Fair Learning with Private Demographic Data

Hussein Mozannar [1] **Mesrob I. Ohannessian** [2] **Nathan Srebro** [3]

Abstract

Sensitive attributes such as race are rarely available to learners in real world settings as their collection is often restricted by laws and regulations. We give a scheme that allows individuals to release their sensitive information privately while still allowing any downstream entity to learn non-discriminatory predictors. We show how to adapt non-discriminatory learners to work with privatized protected attributes giving theoretical guarantees on performance. Finally, we highlight how the methodology could apply to learning fair predictors in settings where protected attributes are only available for a subset of the data.

[1] IDSS, Massachusetts Institute of Technology, MA, USA [2] Department of Electrical and Computer Engineering, University of Illinois at Chicago, IL, USA [3] Toyota Technological Institute, IL, USA. Correspondence to: Hussein Mozannar <mozannar@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

A. Deferred Proofs

Two important notation we use throughout are: for empirical versions of quantities based on data set S we use a superscript S and the "probabilistic" inequality $a \leq_\delta b$ signifies that a is less than b with probability greater than $1-\delta$.

A.1. Section 4

The below example illustrates that non-discrimination with respect to A and Z are not equivalent for general predictors.

Example 1. *Let $|\mathcal{A}| = 2$, consider the predictors $\hat{Y}_1 = h(X, Z)$ and $\hat{Y}_2 = h(X, Z)$ with the conditional probabilities for $y \in \{0, 1\}$ defined in table 1 with the function $h(x) = \begin{cases} 0 \textit{ if } x \leq 1/2 \\ \frac{1}{2x} \textit{ if } x > 1/2 \end{cases}$, note that $h(x) \in [0, 1]$ so that the predictor $\hat{Y}_2$ is valid.*

(a,z)	$\mathbb{P}(\hat{Y}_1 = 1 \mid A = a, Z = z, Y = y)$	$\mathbb{P}(\hat{Y}_2 = 1 \mid A = a, Z = z, Y = y)$
(0,0)	$\frac{1}{2\pi}$	$h(\mathbb{P}(A = 0 \mid Z = 0, Y = y))$
(0,1)	0	$h(\mathbb{P}(A = 0 \mid Z = 1, Y = y))$
(1,0)	0	$h(\mathbb{P}(A = 1 \mid Z = 0, Y = y))$
(1,1)	$\frac{1}{2\pi}$	$h(\mathbb{P}(A = 1 \mid Z = 1, Y = y))$

Table 1. Predictors used to show non-equivalence of discrimination with respect to A and Z when predictors are a function of Z.

The predictors $\hat{Y}_1$ and $\hat{Y}_2$ are designed by construction to show that non discrimination with respect to A and Z are not statistically equivalent. We show that $\hat{Y}_1$ satisfies EO with respect to A but violates it with respect to Z and $\hat{Y}_2$ is non-discriminatory with respect to Z but is for A.

Proof. For $\hat{Y}_1$: first we show it satisfies EO for A:

$$\begin{aligned} &\mathbb{P}(\hat{Y}_1 = 1 \mid A = a, Y = y) \\ &= \pi \mathbb{P}(\hat{Y}_1 = 1 \mid A = a, Z = a, Y = y) \\ &+ (1-\pi)\mathbb{P}(\hat{Y}_1 = 1 \mid A = a, Z = \bar{a}, Y = y) = \frac{1}{2} \end{aligned}$$

Since the above is no different for $a, y \in \{0, 1\}$, $\hat{Y}_1$ satisfies EO. Now with respect to Z:

$$\begin{aligned} &\mathbb{P}(\hat{Y}_1 = 1 \mid Z = a, Y = y) \\ &= \mathbb{P}(A = a \mid Z = a, Y = y)\mathbb{P}(\hat{Y}_1 = 1 \mid Z = a, A = a, Y = y) \\ &+ \mathbb{P}(A = \bar{a} \mid Z = a, Y = y)\mathbb{P}(\hat{Y}_1 = 1 \mid Z = a, A = \bar{a}, Y = y) \\ &= \frac{\mathbb{P}(A = a \mid Z = a, Y = y)}{2\pi} \end{aligned}$$

Therefore if and only if $\mathbb{P}(A = 0 \mid Z = 0, Y = y) = \mathbb{P}(A = 1 \mid Z = 1, Y = y)$ is it also non discriminatory with respect to Z.

