{\lambda}{\sqrt{n}} \\ &+ \underbrace{\sup_{\alpha \in (0,1)} \sup_{\pi \in \Pi} \Big| \alpha \hat{W}_0(\pi) + (1-\alpha) \hat{W}_1(\pi) - \max_{\alpha_j \in \{\alpha_1, \dots, \alpha_N\}} \alpha_j \hat{W}_0(\pi) - (1-\alpha_j) \hat{W}_1(\pi) \Big|}_{(II)}. \end{aligned} \tag{A.41}$$

$(I)$ is bounded as in Lemma A.3. $(II)$ is bounded as follows.

$$(II) \leq \varepsilon \sup_{\pi \in \Pi} |\hat{W}_0(\pi)| + \varepsilon \sup_{\pi \in \Pi} |\hat{W}_1(\pi)|. \tag{A.42}$$

Under Assumption 4.3, the estimated conditional mean and propensity score are uniformly bounded. Therefore we obtain that

$$\varepsilon \sup_{\pi \in \Pi} |\hat{W}_0(\pi)| + \sup_{\pi \in \Pi} \varepsilon |\hat{W}_1(\pi)| \leq \bar{C} \varepsilon \frac{M}{\delta^2} \leq \bar{C} \frac{M}{N \delta^2}.$$

### A.6 Proof of Theorem 4.2

Recall the definition of $\bar{W}_\alpha$ in Equation (4). The set of Pareto optimal policies reads as follows

$$\pi : \alpha W_1(\pi) + (1-\alpha) W_0(\pi) \geq \bar{W}_\alpha \text{ for some } \alpha \in (0,1).$$

Now it suffices to show for the claim to hold that

$$P\Big(\forall \alpha \in (0,1), \quad \max_{j \in \{1,\cdots,N\}} \bar{W}_\alpha - \bar{W}_{j,n} + \lambda(\gamma)/\sqrt{n} + \frac{b}{N} \geq 0\Big) \leq \gamma,$$

where $\lambda(\gamma) = \underline{b}(\sqrt{v} + \sqrt{\log(2/\gamma)})$, whenever $N = \sqrt{n}$ (and hence $\lambda = \lambda(\gamma) + \underline{b}$). Observe that since $\{\alpha_1, \cdots, \alpha_N\}$ are equally spaced, we have that for all $\alpha \in (0,1)$

$$\sup_{\pi \in \Pi} \alpha W_1(\pi) + (1-\alpha) W_0(\pi) \geq \sup_{\pi \in \Pi} \alpha_j W_1(\pi) + (1-\alpha_j) W_0(\pi) + M\varepsilon$$

for some $j \in \{1, \cdots, N\}$ by Assumption 4.2 (ii). Taking $\underline{b} \geq M, \varepsilon = 1/N$, we have

$$\begin{aligned} &P\Big(\forall \alpha \in (0,1), \quad \max_{j \in \{1,\cdots,N\}} \bar{W}_\alpha - \bar{W}_{j,n} + \lambda(\gamma)/\sqrt{n} + \frac{b}{N} \geq 0\Big) \\ &\leq P\Big(\max_{j \in \{1,\cdots,N\}} \bar{W}_{\alpha_j} - \bar{W}_{j,n} + \lambda(\gamma)/\sqrt{n} \geq 0\Big). \end{aligned}$$

We now observe that the following inequality holds:

$$\begin{aligned} &\sup_{\pi \in \Pi} \alpha_j W_1(\pi) + (1-\alpha_j) W_0(\pi) - \bar{W}_{j,n} \\ &= \sup_{\pi \in \Pi} \Big\{\alpha_j W_1(\pi) + (1-\alpha_j) W_0(\pi)\Big\} - \sup_{\pi \in \Pi} \Big\{\alpha \hat{W}_1(\pi) + (1-\alpha) \hat{W}_0(\pi)\Big\} \\ &\leq 2 \sup_{\pi \in \Pi} \Big|\alpha_j W_1(\pi) + (1-\alpha_j) W_0(\pi) - \alpha_j \hat{W}_1(\pi) + (1-\alpha_j) \hat{W}_0(\pi)\Big|. \end{aligned}$$

By Lemma A.3, with probability at least $1-\gamma$,

$$\sup_{\pi \in \Pi} \max_{\alpha_j, j \in \{1,\cdots,N\}} \Big|\alpha_j W_1(\pi) + (1-\alpha_j) W_0(\pi) - \alpha_j \hat{W}_1(\pi) + (1-\alpha_j) \hat{W}_0(\pi)\Big| \leq \bar{C}\sqrt{\frac{v}{n}} + \bar{C}\sqrt{\frac{\log(2/\gamma)}{n}}$$

for a finite constant $\bar{C}$ independent of $n$. By choosing $\underline{b} \geq 2\bar{C} + M$, the proof completes.

## A.7 Proof of Theorem 4.3

By Theorem 4.2 with probability at least $1-\gamma$, $\Pi_\circ \subseteq \hat{\Pi}_\circ(\lambda)$ with $\hat{\Pi}_\circ(\lambda)$ in Equation (14). As a result, we can write with probability $1-\gamma$,

$$\text{UnFairness}(\hat{\pi}) - \inf_{\pi \in \Pi_\circ}\text{UnFairness}(\pi) \leq \text{UnFairness}(\hat{\pi}) - \inf_{\pi \in \hat{\Pi}_\circ(\lambda)}\text{UnFairness}(\pi).$$

We then write

$$\text{UnFairness}(\hat{\pi}) - \inf_{\pi \in \hat{\Pi}_\circ(\lambda)}\text{UnFairness}(\pi) = \text{UnFairness}(\hat{\pi}) - \hat{\mathcal{V}}_n(\hat{\pi}) + \hat{\mathcal{V}}_n(\hat{\pi}) - \inf_{\pi \in \hat{\Pi}_\circ(\lambda)}\text{UnFairness}(\pi).$$

Since $\hat{\pi}_\lambda \in \hat{\Pi}_\circ(\lambda)$, we have

$$\begin{aligned}
&\text{UnFairness}(\hat{\pi}) - \hat{\mathcal{V}}_n(\hat{\pi}) + \hat{\mathcal{V}}_n(\hat{\pi}) - \inf_{\pi \in \hat{\Pi}_\circ(\lambda)}\text{UnFairness}(\pi) \\
&\leq 2 \sup_{\pi \in \hat{\Pi}_\circ(\lambda)} \Big|\text{UnFairness}(\pi) - \hat{\mathcal{V}}_n(\pi)\Big| \leq 2 \sup_{\pi \in \Pi} \Big|\text{UnFairness}(\pi) - \hat{\mathcal{V}}_n(\pi)\Big|
\end{aligned}$$

where the last equality follows from the fact that $\hat{\Pi}_\circ(\lambda) \subseteq \Pi$. Assumption 4.4 bounds $\sup_{\pi \in \Pi}\Big|\text{UnFairness}(\pi) - \hat{\mathcal{V}}_n(\pi)\Big|$ completing the proof.

## A.8 Proof of Theorem 4.4

For $\widehat{D}(\pi)$ it suffices to observe that

$$\sup_{\pi \in \Pi}\Big|\widehat{W}_1(\pi_1) - \widehat{W}_0(\pi) - W_1(\pi) + W_0(\pi)\Big| \leq \sup_{\pi \in \Pi}\Big|\widehat{W}_1(\pi_1) - W_1(\pi)\Big| + \sup_{\pi \in \Pi}\Big|W_0(\pi) - \widehat{W}_0(\pi)\Big| \tag{A.43}$$

with each term being bounded with probability at least $1-2\gamma$[2], by $\bar{C}\sqrt{v/n}+\bar{C}\sqrt{\log(2/\gamma)/n}$ for a finite constant $\bar{C} < \infty$, similarly to what discussed in the proof of Lemma A.3.

The UnFairness bound follows as a corollary of Theorem 4.3, where here Assumption 4.4 holds with $\mathcal{K}(\Pi, \gamma)n^{-\eta} \lesssim \bar{C}\sqrt{v/n} + \bar{C}\sqrt{\log(2/\gamma)/n}$ for a finite constant $\bar{C} < \infty$, i.e., the bound of Equation (A.43).

[2] $2\gamma$ follows by the union bound.

For $\widehat{C}(\pi)$ the argument follows similarly, after noticing that we can bound

$$\sup_{\pi \in \Pi} \Big| \frac{1}{n\hat{p}_1} \sum_{i=1}^n \pi(X_i) S_i - \mathbb{E}[\pi(X)|S=1] + \frac{1}{n(1-\hat{p}_1)} \sum_{i=1}^n \pi(X_i)(1-S_i) - \mathbb{E}[\pi(X)|S=0] \Big|$$
$$\leq \underbrace{\sup_{\pi \in \Pi} \Big| \frac{1}{n\hat{p}_1} \sum_{i=1}^n \pi(X_i) S_i - \mathbb{E}[\pi(X)|S=1] \Big|}_{(A)} + \underbrace{\sup_{\pi \in \Pi} \Big| \frac{1}{(1-\hat{p}_1)n} \sum_{i=1}^n \pi(X_i)(1-S_i) - \mathbb{E}[\pi(X)|S=0] \Big|}_{(B)}.$$

We proceed by bounding $(A)$,while $(B)$ follows similarly. We have

$$(A) \leq \underbrace{\sup_{\pi \in \Pi} \Big| \frac{1}{p_1 n} \sum_{i=1}^n \pi(X_i) S_i - \mathbb{E}[\pi(X)|S=1] \Big|}_{(i)} + \underbrace{\Big| \frac{1}{p_1} - \frac{1}{\hat{p}_1} \Big|}_{(ii)},$$

where the second component follows by the triangular inequality and the fact that $\pi(X_i)S_i \in \{0,1\}$. We now observe that each summand in $(i)$ is centered around its expectation. Therefore, we can bound $(i)$ using the Radamacher complexity of $\Pi$, with

$$\mathbb{E}[(i)] \leq \frac{2}{\delta} \mathbb{E}\Big[ \sup_{\pi \in \Pi} \Big| \frac{1}{n} \sum_{i=1}^n \sigma_i \pi(X_i) S_i \Big| \Big],$$

with $\sigma_1, \cdots, \sigma_n$ being independent Radamacher random variables. Using the Dudley's entropy bound (see Wainwright (2019)) it is easy to show that the right-hand side is bounded by $\bar{C}\sqrt{v/n}$ for a constant $\bar{C} < \infty$. Finally, using the bounded difference inequality (Boucheron et al., 2003), with probability at least $1-\gamma$,

$$|(i) - \mathbb{E}[(i)]| \leq \bar{C} \sqrt{\frac{\log(2/\gamma)}{n}},$$

for a finite constant $\bar{C}$. The bound on the second component (ii) follows from standard property of the sample mean and the assumption that $\hat{p}_1 \geq \delta$. The final statement follows as a direct corollary of Theorem 4.3.

For $\mathcal{I}(\pi)$ the claim holds since

$$\begin{aligned} &\sup_{\pi \in \Pi} \Big| I_s(\pi) - \hat{I}_s(\pi) \Big| \leq \\ &\underbrace{\sup_{\pi \in \Pi} \Big| \frac{1}{n} \sum_{i=1}^n (\hat{\Gamma}_{1,s,i} - \hat{\Gamma}_{0,s,i}) \pi(X_i, s') - \mathbb{E}\Big[(\Gamma_{1,s,i} - \Gamma_{0,s,i}) \pi(X_i, s')\Big] \Big|}_{(A)} \\ &+ \underbrace{\sup_{\pi \in \Pi} \Big| \frac{1}{n} \sum_{i=1}^n (\hat{\Gamma}_{1,s,i} - \hat{\Gamma}_{0,s,i}) \pi(X_i, s) - \mathbb{E}[(\Gamma_{1,s,i} - \Gamma_{0,s,i}) \pi(X_i, s)] \Big|}_{(B)}. \end{aligned} \tag{A.44}$$

Observe now that under Assumption 4.3, following the same argument in Lemma A.3, we can bound $(A)$ and $(B)$ as follows

$$(A) \lesssim \sqrt{\frac{v}{n}} + \sqrt{\frac{\log(2/\gamma)}{n}}, \quad (B) \lesssim \sqrt{\frac{v}{n}} + \sqrt{\frac{\log(2/\gamma)}{n}}.$$

with probability at least $1-\gamma$. The reader may refer to the proof of Lemma A.3 for details.

## A.9 Proof of Theorem 4.5

First, since $\pi(x, s)$ is constant in $s$ with an abuse of notation we can write $\pi(x)$ as a function of $x$ only. We first observe that we can write

$$C(\pi) = \mathbb{E}\Big[(\frac{(1-S)}{1-p_1} - \frac{S}{p_1})\pi(X)\Big] = \mathbb{E}\Big[\frac{(p_1 - S)}{(1-p_1)p_1}\pi(X)\Big]$$

For the lower bound it suffices to find one distribution which satisfies the condition. We choose $Y(1) = 0$, and $Y(0) = 0$ almost surely, which satisfies the bounded assumption on $Y$. This condition implies that any $\pi \in \Pi$ satisfies Pareto optimality, hence $\Pi_o = \Pi$.

Observe that the expression for $C(\pi)$ corresponds to the risk associated with a classifier

$\pi(X)$ for classifying the sensitive attribute $S$ with loss

$$l(S,\pi(X)) \propto (p_1 - S)\pi(X) = \begin{cases} p_1 - 1 \text{ if } S = 1, \pi(X) = 1 \\ p_1 \text{ if } S = 0, \pi(X) = 1 \\ 0 \text{ otherwise .} \end{cases} .$$

We now proceed following some of the steps in Theorem 14.5 and Theorem 14.6 in Devroye et al. (2013), but introducing modifications in the construction of the set of distributions under consideration and in the data-generating process due to the different loss function and its dependence with $P(S = 1)$ (which itself depends on the distribution of $(X, S)$).[3] We start by choosing $D$ to be distributed as a Bernoulli random variable independent of $(X, S)$. As a result, $(Y, D)$ are independent of $(X, S)$. Therefore, since $(Y, D)$ is independent of $(X, S)$ it suffices to focus on classifiers $\pi_n(X)$ constructed using information $(X_1, S_1), \cdots, (X_n, S_n)$ only. The rest of the proof consists in constructing a distribution of $(X, S)$ such that the lower bound is attained. Recall that classifiers depend on $X$ only and not on $S$ by assumption.

Consider first the case where $(v - 1)/2$ is an integer. The case where it is not follows similarly to below and discussed at the end of the proof. We construct a family of distributions for $(X, S)$, defined $\mathcal{F}$ as follows: first we find points $x_1, \cdots x_v$ that are shattered by $\Pi_o$. Each distribution in $\mathcal{F}$ is concentrated on the set of these points. A member in $\mathcal{F}$ is described by $v - 1$ bits $b_1, \cdots, b_{v-1}$. This is representated as a bit vector $b \subset \{0, 1\}^{v-1}$. Each bit vector that we consider is assumed to sum to $(v - 1)/2$, namely

$$\sum_{i=1}^{v-1} b_i = \frac{v-1}{2}.$$

Assume that $v - 1 \leq n$. For each vector $b$, we let $X$ put mass $m$ at $x_i, i < v$, and mass $1 - (v - 1)m$ at $x_v$. This imposes the condition $(v - 1)m \leq 1$, which will be satisfied. We choose for all $b$ that we consider $P(S = 1) = p_1 \in (\delta, 1 - \delta)$ which we choose later in the

[3] The lack of restriction on the error of the classifier represents a further difference.

proof. Next, introduce the constant $c \in (0, p_1)$. Let $U$ a uniform random variable on $[0, 1]$,

$$S = \begin{cases} 1 \text{ if } U \leq p_1 - c + 2cb_i, X = x_i, i < v \\ 1 \text{ if } U \leq p_1, X = x_v \\ 0 \text{ otherwise} \end{cases}.$$

Thus for $X = x_i, i < v$, $S$ is one with probability $p_1 - c$ or $p_1 + c$, while for $X = x_v$ $S$ is one with probability $p_1$. Now observe that the choice of $S$ and the fact that $P(S = 1) = p_1$ implies that

$$p_1 = \sum_{i=1}^{v-1} m(p_1 - c + 2cb_i) + p_1(1 - m(v-1)) = (v-1)mp_1 + p_1(1 - m(v-1)), \quad \text{(A.45)}$$

since $c\sum_{i=1}^{v-1} b_i = c\frac{v-1}{2}$ by the restriction on $b \in \mathcal{B}$. The above expression is satisfied for any $m$, so no restrictions on $m$ are implied by the Equation (A.45). With a simple argument, it is easy to show that one of the best rules[4] for $b$ is the one which sets

$$f_b(x) = \begin{cases} 1 \text{ if } x = x_i, i < v, b_i = 1 \\ 0 \text{ otherwise.} \end{cases}$$

Such rule is feasible since it has VC-dimension $v$. Notice now that we can write for the decision rule $f_b(x)$, $\mathbb{E}[l(S, f_b(X))|X = x_i] = -c$ for $i < v$, for fixed $b$. Observe now that we can write for any $\pi_n, X \in \{x_1, \cdots, x_{v-1}\}$ , for fixed $b$,

$$\mathbb{E}[l(S, \pi_n(X))|X] - \mathbb{E}[l(S, f_b(X))|X] \geq 2c1\{\pi_n(X) \neq f_b(X)\},$$

since if $\pi_n(X) = 1 - f_b(X)$, then $\mathbb{E}[l(S, 1 - f_b(X))|X] = c$. Therefore we can bound for

[4] A different which leads to the same objective is the one that classifies one also for $X = x_v$. This would be indifferent with respect to $f_b$ since the loss function at $X = x_v$ is always zero in expectation for either prediction.

any $\pi_n$, and a fixed $b$

$$\begin{aligned}\text{UnFairness}(\pi_n) - \inf_{\pi \in \Pi_o} \text{UnFairness}(\pi) \propto \mathbb{E}[l(S, \pi_n(X))] - \inf_{\pi \in \Pi} \mathbb{E}[l(S, \pi(X))] \\ \geq \sum_{j=1}^{v-1} 2mc1\{\pi_n(x_j) = 1 - f_b(x_j)\} \\ \geq \sum_{j=1}^{v-1} 2mc1\{\pi_n(x_j) = 1 - f_b(x_j)\}.\end{aligned} \tag{A.46}$$

Since we take the supremum over the class of distribution $P_b \in \mathcal{F}$ indexed by the bit-vector $b$, it suffices to provide upper bound with respect to $b$ being a random variable and take expectations over $b$. We replace $b$ by a uniformly distributed random variable $B$ over $\mathcal{B} \subset \{0,1\}^{v-1}$, where $\mathcal{B}$ is the set of bit vectors which sum to $(v-1)/2$. We observe that for any $t \geq 0$,

$$\begin{aligned}&\sup_{(X,S) \in \mathcal{F}} P\Big(\text{UnFairness}(\pi_n) - \inf_{\pi \in \Pi_o} \text{UnFairness}(\pi) > t\Big) \\ &= \sup_{b} P\Big(\text{UnFairness}(\pi_n) - \inf_{\pi \in \Pi_o} \text{UnFairness}(\pi) > t\Big) \\ &\geq \mathbb{E}_b\Big[1\{\text{UnFairness}(\pi_n) - \inf_{\pi \in \Pi_o} \text{UnFairness}(\pi) > t\}\Big] \text{ (with random b)} \\ &\geq \mathbb{E}_b\Big[1\Big\{\sum_{j=1}^{v-1} 2mc1\{\pi_n(x_j) = 1 - f_b(x_j)\} > t\Big\}\Big]\end{aligned}$$

where the last inequality uses Equation (A.46) and the monotonicity of the indicator function. We can now write

$$\begin{aligned}&\mathbb{E}_b\Big[1\Big\{\sum_{j=1}^{v-1} 2mc1\{\pi_n(x_j) = 1 - f_b(x_j)\} > t\Big\}\Big] \\ &= \frac{1}{|\mathcal{B}|} \sum_{(x_1, \cdots, x_n', s_1, \cdots, s_n) \in (\{x_1, \cdots, x_v\} \times \{0,1\})^2} \\ &\sum_{b \in \mathcal{B}} 1\Big\{\sum_{j=1}^{v-1} 2mc1\{\pi_n(x_j) = 1 - f_b(x_j)\} > t\Big\} \prod_{j=1}^{n} p_b(x_j', s_j)\end{aligned}$$

with $p_b(x_j', s_j)$ denoting the joint probability of $x_j', s_j$. For a fixed $b$, define $b^c = (1 -$