For $\hat{Y}_2$: by construction only one of $(\mathbb{P}(\hat{Y}_2 = 1 \mid A = a, Z = a, Y = y), \mathbb{P}(\hat{Y}_2 = 1 \mid A = \bar{a}, Z = a, Y = y))$ is non-zero as only one of $(\mathbb{P}(A = 1 \mid Z = a, Y = y), \mathbb{P}(A = 0, Z = a, Y = y))$ is greater than $1/2$ and so:

$$\begin{aligned} &\mathbb{P}(\hat{Y}_2 = 1 \mid Z = a, Y = y) \\ &= \mathbb{P}(A = a \mid Z = a, Y = y)\mathbb{P}(\hat{Y}_2 = 1 \mid Z = a, A = a, Y = y) \\ &+ \mathbb{P}(A = \bar{a} \mid Z = a, Y = y)\mathbb{P}(\hat{Y}_2 = 1 \mid Z = a, A = \bar{a}, Y = y) \\ &= \frac{1}{2} \end{aligned}$$

Therefore $\hat{Y}_2$ satisfies EO with respect to Z, on the other side:

$$\begin{aligned}
&\mathbb{P}(\hat{Y}_2 = 1|A = a, Y = y)\\
&= \mathbb{P}(Z = a|A = a, Y = y)\mathbb{P}(\hat{Y}_2 = 1|Z = a, A = a, Y = y)\\
&+ \mathbb{P}(Z = \bar{a}|A = a, Y = y)\mathbb{P}(\hat{Y}_2 = 1|Z = \bar{a}, A = a, Y = y)\\
&= \pi \cdot h(\mathbb{P}(A = a|Z = a, Y = y))\\
&+ (1 - \pi) \cdot h(\mathbb{P}(A = a|Z = \bar{a}, Y = y))
\end{aligned}$$

and is discriminatory with respect to A unless $\mathbb{P}(A = a, Y = y) = \mathbb{P}(A = \bar{a}, Y = y)$ for $y \in \{0, 1\}$ as $\mathbb{P}(A = a|Z = a, Y = y) = \frac{\pi\mathbb{P}(A=a,Y=y)}{\mathbb{P}(Z=a,Y=y)}$. □

***Proposition** 1 Consider any exact non-discrimination notion among equalized odds, demographic parity, accuracy parity, or equality of false discovery/omission rates. Let $\hat{Y} := h(X)$ be a binary predictor, then $\hat{Y}$ is non-discriminatory with respect to A if and only if it is non-discriminatory with respect to Z.*

Proof. The proof of the above proposition relies on the fact that if $\hat{Y}$ is independent of Z given A, then the conditional probabilities with respect to Z and A are related via a linear system.

We prove the proposition by considering a general formulation of the constraints we previously mentioned, let $\mathcal{E}_1, \mathcal{E}_2$ be two probability events defined with respect to $(X, Y, \hat{Y})$, then consider the following probability:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, Z = a\right)\\
&= \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, Z = a, A = a'\right)) \mathbb{P}(A = a'|\mathcal{E}_2, Z = z)\\
&\overset{(a)}{=} \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, A = a'\right)) \mathbb{P}(A = a'|\mathcal{E}_2, Z = z)\\
&= \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, A = a'\right)) \frac{\mathbb{P}(Z = z|A = a', \mathcal{E}_2)\mathbb{P}(A = a', \mathcal{E}_2)}{\mathbb{P}(Z = a', \mathcal{E}_2)}\\
&= \mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, A = a\right)) \frac{\pi\mathbb{P}(A = a, \mathcal{E}_2)}{\mathbb{P}(Z = a, \mathcal{E}_2)}\\
&+ \sum_{a' \in \mathcal{A}\setminus\{a\}} \mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, A = a'\right)) \frac{\bar{\pi}\mathbb{P}(A = a', \mathcal{E}_2)}{\mathbb{P}(Z = a', \mathcal{E}_2)} \qquad (1)
\end{aligned}$$

step (a) follows as $(X, Y, \hat{Y})$ are independent of Z given A. We define non-discrimination with respect to A as having (similarly defined with respect to Z):

$$\mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, A = a\right) = \mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, A = a'\right) \quad \forall a, a' \in \mathcal{A}$$