$b_1, \cdots, 1-b_{v-1})$. Observe that if $b \in \mathcal{B}$, then $b^c \in \mathcal{B}$ since we assumed that $(v-1)/2$ is an integer. Now observe that if

$$\frac{t}{2mc} \leq (v-1)/2, \tag{A.47}$$

then

$$1\Big\{\sum_{j=1}^{v-1} 2mc1\{\pi_n(x_j) = 1 - f_b(x_j)\} > t\Big\} + 1\Big\{\sum_{j=1}^{v-1} 2mc1\{\pi_n(x_j) = 1 - f_{b^c}(x_j)\} > t\Big\} \geq 1$$

since it must be that either (or both) indicators are equal to one. Therefore for $t\Big/2mc \leq (v-1)/2$, the last expression in the lower bound above is bounded from below by

$$\frac{1}{|\mathcal{B}|} \sum_{(x_1,\cdots,x'_n,s_1,\cdots,s_n)\in(\{x_1,\cdots,x_v\}\times\{0,1\})^2} \sum_{b\in\mathcal{B}} \frac{1}{2} \min\Big\{\prod_{j=1}^{n} p_b(x'_j, s_j), \prod_{j=1}^{n} p_{b^c}(x'_j, s_j)\Big\}.$$

By LeCam's inequality, we have that the above expression is bounded from below by (see Page 244 in Devroye et al. 2013)

$$\frac{1}{4|\mathcal{B}|} \sum_{b\in\mathcal{B}} \Big(\sum_{(x,s)} \sqrt{p_b(x,s)p_{b^c}(x,s)}\Big)^{2n}.$$

Observe that we have for $x = x_v$,

$$p_b(x,1) = p_{b^c}(x,1) = p_1(1 - m(v-1)), \quad p_b(x,1) = p_{b^c}(x,1) = (1-p_1)(1-m(v-1)).$$

For $x = x_i, i < v$, we have

$$p_b(x,s)p_{b^c}(x,s) = m^2(p_1^2 - c^2), \quad s \in \{0,1\}.$$

Therefore, we obtain

$$\begin{aligned}\sum_{(x,s)} \sqrt{p_b(x,s)p_{b^c}(x,s)} &= (1 - m(v-1)) + 2(v-1)m\sqrt{(p_1^2 - c^2)} \\ &= (1 - (v-1)m) + 2(v-1)m\sqrt{(p_1^2 - c^2)}.\end{aligned}$$

Hence we can write

$$\frac{1}{4|\mathcal{B}|}\sum_{b\in\mathcal{B}}\sum_{(x,s)}\sqrt{p_b(x,s)p_{b^c}(x,s)} = \frac{1}{4}\Big\{(1-(v-1)m)+2(v-1)m\sqrt{(p_1^2-c^2)}\Big\}.$$

Define $F = m(v-1)(p_1 - c)$. Then we can write

$$\begin{aligned}\frac{1}{4|\mathcal{B}|}\sum_{b\in\mathcal{B}}\Big(\sum_{(x,s)}\sqrt{p_b(x,s)p_{b^c}(x,s)}\Big)^{2n} &= \frac{1}{4}\Big\{(1-(v-1)m)+2(v-1)m\sqrt{(p_1^2-c^2)}\Big\}\\ &= \frac{1}{4}\Big\{1-\frac{F}{p_1-c}\Big(1-\sqrt{4p_1^2-4c^2}\Big)\Big\}^{2n}.\end{aligned}$$

We now choose $p_1 = 1/2$. We can now follow Devroye et al. (2013), end of Page 244 and write

$$\begin{aligned}\frac{1}{4}\Big\{1-\frac{F}{p_1-c}\Big(1-\sqrt{4p_1^2-4c^2}\Big)\Big\}^{2n} &= \frac{1}{4}\Big\{1-\frac{F}{p_1-c}\Big(1-\sqrt{1-4c^2}\Big)\Big\}^{2n}\\ &\geq \frac{1}{4}\Big\{1-\frac{F}{p_1-c}4c^2\Big\}^{2n}\\ &\geq \frac{1}{4}\exp\Big(-\frac{16nFc^2}{1-2c}\Big/\Big(1-\frac{8Fc^2}{1-2c}\Big)\Big),\end{aligned}$$

where we used $1-x \geq e^{-x/(1-x)}$.

We now choose $c = \frac{t}{(v-1)m}$, which satisfies Equation (A.47), and where we need the condition that $0 < t \leq \frac{(v-1)m}{2}$ which we check later in the proof. We write

$$\frac{16nFc^2}{1-2c}\Big/\Big(1-\frac{8Fc^2}{1-2c}\Big) = \frac{16nFc^2}{1-2c-8Fc^2}.$$

Fix a constant $h \in (0,1)$ whose conditions will be discussed below together with the conditions for $t$. Take $t, h$ such that $1-2c-8Fc^2 \geq h \in (0,1)$. Then it follows that (since $c = \frac{t}{(v-1)m}$)

$$\frac{16nFc^2}{1-2c-8Fc^2} \leq \frac{16nt^2F}{(v-1)^2m^2h}.$$

Hence, the lower bound reads as follows:

$$\sup_{(X,S)\in\mathcal{F}} P\Big(\text{UnFairness}(\pi_n) - \inf_{\pi\in\Pi_o} \text{UnFairness}(\pi) > t\Big) \geq \frac{1}{4}\exp\Big(-\frac{16nt^2F}{(v-1)^2m^2h}\Big).$$

Let $\frac{1}{4}\exp\Big(-\frac{16nt^2F}{(v-1)^2m^2h}\Big) = \kappa$. By re-arranging the expression, we write with probability at least $\kappa$, for some distribution in $\mathcal{F}$, for all $\pi_n$,

$$\text{UnFairness}(\pi_n) - \inf_{\pi\in\Pi_o} \text{UnFairness}(\pi) \geq \sqrt{\frac{F(v-1)^2m^2\log(\frac{1}{4\kappa})}{16nh}} \tag{A.48}$$

where we chose $t = \sqrt{\frac{F(v-1)^2m^2\log(\frac{1}{4\kappa})}{16nh}}$.

Next, we check the condition for $t, h$, and characterize the constants $m, h, F$. Recall that the conditions are the following:

$$0 < t \leq \frac{(v-1)m}{2}, \quad 1 - 2c - 8Fc^2 \geq h, \quad c = \frac{t}{(v-1)m}, \quad F = m(v-1)(\frac{1}{2} - c), \quad 0 < m \leq \frac{1}{v-1},$$
$$t = \sqrt{\frac{F(v-1)^2m^2\log(\frac{1}{4\kappa})}{16nh}}, \quad h \in (0,1),$$

where the first condition on $t$ follows from Equation (A.47). Take first $h = F/8$. Then the first condition on $t$ implies that $n \geq \log(1/4\kappa)$. The second condition on $h$ (with $h = F/8$) is satisfied if the first inequality holds

$$1 - F/8 \geq c(2 + 4F) \geq c(2 + 8Fc)$$

since $c \in (0, 1/2)$. Now, observe that $F \leq 1/2$, hence it suffices to show that

$$c \leq \frac{1 - 1/16}{4} \Rightarrow \sqrt{\frac{\log(\frac{1}{4\kappa})}{2n}} \leq \frac{15}{64} \Rightarrow n \geq \bar{C}\log(1/4\kappa),$$

for a finite constant $\bar{C}$. The proof completes since the remaining conditions can be satisfied for an arbitrary choice of $0 < m < 1/(v-1)$.

We are left to show that the claim holds if $(v-1)/2$ is not an integer. For this case we follow the same steps of the proof where we construct a set of distributions $\mathcal{F}$ which puts mass $m$ on $v-2$ $x_i, i < v-1$ and mass $\frac{1-(v-2)m}{2}$ on the remaining $x_{v-1}, x_v$. We construct a bit vector $b \in \mathcal{B} \subset \{0,1\}^{v-2}$ with $\sum_{i=1}^{v-2} b_i = \frac{v-2}{2}$ which must be equal to an integer since $\frac{v-1}{2}$ is not. We construct (since $v \geq 3$)

$$S = \begin{cases} 1 \text{ if } U \leq p_1 - c + 2cb_i, X = x_i, i < v-1 \\ 1 \text{ if } U \leq p_1, X = x_i, i \in \{v-1, v\} \\ 0 \text{ otherwise} \end{cases},$$

while the remaining part of the proof follows similarly to above.

## A.10 Regret bounds for $|D(\pi)|$, and $|C(\pi)|$

To obtain UnFairness bounds for unfairness being defined as either $D(\pi)$ or $C(\pi)$ *in absolute value* it suffices to bound the following empirical processes

$$\sup_{\pi \in \Pi} \Big| |\hat{C}(\pi)| - |C(\pi)| \Big|, \quad \sup_{\pi \in \Pi} \Big| |\hat{D}(\pi)| - |D(\pi)| \Big|.$$

We bound the first on the left-hand side while the second follows similarly. We write by the reverse triangular inequality

$$\sup_{\pi \in \Pi} \Big| |\hat{C}(\pi)| - |C(\pi)| \Big| \leq \sup_{\pi \in \Pi} \Big| \hat{C}(\pi) - C(\pi) \Big|.$$

The rest of the proof follows similarly to Theorem B.3.

## A.11 Proofs in Section 4.4

### A.11.1 Proof of Proposition 4.6

For simplicity we assume that $\beta_0 = \beta_1 = \beta \in [0,1]^p$, while our reasoning directly extend to different $\beta_0, \beta_1$. To analyze the computational complexity of the algorithm, we first, need to compute the computational complexity of each operation needed to estimated $\bar{W}_{j,n}$. Note that each optimization problem to estimate $\bar{W}_{j,n}$ is a linear program with $p$ variables and constraint $\beta^{(j)} \in [0,1], 1 \leq j \leq p$. Therefore, using standard arguments (Papadimitriou and Steiglitz, 1998, Theorems 8.2, 8.5), each program admits an exact solution in $\mathcal{O}(p^\omega)$ running time, for a finite constant $\omega$. There are $\sqrt{n}$ many of such programs, with overall running time $\mathcal{O}(\sqrt{n}p^\omega)$. Consider now the optimization program in Equation (16). Suppose first that $g(x) = x$. Then we can write the program as follow: for each constraint (i.e., each $\alpha_j$ with corresponding $\bar{W}_{j,n}$) in (B), we construct one program where (B) must hold for a single $\alpha_j$ only, and where we drop (C), (E) and replace (D) with $\beta^{(j)} \in [0,1], 1 \leq j \leq p$. We have in total $\sqrt{n}$ many of such programs. These programs are linear programs with $p+1$ constraints and $p$ variables. Similarly to what discussed above, each of this program can be solved with running time $\mathcal{O}(p^\omega)$ for some finite $\omega$. Once we solve each of this program, the solution to Equation (16) is obtained by finding the smallest objective among the $\sqrt{n}$ many programs. The running time for finding the minimum from $\sqrt{n}$ many elements is $\mathcal{O}(\sqrt{n})$. Therefore, the overall complexity of the optimization is $\mathcal{O}(\sqrt{n}p^\omega)$. Consider now the case where $g(x) = |x|$. In such a case, for each sub-problem which substitute (B) in Equation (16) with a single constraint for a given $\alpha_j$, we can write two sub-problems. The first, is the optimization under the constraint that $x \geq 0$ and the second is the optimization under the constraint that $x \leq 0$, with objective function multiplied by $-1$, where $x$ denotes the argument of the function $g(\cdot)$ (i.e., $\sum_{i=1}^n \hat{F}_i \pi(X_i)$). Again, each subproblem is a linear program with computational complexity $\mathcal{O}(p^\omega)$ which completes the proof.