Assume first that the predictor $\hat{Y}$ is non-discriminatory with respect to A, hence $\exists c$ where $\forall a \in \mathcal{A}$ we have $\mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = a) = c$, hence by (16) for all $a \in \mathcal{A}$:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, Z = a\right)\\
&= c\frac{\pi\mathbb{P}(A = a, \mathcal{E}_2)}{\mathbb{P}(Z = a, \mathcal{E}_2)} + \sum_{a' \in \mathcal{A}\setminus\{a\}} c\frac{\bar{\pi}\mathbb{P}(A = a', \mathcal{E}_2)}{\mathbb{P}(Z = a', \mathcal{E}_2)} = c
\end{aligned}$$

which proves that $\hat{Y}$ is also non-discriminatory with respect to A.

Now, assume instead that the predictor $\hat{Y}$ is non-discriminatory with respect to Z, hence $\exists c$ where $\forall a \in \mathcal{A}$ we have $\mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z = a) = c$. Let P be the following $|\mathcal{A}| \times |\mathcal{A}|$ matrix:

$$\begin{cases} P_{i,i} = \frac{\pi\mathbb{P}(A=i,\mathcal{E}_2)}{\mathbb{P}(Z=i,\mathcal{E}_2)} \text{ for } i \in \mathcal{A} \\ P_{i,j} = \frac{\bar{\pi}\mathbb{P}(A=i,\mathcal{E}_2)}{\mathbb{P}(Z=j,\mathcal{E}_2)} \text{ for } i, j \in \mathcal{A} \text{ s.t.} i \neq j \end{cases}$$

Then we have the following linear system of equations:

$$\begin{bmatrix} \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z=0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z=|\mathcal{A}|-1) \end{bmatrix} = P \begin{bmatrix} \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A=|\mathcal{A}|-1) \end{bmatrix}$$

denoted by $\mathbf{z} = P\mathbf{a}$

In our case $\mathbf{a} = c \cdot \mathbf{1}$, and we show that also $\mathbf{z} = c \cdot \mathbf{1}$. Let us state some properties of the matrix P:

- P is row-stochastic
- P is invertible (we later show the exact form of this inverse implying its existence, however its existence is easy to see as all rows are linearly independent as $\pi \neq \bar{\pi}$ and $\forall a$, $\mathbb{P}(Z=a, \mathcal{E}_2) > 0$).
- As P is row-stochastic and invertible, the rows of P^{-1} sum to 1, this is as $P\mathbf{1} = \mathbf{1} \iff \mathbf{1} = P^{-1}\mathbf{1}$

By the second property $z = c \cdot P^{-1}\mathbf{1}$ and by the third property we have $P^{-1}\mathbf{1} = \mathbf{1}$ which in turn means that $z = c \cdot \mathbf{1}$ and implies that $\hat{Y}$ is non-discriminatory with respect to Z.

As an extension, consider fairness notions formulated as:

$$\mathbb{P}\left(\mathcal{E}_1, A=a|\mathcal{E}_2\right) = \mathbb{P}\left(\mathcal{E}_1, A=a'|\mathcal{E}_2\right) \quad \forall a, a' \in \mathcal{A}$$

Then we have

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_1, Z=a|\mathcal{E}_2\right) \\
&= \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\mathcal{E}_1, Z=a|\mathcal{E}_2, A=a'\right) \mathbb{P}(A=a'|\mathcal{E}_2) \\
&= \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\mathcal{E}_1|\mathcal{E}_2, A=a'\right) \mathbb{P}\left(Z=a|A=a'\right)) \mathbb{P}(A=a'|\mathcal{E}_2) \\
&= \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\mathcal{E}_1, A=a'|\mathcal{E}_2\right) \mathbb{P}\left(Z=a|A=a'\right)) \\
&= \pi \mathbb{P}\left(\mathcal{E}_1, A=a'|\mathcal{E}_2\right) \sum_{a' \in \mathcal{A} \setminus \{a\}} \bar{\pi} \mathbb{P}\left(\mathcal{E}_1, A=a'|\mathcal{E}_2\right)
\end{aligned}$$

By the same arguments as above, for these notions of fairness $\hat{Y}$ is non-discriminatory with respect to A if and only if it is non-discriminatory with respect to Z.

For concreteness, we derive equation (16) for each of the fairness notions we mentioned. First a detailed derivation for equalized odds, we let $\mathcal{E}_1 = \{\hat{Y} = 1\}$ and for EO we need to apply the above reasoning for $|\mathcal{Y}|$ events $\mathcal{E}_{2_y} = \{Y = y\}$:

$$
\begin{aligned}
&\mathbb{P}(\hat{Y}=1|Y=y,Z=a)\\
&=\sum_{a'}\mathbb{P}(\hat{Y}=1|Y=y,Z=a,A=a')\mathbb{P}(A=a'|Z=a,Y=y)\\
&\overset{(a)}{=}\sum_{a'}\mathbb{P}(\hat{Y}=1|Y=y,A=a')\mathbb{P}(A=a'|Z=a,Y=y)\\
&=\sum_{a'}\mathbb{P}(\hat{Y}=1|Y=y,A=a')\frac{\mathbb{P}(Z=a,Y=y|A=a')\mathbb{P}(A=a')}{\mathbb{P}(Z=a,Y=y)}\\
&\overset{(b)}{=}\sum_{a'}\mathbb{P}(\hat{Y}=1|Y=y,A=a')\frac{\mathbb{P}(Z=a|A=a')\mathbb{P}(Y=y|A=a')\mathbb{P}(A=a')}{\mathbb{P}(Z=a,Y=y)}\\
&=\mathbb{P}(\hat{Y}=1|Y=y,A=a)\frac{\pi\mathbf{P}_{ya}}{\pi\mathbf{P}_{ya}+\sum_{a''\backslash a}\bar{\pi}\mathbf{P}_{ya''}}\\
&+\sum_{a'\backslash a}\mathbb{P}(\hat{Y}=1|Y=y,A=a')\frac{\bar{\pi}\mathbf{P}_{ya'}}{\pi\mathbf{P}_{ya}+\sum_{a''\backslash a}\bar{\pi}\mathbf{P}_{ya''}}
\end{aligned}
$$

First line by conditioning on A and then taking expectation, (a) is by our assumption of the conditional independence of $Z, \hat{Y}$ given A and step (b) by the independence of Z and Y given A.

Similarly for demographic parity with denoting $p_a = \mathbb{P}(A=a)$:

$$
\begin{aligned}
&\mathbb{P}(\hat{Y}=1|Z=a)\\
&=\sum_{a'}\mathbb{P}(\hat{Y}=1|Z=a,A=a')\mathbb{P}(A=a'|Z=a)\\
&=\sum_{a'}\mathbb{P}(\hat{Y}=1|A=a')\mathbb{P}(A=a'|Z=a)\\
&=\sum_{a'}\mathbb{P}(\hat{Y}=1|A=a')\frac{\mathbb{P}(Z=a|A=a')p_{a'}}{\sum_{a''}\mathbb{P}(Z=a|A=a'')p_{a''}}\\
&=\sum_{a'}\mathbb{P}(\hat{Y}=1|A=a')\frac{\mathbb{P}(Z=a|A=a')p_{a'}}{\sum_{a''}\mathbb{P}(Z=a|A=a'')p_{a''}}\\
&=\mathbb{P}(\hat{Y}=1|A=a)\frac{\pi p_a}{\pi p_a+\sum_{a''\backslash a}\bar{\pi}p_{a''}}+\sum_{a'\backslash a}\mathbb{P}(\hat{Y}=1|A=a')\frac{\bar{\pi}p_{a'}}{\pi p_a+\sum_{a''\backslash a}\bar{\pi}p_{a''}}
\end{aligned}
$$