## A.12 Proof of Proposition 4.7

Note first that $\bar{W}_{j,n}^\delta \leq \bar{W}_{j,n}$ for all $j, \delta$. Therefore, we obtain that the constraint in (B) imposed when estimating $\hat{\pi}_\lambda^\delta$ is less restrictive than the constraint when solving Equation (16). It follows that by construction of the algorithm and the early stopping time to optimize UnFairness,

$$\mathcal{V}_n(\hat{\pi}_\lambda^\delta) - \mathcal{V}_n(\hat{\pi}_\lambda) \leq \delta.$$

We can then write

$$\begin{aligned}\text{UnFairness}(\hat{\pi}_\lambda^\delta) - \inf_{\pi \in \Pi_o} \text{UnFairness}(\pi) = &\underbrace{\text{UnFairness}(\hat{\pi}_\lambda^\delta) - \mathcal{V}_n(\hat{\pi}_\lambda^\delta)}_{(A)} + \underbrace{\mathcal{V}_n(\hat{\pi}_\lambda^\delta) - \mathcal{V}_n(\hat{\pi}_\lambda)}_{(B)} \\ &+ \underbrace{\mathcal{V}_n(\hat{\pi}_\lambda) - \text{UnFairness}(\hat{\pi}_\lambda)}_{(C)} \\ &+ \underbrace{\text{UnFairness}(\hat{\pi}_\lambda) - \inf_{\pi \in \Pi_o} \text{UnFairness}(\pi)}_{(D)}.\end{aligned}$$

We can now bound

$$(A) + (C) \leq 2 \sup_{\pi \in \Pi} \Big| \text{UnFairness}(\pi) - \mathcal{V}_n(\pi) \Big|, \quad (B) \leq \delta.$$

The rest of the proof follows similarly to the one of Theorem 4.4.

# B Extensions and Mathematical Details

## B.1 Comparison under strong duality

We now sketch the differences in the optimization problem with the one in Equation (9) assuming strong duality for expositional convenience, and providing an intuition on the result in Corollary 1 for $\tilde{\pi}$. Assuming strong-duality, the optimization problem of maximizing welfare under fairness constraint in Equation (9) can be equivalently re-written as:

$$\tilde{\pi} \in \arg\min_{\pi \in \Pi} \text{UnFairness}(\pi), \quad \text{such that } p_1 W_1(\pi) + (1 - p_1) W_0(\pi) \geq \lambda(\kappa) \tag{B.1}$$

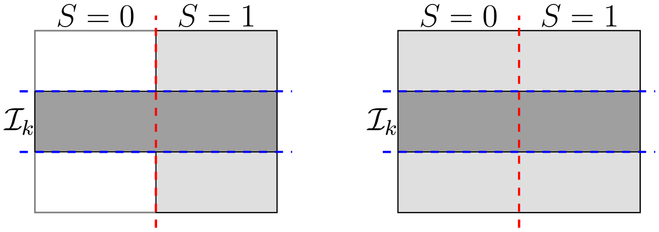
for some constant $\lambda(\kappa) \leq \bar{W}_{p_1}$ which depends on $\kappa$. We now constrast Equation (B.1) to our proposed approach (Equation (7)). Suppose first that $\lambda(\kappa) = \bar{W}_{p_1}$, i.e., $\widetilde{\pi}$ *is* Pareto optimal. Then the constraint in Equation (B.1) is *stricter* than the constraint in Equation (7), since the latter case imposes that $\alpha W_1(\pi) + (1 - \alpha) W_0(\pi) \geq \bar{W}_\alpha$, for *some* $\alpha$, instead of for a particular chosen weight (e.g., $p_1$). As a result, $\pi^\star$ leads to a lower level of UnFairness whenever $\widetilde{\pi}$ *is* Pareto optimal, since $\pi^\star$ minimizes UnFairness under weaker constraints compared to $\widetilde{\pi}$. When instead $\widetilde{\pi}$ is *not* Pareto optimal, i.e., $\lambda(\kappa) < \bar{W}_{p_1}$, $\widetilde{\pi}$ is Pareto dominated by some other allocation $\widetilde{\widetilde{\pi}}$. However $\widetilde{\widetilde{\pi}}$ leads to a larger UnFairness than $\pi^\star$, while not Pareto dominating $\pi^\star$.

The key intuition is the following: under strong duality, the dual of $\widetilde{\pi}$ corresponds to minimize UnFairness for *one particular* weighted combination of welfare exceeding a certain threshold. In contrast, our decision problem imposes the constraint that *some* weighted combination of welfares exceeding a certain threshold. This difference reflects the difference between the lexicographic preferences that we propose as opposed to an additive social planner's utility. It guarantees that whenever $\widetilde{\pi}$ is Pareto optimal, its fairness is dominated from the one under $\pi^\star$.

## B.2 Cross-fitting with UnFairness

In this section we discuss cross-fitting with fairness. Two alternative cross-fitting procedures are available to the researcher. The first one, consists in dividing the sample into $K$ folds and estimating the conditional mean $\hat{m}_{d,s}^{(-k(i))}(X_i)$ using observations for which $S = s$ only, after excluding the fold $k$ corresponding to unit $i$ (panel on the right in Figure B.1). Formally, let $i \in \mathcal{I}_k \cap \mathcal{S}_1$ where $\mathcal{I}_k$ is the $k$-th fold of the data and $\mathcal{S}_1 = \{i : S_i = s_1\}$. Let $\hat{m}_{d,s_2}^{(-k(i))}$ be an estimator obtained using samples not in the fold $k$, $\mathcal{I}_k^c \cap \mathcal{S}_1^c$ for which $\mathcal{S}_1^c = \{i : S_i = s_2\}$; for example by a random forest or linear regression of $Y_j$ onto $X_j$ for $S_j = s_2$, and $j \notin \mathcal{I}_k$. Such an approach does not impose parametric restrictions on the dependence of $m_{d,s}$ on the attribute $s$, at the expense of shrinking the effective sample size used for estimation. The second approach consists in further imposing additional parametric restrictions on the depends of $m_{d,s}$ on $s$ and using all observations in all folds except $k$ for estimating $\hat{m}_{d,s}^{(-k(i))}(X_i)$ (panel on the right in Figure B.1).

Figure B.1: Graphical representation of cross-fitting under two alternative model formulation. The light gray area is the training set, used to construct an estimator of $\hat{m}_{d,s=1}$, whereas the darker gray area is an evaluation set, area in which a prediction of $\hat{m}_{d,s=1}$ is computed.



## B.3 Linear or quadratic constraints for the policy function space representation

In this section we discuss mixed integer formualtions of probabilistic and deterministic decisions rules. Consider first a deterministic decision rule of the form

$$\Pi = \Big\{\pi_\beta(X, S) = 1\{X^\top \beta + S\beta_0 > 0\}, \quad \beta \in \mathcal{B}\Big\}.$$

Then we can write the constraint (A) in Equation (16) as (Kitagawa and Tetenov, 2018)

$$\frac{X_i^\top \beta + s\beta_0}{|C_i|} < z_{s,i} \leq \frac{X_i^\top \beta + s\beta_0}{|C_i|} + 1, \quad C_i > \sup_{\beta \in \mathcal{B}} |X_i^\top \beta| + |\beta_0|, \quad z_{s,i} \in \{0, 1\}.$$

Consider now the following probabilistic decision

$$\Pi = \Big\{\pi_\beta(X, S) = p_1 1\{X_i^\top \beta + S\beta_0 > 0\} + p_0 1\{X^\top \beta + S\beta_0 \leq 0\}, p_1, p_0 \in [0, 1], \beta \in \mathcal{B}\Big\}. \tag{B.2}$$

Then we can represent each decision variable as follows

$$\begin{aligned} & z_{s,i} = p_1 \xi_{s,i} + p_0(1 - \xi_{s,i}) \\ & \frac{X_i^\top \beta + s\beta_0}{|C_i|} < \xi_{s,i} \leq \frac{X_i^\top \beta + s\beta_0}{|C_i|} + 1, \quad C_i > \sup_{\beta \in \mathcal{B}} |X_i^\top \beta| + |\beta_0|, \quad \xi_{s,i} \in \{0, 1\}. \end{aligned}$$

where we introduced the additional variables $\xi_{s,i}$. We use this probablistic rule in the empirical application.

One last type of function class of interest is a linear probability rule of the following form

$$z_{s,i} = X_i^\top \beta + \beta_0 S_i, \quad z_{s,i} \in [0,1]$$

which leads to fast computations due lack of integer variables in the program.

## B.4 Extension: Additional Notions of UnFairness

### B.4.1 Predictive Parity

Predictive parity has been discussed in Kasy and Abebe (2021) among others. Here we consider its definition within the context of policy-targeting. Its notion requires additional assumption for its implementation, assuming *deterministic* treatment assignments $\pi(X_i) \in \{0,1\}$ (i.e., $\mathcal{T} = \{0,1\}$). The notion reads as follows:

$$P_s(\pi) = \Big|\mathbb{E}\Big[Y(1)\Big|\pi(X) = 1, S = s\Big] - \mathbb{E}\Big[Y(1)\Big|\pi(X) = 1\Big]\Big|.$$

Larger values of $P_s(\pi)$ increase UnFairness. Using the definition of the conditional expectation, and using consistency of potential outcomes, the following lemma holds.

**Lemma B.1.** *Let $\mathcal{T} = \{0,1\}$. Then following holds.*

$$P_s(\pi) = (1-p_s)\Big|\frac{\mathbb{E}\Big[Y(1)1\{S_i = s\}\pi(X)\Big]}{p_s\mathbb{P}(\pi(X) = 1|S = s)} - \frac{\mathbb{E}\Big[Y(1)\pi(X)1\{S = s'\}\Big]}{(1-p_s)\mathbb{P}(\pi(X) = 1|S = s')}\Big|.$$

*Proof of Lemma B.1.* Using the definition of conditional expectation:

$$\mathbb{E}\Big[Y\Big|\pi(X) = 1, S = s\Big] = \mathbb{E}\Big[\frac{Y(1)1\{S = s\}\pi(X)}{p_s P(\pi(X) = 1|S = s)}\Big]. \tag{B.3}$$

We also write

$$\mathbb{E}\Big[Y\Big|\pi(X) = 1\Big] = p_s\mathbb{E}\Big[Y\Big|\pi(X) = 1, S = s\Big] + (1-p_s)\mathbb{E}\Big[Y\Big|\pi(X) = 1, S = s'\Big].$$

Combining the expression with Equation (B.3) completes the proof. □

Given two sensitive groups $\mathcal{S} = \{0, 1\}$, the corresponding notion of UnFairness we consider takes the following form:

$$P(\pi) \propto \frac{P_1(\pi)}{1-p_1} = \frac{P_0(\pi)}{p_1}. \tag{B.4}$$

We consider a double-robust estimator which takes the following form:

$$\widehat{\mathcal{V}}_n(\pi) = \Bigg| \frac{\sum_{i=1}^n \pi(X_i)S_i\Big\{\frac{(Y_i-\hat{m}_1(X_i,S_i))D_i}{\hat{e}(X_i,S_i)} + \hat{m}_1(X_i,S_i)\Big\}}{np_1\mathbb{P}(\pi(X_i)=1|S_i=1)} - \frac{\sum_{i=1}^n(1-S_i)\pi(X_i)\Big\{\frac{(Y_i-\hat{m}_1(X_i,S_i))D_i}{\hat{e}(X_i,S_i)} + \hat{m}_1(X_i,S_i)\Big\}}{n(1-p_1)\mathbb{P}(\pi(X_i)=1|S_i=0)} \Bigg|. \tag{B.5}$$

Observe that the estimator depends on the estimated conditional mean function and propensity score, whereas $p_s$ and $\mathbb{P}(\pi(X) = 1|S = s)$ are assumed to be known. These two components can be obtained, for instance from census data, since $p_s$ and $\mathbb{P}(\pi(X) = 1|S = s)$ only depend on the distribution of covariates and sensitive attributes. Whenever $\mathbb{P}(\pi(X_i) = 1|S_i = s)$ is replaced by its sampled analog $\mathbb{P}_n(\pi(X_i) = 1|S_i = s) = \frac{1}{np_s}\sum_{i=1}^n \pi(X_i)1\{S_i = s\}$, we require that $\mathbb{P}_n(\pi(X_i) = 1|S_i = s)$ is bounded away from zero almost surely.

**Theorem B.2** (Predictive parity)**.** *Let Assumptions 2.1, 4.1,4.2, 4.3 hold. Let either* UnFairness$(\pi)$ *be defined using the notion of Predictive (dis)-parity. Assume that* $P(\pi(X, S) = 1|S = 1), P(\pi(X, S) = 1|S = 0) \in (\kappa, 1-\kappa)$ *for all* $\pi \in \Pi$, $\kappa \in (0, 1)$*. Then for some constant* $c_0 < \infty$*, for any* $\gamma \in (0, 1), \lambda \geq \underline{b}\sqrt{\frac{v\log(2/\gamma)}{n}}$*, for a constant* $\underline{b} > 0$*, independent of the sample size with probability at least* $1 - 2\gamma$*,*

$$\text{UnFairness}(\hat{\pi}) - \inf_{\pi\in\Pi_o} \text{UnFairness}(\pi) \leq c_0\sqrt{\frac{\log(2/\gamma)}{n}} + c_0\sqrt{\frac{v}{n}},$$

*for a finite constant* $c_0 < \infty$*.*

**Remark 1** (Mixed-integer linear representation of Predictive Parity)**.** Let $\mathbb{P}(\pi(X_i)|S_i = 1) = \frac{1}{p_sN}\sum_{i=1}^N \pi(X_i)S_i$ where $N$ denotes the number of individuals whose census-information (i.e., baseline covariates and sensitive attributes) are observed. The optimization problem can be formulated as a mixed-integer *fractional* linear program for $\pi(X)$ satisfying a linear representation. This follows after the linearization of the constraint (B), which can be achieved by introducing $2N \times n$ many additional binary variables. Since fractional linear programs admit a mixed-integer linear program representations (Charnes and Cooper, 1962), the optimization problem can be solved as a mixed integer linear program.

*Proof of Theorem B.2.* We write

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{\widehat{P}_s(\pi)}{1-p_s} - \frac{P_s(\pi)}{1-p_s}\Big|\Big] \\
&\le \underbrace{\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{\sum_{i=1}^n \pi(X_i)S_i\Big\{\frac{(Y_i-\hat{m}_1(X_i,S_i))D_i}{\hat{e}(X_i,S_i)} + \hat{m}_1(X_i,S_i)\Big\}}{np_1\mathbb{P}(\pi(X)=1|S=1)} - \mathbb{E}\Big[Y|\pi(X)=1,S=1\Big]\Big|\Big]}_{(A)} \\
&+ \underbrace{\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{\sum_{i=1}^n \pi(X_i)(1-S_i)\Big\{\frac{(Y_i-\hat{m}_1(X_i,S_i))D_i}{\hat{e}(X_i,S_i)} + \hat{m}_1(X_i,S_i)\Big\}}{n\mathbb{P}(\pi(X)=1|S=0)(1-p_1)} - \mathbb{E}\Big[Y|\pi(X)=1,S=0\Big]\Big|\Big]}_{(B)}.
\end{aligned}
$$

We study $(A)$ while $(B)$ follows similarly. First, we write

$$
\begin{aligned}
(A) \le& \underbrace{\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{\sum_{i=1}^n \pi(X_i)S_i\Big\{\frac{(Y_i-\hat{m}_1(X_i,S_i))D_i}{\hat{e}(X_i,S_i)} + \hat{m}_1(X_i,S_i) - \frac{(Y_i-m_1(X_i,S_i))D_i}{e(X_i,S_i)} - m_1(X_i,S_i)\Big\}}{np_1\mathbb{P}(\pi(X)=1|S=1)}\Big|\Big]}_{(I)} \\
&+ \underbrace{\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{\sum_{i=1}^n \pi(X_i)S_i\Big\{\frac{(Y_i-m_1(X_i,S_i))D_i}{e(X_i,S_i)} + m_1(X_i,S_i)\Big\}}{np_1\mathbb{P}(\pi(X)=1|S=1)} - \mathbb{E}\Big[Y|\pi(X)=1,S=1\Big]\Big|\Big]}_{(II)}.
\end{aligned}
$$

We study $(I)$ first. Define

$$
V_n(\pi) = \frac{1}{np_1}\sum_{i=1}^n \pi(X_i)S_i\Big\{\frac{(Y_i-\hat{m}_1(X_i,S_i))D_i}{\hat{e}(X_i,S_i)} + \hat{m}_1(X_i,S_i) - \frac{(Y_i-m_1(X_i,S_i))D_i}{e(X_i,S_i)} - m_1(X_i,S_i)\Big\}.
$$

We have

$$
(I) \le \frac{1}{\kappa}\mathbb{E}\Big[\underbrace{\sup_{\pi\in\Pi}|V_n(\pi)|}_{(a)}\Big].
$$

We write

$$
\begin{aligned}
(a) \leq & \underbrace{\frac{1}{\delta} \sup_{\pi \in \Pi} \Big| \frac{1}{n} \sum_{i=1}^n S_i D_i (Y_i - m_{1,S_i}(X_i)) \Big( \frac{1}{e(X_i, S_i)} - \frac{1}{\hat{e}(X_i, S_i)} \Big) \pi(X_i, S_i) \Big|}_{(j)} \\
& + \underbrace{\frac{1}{\delta} \sup_{\pi \in \Pi} \Big| \frac{1}{n} \sum_{i=1}^n \Big( \frac{D_i}{\hat{e}(X_i, S_i)} - 1 \Big) (m_{1,S_i}(X_i) - \hat{m}_{1,S_i}(X_i)) \pi(X_i, S_i) S_i \Big|}_{(jj)} .
\end{aligned}
\tag{B.6}
$$

We study $(j)$ and $(jj)$ separately. We start from $(j)$. Recall, that by cross fitting $\hat{e}(X_i, S_i) = \hat{e}^{-k(i)}(X_i, S_i)$, where $k(i)$ is the fold containing unit $i$. Therefore, observe that given the $K$ folds for cross-fitting, we have

$$
\begin{aligned}
& \Big| \frac{1}{n} \sum_{i=1}^n S_i D_i (Y_i - m_{1,S_i}(X_i)) \Big( \frac{1}{e(X_i, S_i)} - \frac{1}{\hat{e}(X_i, S_i)} \Big) \pi(X_i, S_i) \Big| \\
& \leq \sum_{k \in \{1, \dots, K\}} \Big| \frac{1}{n} \sum_{i \in \mathcal{I}_k} S_i D_i (Y_i - m_{1,S_i}(X_i)) \Big( \frac{1}{e(X_i, S_i)} - \frac{1}{\hat{e}^{(-k(i))}(X_i, S_i)} \Big) \pi(X_i, S_i) \Big| .
\end{aligned}
\tag{B.7}
$$

In addition, we have that

$$
\begin{aligned}
& \mathbb{E}\Big[ \sum_{i \in \mathcal{I}_k} S_i D_i (Y_i - m_{1,S_i}(X_i)) \Big( \frac{1}{e(X_i, S_i)} - \frac{1}{\hat{e}^{(-k(i))}(X_i, S_i)} \Big) \pi(X_i, S_i) \Big] \\
& = \mathbb{E}\Big[ \mathbb{E}\Big[ \sum_{i \in \mathcal{I}_k} S_i D_i (Y_i - m_{1,S_i}(X_i)) \Big( \frac{1}{e(X_i, S_i)} - \frac{1}{\hat{e}^{(-k(i))}(X_i, S_i)} \Big) \pi(X_i, S_i) \Big| \hat{e}^{(-k(i))} \Big] \Big] = 0,
\end{aligned}
\tag{B.8}
$$

by cross-fitting. By Assumption 4.3, we know that

$$
\sup_{x \in \mathcal{X}, s \in \mathcal{S}} \Big| \frac{1}{e(x, s)} - \frac{1}{\hat{e}^{(-k(i))}(x, s)} \Big| \leq 2/\delta^2 \tag{B.9}
$$

and therefore each summand in Equation (B.7) is bounded by a finite constant $2/\delta^2$. We now obtain, using the symmetrization argument (Van Der Vaart and Wellner, 1996), and the Dudley's entropy integral (Wainwright, 2019)

$$\mathbb{E}\Big[\sup_{\pi\in\Pi}|\frac{1}{n}\sum_{i\in\mathcal{I}_k} S_i D_i(Y_i - m_{1,S_i}(X_i))\Big(\frac{1}{e(X_i,S_i)} - \frac{1}{\hat{e}^{(-k(i))}(X_i,S_i)}\Big)\pi(X_i,S_i)|\Big|\hat{e}^{(-k(i))}\Big] \lesssim \frac{M}{\delta^2}\sqrt{v/n}. \tag{B.10}$$