Now for equal accuracy among groups:

$$
\begin{aligned}
&\mathbb{P}(\hat{Y}\neq Y|Z=a)\\
&=\sum_{a'}\mathbb{P}(\hat{Y}\neq Y|Z=a,A=a')\mathbb{P}(A=a'|Z=a)\\
&=\mathbb{P}(\hat{Y}\neq Y|A=a')\frac{\pi p_a}{\pi p_a+\sum_{a''\backslash a}\bar{\pi}p_{a''}}+\sum_{a'\backslash a}\mathbb{P}(\hat{Y}\neq Y|A=a')\frac{\bar{\pi}p_{a'}}{\pi p_a+\sum_{a''\backslash a}\bar{\pi}p_{a''}}
\end{aligned}
$$

And finally for equality of false discovery/omission rates, denote $p_{\hat{y},a} := \mathbb{P}(\hat{Y}=\hat{y}, A=a)$:

$$
\begin{aligned}
&\mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, Z = a) \\
&= \sum_{a'} \mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, Z = a, A = a') \mathbb{P}(A = a' | Z = a, \hat{Y} = \hat{y}) \\
&= \sum_{a'} \mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, A = a') \mathbb{P}(A = a' | Z = a, \hat{Y} = \hat{y}) \\
&= \sum_{a'} \mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, A = a') \frac{\mathbb{P}(Z = a, \hat{Y} = \hat{y} | A = a') p_{a'}}{\sum_{a''} \mathbb{P}(Z = a, \hat{Y} = \hat{y} | A = a'') p_{a''}} \\
&= \mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, A = a') \frac{\pi p_{\hat{y},a}}{\pi p_{\hat{y},a} + \sum_{a'' \setminus a} \bar{\pi} p_{\hat{y},a''}} + \sum_{a' \setminus a} \mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, A = a') \frac{\bar{\pi} p_{\hat{y},a'}}{\pi p_{\hat{y},a} + \sum_{a'' \setminus a} \bar{\pi} p_{\hat{y},a''}}
\end{aligned}
$$

Note that we did not need the independence of $\hat{Y}$ and Z given A to express $\mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, Z = a)$ in terms of $\mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, A = a)$ so that the equivalence follows without our assumption for equality of FDR. However, to be able to do the inversion of statistics we require the assumption. □

The version of the below Lemma that appears in the text is obtained by plugging in $\pi = \frac{e^\epsilon}{|\mathcal{A}|-1+e^\epsilon}$.

Lemma 1 *For any $\delta \in (0, 1/2)$, any binary predictor $\hat{Y} := h(X)$, denote by $\mathbf{P}_{ya} := \mathbb{P}(Y = y, A = a)$, $\Gamma_{ya} := \left| q_{y,a}(\hat{Y}) - \gamma_{y,0}(\hat{Y}) \right|$ and $\widetilde{\Gamma}^S_{ya}$ our proposed estimator based on S, let $C = \frac{\pi + |\mathcal{A}| - 1}{|\mathcal{A}|\pi - 1}$, then if $n \geq \frac{8 \log(|8\mathcal{A}|/\delta)}{\min_{ya} \mathbf{P}_{ya}}$, we have:*

$$
\mathbb{P}\left(\max_{ya} |\widetilde{\Gamma}^S_{ya} - \Gamma_{ya}| > \sqrt{\frac{\log(16/\delta)}{2n} \frac{4|\mathcal{A}|C^2}{\min_{ya} \mathbf{P}^2_{ya}}} \right) \leq \delta
$$

Proof. **Step 1:** Deriving our estimator

The following equality allows to invert the statistics of the population with respect to Z that we have sample estimates of to get population estimates of the true statistics with respect to A. We write