We now consider the term $(jj)$. Observe that we can write

$$\begin{aligned}(jj) \leq & \underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{D_i}{\hat{e}(X_i,S_i)} - \frac{D_i}{e(X_i,S_i)}\Big)(m_{1,S_i}(X_i) - \hat{m}_{1,S_i}(X_i))S_i\pi(X_i,S_i)\Big|}_{(v)} \\ & + \underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{D_i}{e(X_i,S_i)} - 1\Big)(m_{1,S_i}(X_i) - \hat{m}_{1,S_i}(X_i))\pi(X_i,S_i)S_i\Big|}_{(vv)}.\end{aligned} \tag{B.11}$$

We consider each term seperately. Consider $(vv)$ first. Using the cross-fitting argument we obtain

$$\begin{aligned}&\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{D_i}{e(X_i,S_i)} - 1\Big)(m_{1,S_i}(X_i) - \hat{m}_{1,S_i}(X_i))\pi(X_i,S_i)S_i\Big| \\ &\leq \sum_{k\in\{1,\dots,K\}}\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k}\Big(\frac{D_i}{p_s e(X_i,S_i)} - 1\Big)(m_{1,S_i}(X_i) - \hat{m}_{1,S_i}^{(-k(i))}(X_i))\pi(X_i,S_i)S_i\Big|.\end{aligned} \tag{B.12}$$

Observe now that

$$\mathbb{E}\Big[\Big(\frac{D_i}{e(X_i,S_i)} - 1\Big)(m_{1,S_i}(X_i) - \hat{m}_{1,S_i}^{(-k(i))}(X_i))\pi(X_i,S_i)S_i\Big|\hat{m}_{1,S_i}^{(-k(i))}\Big] = 0. \tag{B.13}$$

Therefore, following the same argument discussed before,

$$\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i\in\mathcal{I}_k}\Big(\frac{D_i}{e(X_i,S_i)} - 1\Big)(m_{1,S_i}(X_i) - \hat{m}_{1,S_i}^{(-k(i))}(X_i))\pi(X_i,S_i)S_i\Big|\Big] \lesssim \frac{M}{\delta^2}\sqrt{\frac{v}{n}}. \tag{B.14}$$

We are now left to bound $(v)$. We obtain that

$$(v) \leq \sqrt{\frac{1}{n}\sum_{i=1}^n\Big(\frac{1}{\hat{e}(X_i,S_i)} - \frac{1}{e(X_i,S_i)}\Big)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n(m_{1,S_i}(X_i) - \hat{m}_{1,S_i}(X_i))^2}. \tag{B.15}$$

Using Jensen inequality and Assumption 4.3 $\mathbb{E}[(v)] \lesssim n^{-1/2}$.

We now move to bound the expectation of $(II)$. First, observe that by Lemma B.1, and standard properties of the double-robust estimator, we have that

$$\mathbb{E}\Big[\frac{\frac{1}{p_1 n}\sum_{i=1}^n \pi(X_i)S_i\Big\{\frac{(Y_i - m_1(X_i,S_i))D_i}{e(X_i,S_i)} + m_1(X_i,S_i)\Big\}}{\mathbb{P}(\pi(X)=1|S=s)}\Big] = \mathbb{E}\Big[Y(1)|\pi(X)=1, S=1\Big].$$

Using the symmetrization argument (see Van Der Vaart and Wellner (1996)), we have

$$(II) \leq 2\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{\frac{1}{p_1 n}\sum_{i=1}^n \pi(X_i)\sigma_i S_i\Big\{\frac{(Y_i - m_1(X_i,S_i))D_i}{e(X_i,S_i)} + m_1(X_i,S_i)\Big\}}{\mathbb{P}(\pi(X)=1|S=s)}\Big|\Big],$$

where $\{\sigma_i\}$ are *i.i.d.* exogenous Radamacher random variables. Using the assumption that $P(\pi(X)=1|S=s)\in(\kappa, 1-\kappa)$, we write

$$\begin{aligned}&\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{\frac{1}{p_1 n}\sum_{i=1}^n \pi(X_i)\sigma_i S_i\Big\{\frac{(Y_i - m_1(X_i,S_i))D_i}{e(X_i,S_i)} + m_1(X_i,S_i)\Big\}}{\mathbb{P}(\pi(X)=1|S=s)}\Big|\Big]\\ &\leq \frac{1}{\kappa}\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{p_1 n}\sum_{i=1}^n \pi(X_i)\sigma_i S_i\Big\{\frac{(Y_i - m_1(X_i,S_i))D_i}{e(X_i,S_i)} + m_1(X_i,S_i)\Big\}\Big|\Big].\end{aligned}$$

We now proceed using a standard argument. Using the fact that each summand in the above expression are uniformly bounded, and $\Pi$ has finite VC-dimension, using the Dudley's entropy integral bound, it directly follows that

$$\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{p_1 n}\sum_{i=1}^n \pi(X_i)\sigma_i S_i\Big\{\frac{(Y_i - m_1(X_i,S_i))D_i}{e(X_i,S_i)} + m_1(X_i,S_i)\Big\}\Big|\Big] \lesssim \sqrt{\frac{v}{n}}$$

which concludes the proof.

□

### B.4.2 Counterfactual envy-freeness

In this section we discuss estimation and guarantees for the counterfactual notion of fairness in Section 5.

We estimate $\mathcal{A}(\cdot)$ as:

$$\begin{aligned}\mathcal{A}_n(s, s'; \pi) =& \frac{1}{n\hat{p}_s} \sum_{i:S_i = s} \Big\{ \hat{m}_{1,s'}(X_i)\pi(X_i, s) + \hat{m}_{0,s'}(X_i)(1 - \pi(X_i, s)) \Big\} \\ &- \frac{1}{n} \sum_{i=1}^{n} \Big\{ \hat{\Gamma}_{1,s,i}\pi(X_i, s) - \hat{\Gamma}_{0,s,i}(1 - \pi(X_i, s)) \Big\}.\end{aligned} \tag{B.16}$$

Whenever we aim not to discriminate in either direction, we take the sum of the effects $\mathcal{A}(s_1, s_2; \pi)$ and $\mathcal{A}(s_2, s_1; \pi)$,[5] and define counterfactual envy-freeness and its estimator as

$$\mathrm{E}(\pi) = \mathcal{A}(1, 0; \pi) + \mathcal{A}(0, 1; \pi), \quad \hat{\mathrm{E}}(\pi) = \mathcal{A}_n(1, 0; \pi) + \mathcal{A}_n(0, 1; \pi). \tag{B.17}$$

**Assumption B.1.** Assume that for some $\zeta > 0$, $\mathbb{E}\Big[\Big(\hat{m}_{d,s_1}(X_i(s_2)) - m_{d,s_1}(X_i(s_2))\Big)^2\Big] = \mathcal{O}(n^{-2\zeta}), \forall s_1, s_2 \in \{0, 1\}, d \in \{0, 1\}$.

Assumption B.1 states that the estimator of the conditional mean function for each sensitive attribute and treatment status $s, d \in \{0, 1\}$, must converge to the true conditional mean function in mean-squared error at some arbitrary rate $2\zeta > 0$. Here, we require convergence in $l_2$ for a given sensitive attribute conditional on the *opposite* sensitive attribute, due to the particular notion of fairness considered.[6] Examples include (i) linear regression models, of the form $m_{d,s}(x) = x\beta_s$, with $\beta_s$ being potentially high-dimensional, and bounded covariates; (ii) local polynomial estimators (Fan and Gijbels, 1996).

**Theorem B.3.** *Let Assumptions 4.1-4.3, 5.1 and B.1 hold. Let* $\mathrm{UnFairness}(\cdot) = \mathrm{E}(\cdot)$ *and* $\hat{\mathcal{V}}_n(\cdot) = \hat{E}(\cdot)$. *Then for some constants* $0 < \underline{b}, c_0 < \infty$ *independent of the sample size, for any* $\gamma \in (0, 1), \lambda \geq \underline{b}(\sqrt{v} + \sqrt{\log(2/\gamma)} + 1), N = \sqrt{n}$, *with probability at least* $1 - 2\gamma$,

$$\mathrm{UnFairness}(\hat{\pi}) - \inf_{\pi \in \Pi_o} \mathrm{UnFairness}(\pi) \leq c_0 \sqrt{\frac{v}{n^{2\zeta}}} + c_0 \sqrt{\frac{\log(2/\gamma)}{n}}.$$

The proof is in Appendix B.4.3. A corollary of Theorem B.3 is that under the parametric rate of convergence of the conditional mean function, the regret bound scales at rate $n^{-1/2}$. Interestingly, the convergence rate is of order slower than $n^{-1/2}$ for non-parametric

[5] Such an approach builds on the notion of "social envy" discussed in Feldman and Kirman (1974).

[6] Namely, to estimate fairness, we need to extrapolate relative to the *opposite* group.

estimators compared to the notions of UnFairness discussed in Section 4. The slower convergence rate is because counterfactual envy-freeness requires estimating the conditional mean function on the population with attribute $S = s_1$ while averaging over the covariates' distribution with the opposite attribute, therefore requiring extrapolation. This result showcases the *trade-off* in the choice of a counterfactual notion of unfairness relative to predictive ones.

### B.4.3 Envy-freeness UnFairness: Proofs

**Lemma B.4.** *Under Assumption4.1, 4.2, 4.3, 5.1, B.1, the following holds: with probability at least* $1 - \gamma$ ,

$$\sup_{\pi \in \Pi} \Big| \mathcal{A}(s, s'; \pi) - \mathcal{A}_n(s, s'; \pi) \Big| \leq \frac{cM}{\delta^2} \sqrt{\frac{\log(2/\gamma)}{n}} + \frac{c}{\delta} n^{-\eta} + \sqrt{\frac{v}{n}} \tag{B.18}$$

*for a universal constant* $c < \infty$.

*Proof of Lemma B.4.* We consider the case where $s' \neq s$, whereas $s' = s$ follows trivially. Observe that we can write

$$\sup_{\pi \in \Pi} \Big| \mathcal{A}(s, s'; \pi) - \mathcal{A}_n(s, s'; \pi) \Big| \leq$$
$$\underbrace{\sup_{\pi \in \Pi} \Big| \mathbb{E}_{X(s)} \Big[ V_{\pi(X(s), s)}(X(s), s') \Big] - \frac{1}{n} \sum_{i=1}^n \Big( \frac{1\{S_i = s\}}{\hat{p}_s} \hat{m}_{1,s'}(X_i) \pi(X_i, s) + \frac{1\{S_i = s\}}{\hat{p}_s} \hat{m}_{0,s'}(X_i)(1 - \pi(X_i, s)) \Big) \Big|}_{(A)}$$
$$+ \underbrace{\sup_{\pi \in \Pi} |\hat{W}_{s'}(\pi) - W_{s'}(\pi)|}_{(B)}. \tag{B.19}$$

The term (B) is bounded as discussed in Lemma A.3. Therefore, we are only left to discuss bounds on (A). To derive bounds in such a scenario, we first observe that we can write

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\Big|\mathbb{E}_{X(s)}\Big[V_{\pi(X(s),s)}(X(s),s')\Big] - \frac{1}{n}\sum_{i=1}^n\Big(\frac{1\{S_i=s\}}{\hat{p}_s}\hat{m}_{1,s'}(X_i)\pi(X_i,s) + \frac{1\{S_i=s\}}{\hat{p}_s}\hat{m}_{0,s'}(X_i)(1-\pi(X_i,s))\Big)\Big| \\
&\le \underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s) - \mathbb{E}\Big[\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s)\Big]\Big|}_{(I)} \\
&+ \underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\frac{1\{S_i=s\}}{p_s}m_{0,s'}(X_i)(1-\pi(X_i,s)) - \mathbb{E}\Big[\frac{1\{S_i=s\}}{p_s}m_{0,s'}(X_i)(1-\pi(X_i,s))\Big]\Big|}_{(II)} \\
&+ \underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{1\{S_i=s\}}{\hat{p}_s}\hat{m}_{1,s'}(X_i) - \frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\Big)\pi(X_i,s)\Big|}_{(III)} \\
&+ \underbrace{\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{1\{S_i=s\}}{\hat{p}_s}\hat{m}_{0,s'}(X_i) - \frac{1\{S_i=s\}}{p_s}m_{0,s'}(X_i)\Big)(1-\pi(X_i,s))\Big|}_{(IV)}.
\end{aligned}
$$

We discuss (I) and (III), whereas (II) and (IV) follow similarly. Observe first that by Assumption 4.1 and the bounded difference inequality, with probability $1-\gamma$,

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s) - \mathbb{E}\Big[\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s)\Big]\Big| \\
&\le \mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s) - \mathbb{E}\Big[\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s)\Big]\Big|\Big] + \bar{C}\frac{M}{\delta}\sqrt{\log(2/\gamma)/n}
\end{aligned}
$$

for a constant $\bar{C} < \infty$. Under Assumption 5.1 each summand is centered around zero. Using the symmetrization argument (Van Der Vaart and Wellner, 1996), we have

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s) - \mathbb{E}\Big[\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s)\Big]\Big|\Big] \le \\
&2\mathbb{E}\Big[\sup_{\pi\in\Pi}\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i\frac{1\{S_i=s\}}{p_s}m_{1,s'}(X_i)\pi(X_i,s)\Big|\Big]
\end{aligned}
$$

where $\sigma_i$ are *i.i.d.* Radamacher random variables. Since $m_{1,s}$ is uniformly bounded and similarly $p_s$ is bounded, and by Assumption 4.1, we obtain by the properties of the Dudley's entropy integral (Wainwright, 2019),

$$\mathbb{E}\Big[\sup_{\pi \in \Pi}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i \frac{1\{S_i = s\}}{p_s} m_{1,s'}(X_i)\pi(X_i, s)\Big|\Big] \leq \bar{C}\frac{M}{\delta}\sqrt{v/n}$$

for a universal constant $\bar{C} < \infty$. We now move to bound (III). Using the triangular inequality and Holder's inequality, we obtain

$$(III) \leq \frac{1}{n}\sum_{i=1}^n \frac{1\{S_i = s\}}{p_s}\Big|m_{1,s'}(X_i) - \hat{m}_{1,s'}(X_i)\Big| \tag{B.20}$$

The above bound is deterministic and it does not depend on $\pi$. Observe now that by consistency of potential outcomes and covariates

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i = s\}}{p_s}\Big|m_{1,s'}(X_i) - \hat{m}_{1,s'}(X_i)\Big| \\ =&\frac{1}{n}\sum_{i=1}^n \frac{1\{S_i = s\}}{p_s}\Big|m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}(X_i(s))\Big| \leq \frac{1}{n\delta}\sum_{i=1}^n \Big|m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}(X_i(s))\Big|.\end{aligned} \tag{B.21}$$

We now separate the contribution of each of the $K$ folds using in the cross-fitting algorithm. Namely, we define

$$\frac{1}{n\delta}\sum_{i=1}^n \Big|m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}(X_i(s))\Big| \leq \sum_{k \in \{1,\dots,K\}} \frac{1}{n\delta}\sum_{i \in \mathcal{I}_k}\Big|m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}^{(-k(i))}(X_i(s))\Big| \tag{B.22}$$

where $\mathcal{I}_k$ denotes the set of indexes in fold $k$, and $\hat{m}_{1,s'}^{(-k(i))}$ denotes the estimator obtained from all folds except $k$. Next, we bound the following term using Liaponuv inequality:

$$\frac{1}{n}\sum_{i \in \mathcal{I}_k}\mathbb{E}\Big[|m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}(X_i(s))|\Big] \lesssim \sqrt{\mathbb{E}\Big[|m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}(X_i(s))|^2\Big]} \leq cn^{-\eta}. \tag{B.23}$$

The last inequality follows by Assumption B.1, for a universal constant $c < \infty$. Finally, we discuss exponential concentration of the empirical counterpart. By boundeness of $\hat{m}$ in

Assumption 4.3, we have

$$\sup_{x\in\mathcal{X}} \Big| m_{d,s'}(x) - \hat{m}_{d,s'}(x) \Big| \le 2M. \tag{B.24}$$

By the bounded difference inequality, with probability at least $1-\gamma$,

$$\frac{1}{n}\sum_{i\in\mathcal{I}_k} \Big| m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}(X_i(s)) \Big| \le \mathbb{E}\Big[\Big| m_{1,s'}(X_i(s)) - \hat{m}_{1,s'}(X_i(s)) \Big|\Big] + 4M\sqrt{\log(2/\gamma)/n}. \tag{B.25}$$

Combining the above bounds, the proof completes. □

**Corollary.** *Theorem B.3 holds.*

*Proof.* This follows from Theorem B.3 and Lemma A.2. □

## B.5 Multi-action policies

In this subsection, we discuss how our results extend to multi-action policies. Using Theorem 1 in Zhou et al. (2018), assuming a bounded entropy integral with respect to $\Pi$ (Assumption 3 in Zhou et al. (2018)), and assumptions in Theorem 4.2, it is easy to show that with probability at least $1-\gamma$

$$\Big| W_d(\pi) - \hat{W}_d(\pi) \Big| = \mathcal{O}\Big(\frac{\kappa(\Pi)}{\sqrt{n}}\Big) + o(1/\sqrt{n}) + \mathcal{O}\Big(\sqrt{\frac{\log(2/\gamma)}{n}}\Big)$$

where $\kappa(\Pi) = \int_0^1 \sqrt{\log(\mathcal{N}(\Pi, \epsilon^2))} d\epsilon$ and $\mathcal{N}(\Pi,\epsilon)$ is the covering number for the function class $\Pi$.[7] Here the first two terms follow directly from the bound on the Rademacher complexity in Zhou et al. (2018) and standard symmetrization arguments (Van Der Vaart and Wellner, 1996), while the last term follows from the bounded difference inequality (Wainwright, 2019) and leverages bounded estimated conditional mean and propensity score.[8] Given such concentration result, it is easy to show that the rest of our proofs

[7] The reader may refer to Wainwright (2019) and Definition 4 in Zhou et al. (2018) for more discussion.

[8] Bounded estimated nuisance functions is not required if we interpret our results as asymptotic in the spirit of Athey and Wager (2021).

of Theorems 4.1, 4.2, 4.3, do not require binary actions, and follow without additional modifications for a finite number of actions.

# C Numerical Studies and Empirical Application: Further Results and Details

## C.1 Empirical Application

**Estimation details** We control for confounding of the treatment assignment by estimating the probability of treatment using a penalized logistic regression, where we condition on the non-Caucasian attribute, gender, the average score, years to graduation, whether the individual had previously had entrepreneurship activities, the startup region (which a dummy since only two regions are considered), the degree (either engineer or business) and the school rank. We estimate the outcome using a penalized logistic regression, after conditioning on the above covariates, and any interaction term between gender, treatment assignment, and a vector of covariates, which include years to graduation, prior entrepreneurship, startup region, and the school rank. We estimate treatment effects using a doubly robust estimator. We use cross-fitting with five folds in our estimation.

**Additional results for probabilistic treatment assignments** We also consider in our analysis the class of *probabilistic* assignment rules, which assign treatments with a probability decision as in Equation (B.2). Results are collected in Figure C.1, where we observe that the set of probabilistic decision Pareto dominates the determinitic ones up-to a small optimization error.

## C.2 Numerical Studies

In this subsection, we include additional results for the numerical studies. In Figure C.2 we report the computational time for different number of covariates. The figure shows that the linear probability rule and optimal tree present better scalability than the maximum

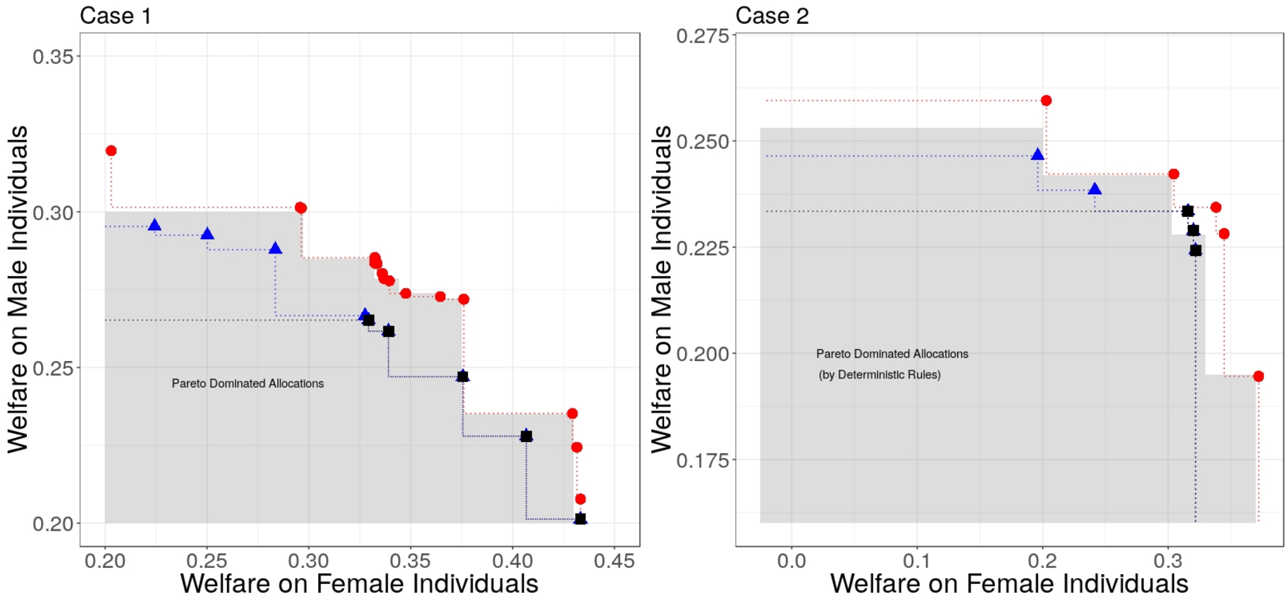


Figure C.1: Empirical application. (Discretized) Pareto frontier under *probabilistic* linear policy rule estimated through MIQP. Dots denote Pareto optimal allocations. Red dots (circle) correspond to $\Pi_1$, blue dots (triangle) to $\Pi_2$ and black dots (square) to $\Pi_3$. The gray area denotes the set of allocations dominated by a *deterministic* decision rule.

score, while the four methods can still be feasibly implemented for $n = 600$. Figure C.3 presents results for the different function classes for $p = 3$ (instead of $p = 4$) covariates. Figure C.4 reports welfare comparisons for the disparate impact method for female and male participants. Table 2 reports comparisons for $n = 400$. In the table, we observe that a smaller sample size tends to decrease the performance of each method, as expected. We still observe that the proposed method leads to the largest fairness across all designs. While for $n = 600$, the proposed method is never Pareto dominated, here also the method is never Pareto dominated with a single exception occurring for the maximum score, where we observe a slight dominance for welfare but not fairness which might occur with small sample sizes. In all remaining cases, the proposed method leads to strictly larger welfare for the female students with respect to all competitors.

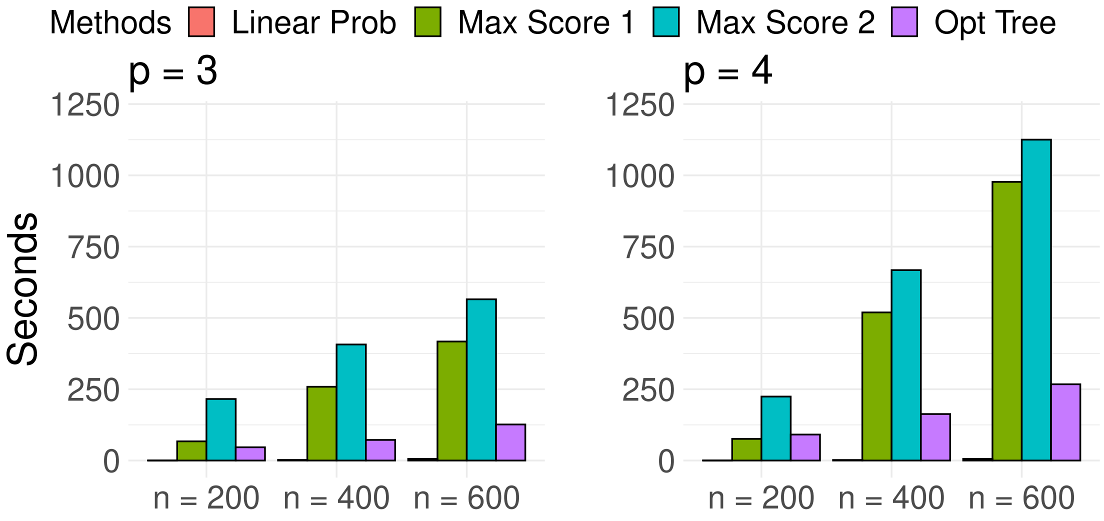


Figure C.2: Computational time in seconds for different number of covariates $p$ and sample sizes $n$. Here, Linear Prob is a linear probability rule estimated via linear programming, maximum score is estimated with mixed-integer linear program and optimal tree via exhaustive search. The maximum score algorithm presents two different stopping times (Max Score 1 and Max Score 2).

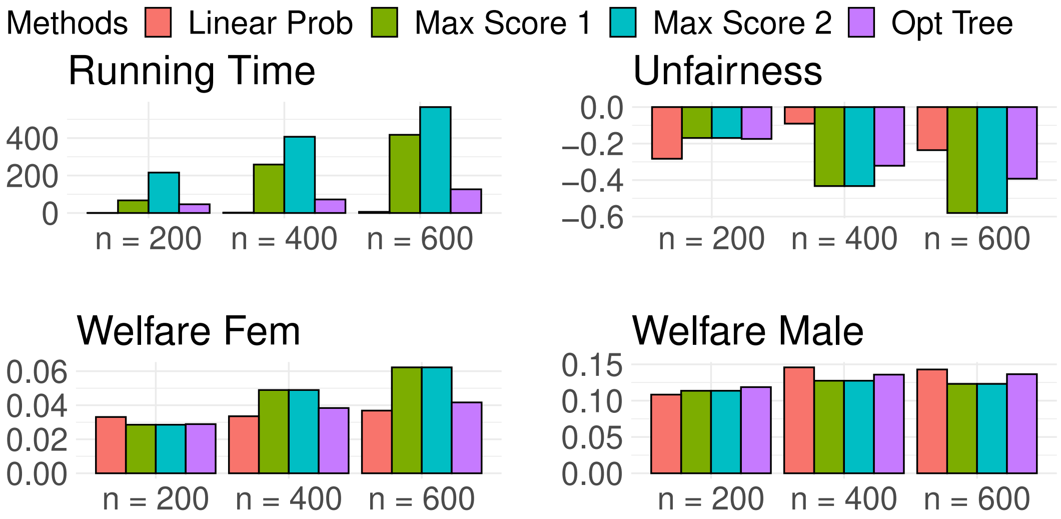


Figure C.3: $p = 3$. Running time, unfairness and welfare as a function of the sample size and covariates. Here, Linear Prob is a linear probability rule estimated via linear programming, maximum score is estimated with mixed-integer linear program and optimal tree via exhaustive search. The maximum score algorithm presents two different stopping times (Max Score 1 and Max Score 2).

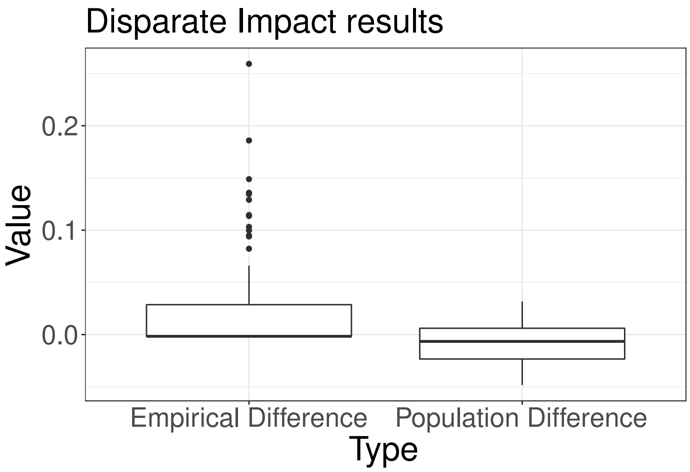


Figure C.4: Difference between females and males' welfare for the method Disparate impact in Table 2, with $\kappa = 1$. The left-hand side panel shows the empirical difference (showing that the constraint is attained) and the right-hand side panel the population counterpart.

Table C.1: Statistical disparity (Unfairness), welfare of male ($W_0$) and female ($W_1$) participants of the proposed method (Fair Targeting) and of the alternative procedures in *percentage points*. Weighted average maximizes a weighted average of females and males' welfare with weight $\alpha = 1/2$; Utilitarian average uses instead $\alpha = \mathbb{E}_n[S]$; Constrained Max maximizes welfare under fairness constrain and Disparate Impact maximizes welfare under constraints on disparate welfare impact between the two groups. $n = 400, p = 4$. The constraint is $\kappa = 10$ for the methods in the fourth and fifth row and $\kappa = 1$ for the last two rows.

| | Linear Rule | | | Maximum Score | | | Tree | | |
|---|---|---|---|---|---|---|---|---|---|
| | UnFair | $W_0$ | $W_1$ | UnFair | $W_0$ | $W_1$ | UnFair | $W_0$ | $W_1$ |
| Fair Targeting | −19.7 | 12.2 | 6.1 | −46.4 | 12.7 | 5.3 | −32.7 | 12.7 | 6.0 |
| Weighted Average | 20.7 | 15 | 4.4 | 7.5 | 15.3 | 4.3 | −1.9 | 14.5 | 5.5 |
| Utilitarian Average | 7.5 | 15.3 | 4.3 | 7.5 | 15.3 | 4.3 | 7.5 | 15.3 | 4.3 |
| Constrained Max | −8.5 | 14.6 | 4.6 | −18.7 | 14 | 5 | −16.8 | 14.5 | 5 |
| Disparate Impact | 6.3 | 12.8 | 4.6 | −0.2 | 14.4 | 5.5 | 2.9 | 14.1 | 5 |
| Constrained Max2 | −12.3 | 14.2 | 4.7 | −21.7 | 13.7 | 4.9 | −19.3 | 14.5 | 5 |
| Disparate Impact2 | −1.6 | 11.2 | 4.8 | −6.2 | 14.2 | 5.4 | −2.2 | 13.9 | 5.2 |

# References

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

Boucheron, S., O. Bousquet, and G. Lugosi (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics 9*, 323–375.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Boucheron, S., G. Lugosi, P. Massart, et al. (2003). Concentration inequalities using the entropy method. *The Annals of Probability 31*(3), 1583–1614.

Charnes, A. and W. W. Cooper (1962). Programming with linear fractional functionals. *Naval Research logistics quarterly 9*(3-4), 181–186.

Devroye, L., L. Györfi, and G. Lugosi (2013). *A probabilistic theory of pattern recognition*, Volume 31. Springer Science & Business Media.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Volume 66. CRC Press.

Feldman, A. and A. Kirman (1974). Fairness and envy. *The American Economic Review*, 995–1005.

Kasy, M. and R. Abebe (2021). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 576–586.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Mas-Colell, A., M. D. Whinston, J. R. Green, et al. (1995). *Microeconomic theory*, Volume 1. Oxford university press New York.

Papadimitriou, C. H. and K. Steiglitz (1998). *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Van Der Vaart, A. W. and J. A. Wellner (1996). Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Zhou, Z., S. Athey, and S. Wager (2018). Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*.