$$
\begin{aligned}
&\mathbb{P}(\hat{Y} = 1 | Y = y, Z = a) = \\
&\sum_{a'} \mathbb{P}(\hat{Y} = 1 | Y = y, Z = a, A = a') \mathbb{P}(A = a' | Z = a, Y = y) \\
&\overset{(a)}{=} \sum_{a'} \mathbb{P}(\hat{Y} = 1 | Y = y, A = a') \mathbb{P}(A = a' | Z = a, Y = y) \\
&= \sum_{a'} \mathbb{P}(\hat{Y} = 1 | Y = y, A = a') \frac{\mathbb{P}(Z = a, Y = y | A = a') \mathbb{P}(A = a')}{\mathbb{P}(Z = a, Y = y)} \\
&\overset{(b)}{=} \sum_{a'} \mathbb{P}(\hat{Y} = 1 | Y = y, A = a') \frac{\mathbb{P}(Z = a | A = a') \mathbb{P}(Y = y | A = a') \mathbb{P}(A = a')}{\mathbb{P}(Z = a, Y = y)} \\
&= \pi \mathbb{P}(\hat{Y} = 1 | Y = y, A = a) \frac{\mathbb{P}(Y = y, A = a)}{\mathbb{P}(Z = a, Y = y)} + \sum_{a' \setminus a} \bar{\pi} \mathbb{P}(\hat{Y} = 1 | Y = y, A = a') \frac{\mathbb{P}(Y = y, A = a')}{\mathbb{P}(Z = a, Y = y)} \qquad (2)
\end{aligned}
$$

First line is by conditioning on A and then taking expectation, step (a) is by our assumption of the conditional independence of $Z, \hat{Y}$ given A and step (b) by the independence of Z and Y given A.

Let G be the $\mathcal{A} \times \mathcal{A}$ matrix be as such: $\begin{cases} G_{i,i} = \pi \frac{\mathbb{P}(Y=y,A=i)}{\mathbb{P}(Z=i,Y=y)} \text{ for } i \in \mathcal{A} \\ G_{i,j} = \bar{\pi} \frac{\mathbb{P}(Y=y,A=j)}{\mathbb{P}(Z=i,Y=y)} \text{ for } i, j \in \mathcal{A} \text{ s.t.} i \neq j \end{cases}$. Then we can write equation (17) as a linear system with $q_{ya}(\hat{Y}) = \mathbb{P}(\hat{Y} = 1|Y = y, Z = a)$:

$$\begin{bmatrix} q_{y0} \\ \vdots \\ q_{y,|\mathcal{A}-1|} \end{bmatrix} = G \begin{bmatrix} \mathbb{P}(\hat{Y} = 1|Y = y, A = 0) \\ \vdots \\ \mathbb{P}(\hat{Y} = 1|Y = y, A = |\mathcal{A}| - 1) \end{bmatrix}$$

$$q_{y,.} = G\,\mathbb{P}\left(\hat{Y} = 1|Y = y, A\right) \textit{ (notation)}$$

And thus by inverting G we can recover the population statistics. We show that the inverse of G takes the following form:

$$\begin{cases} G^{-1}_{i,i} = \frac{\pi+|\mathcal{A}|-2}{|\mathcal{A}|\pi-1} \frac{\mathbb{P}(Z=i,Y=y)}{\mathbb{P}(Y=y,A=i)} \text{ for } i \in \mathcal{A} \\ G^{-1}_{i,j} = \frac{\pi-1}{|\mathcal{A}|\pi-1} \frac{\mathbb{P}(Z=j,Y=y)}{\mathbb{P}(Y=y,A=i)} \text{ for } i,j \in \mathcal{A} \text{ s.t.} i \neq j \end{cases}$$

Let $i \neq j \in \mathcal{A}$:

$$\begin{aligned} &G_i G^{-1}_{,j} \\ &= \sum_k G_{i,k} G^{-1}_{k,j} \\ &= \pi \frac{\mathbb{P}(Y = y, A = i)}{\mathbb{P}(Z = i, Y = y)} \cdot \frac{\pi - 1}{|\mathcal{A}|\pi - 1} \frac{\mathbb{P}(Z = j, Y = y)}{\mathbb{P}(Y = y, A = i)} + \bar{\pi} \frac{\mathbb{P}(Y = y, A = j)}{\mathbb{P}(Z = i, Y = y)} \frac{\pi + |\mathcal{A}| - 2}{|\mathcal{A}|\pi - 1} \frac{\mathbb{P}(Z = j, Y = y)}{\mathbb{P}(Y = y, A = j)} \\ &+ \sum_{k \setminus \{i,j\}} \bar{\pi} \frac{\mathbb{P}(Y = y, A = k)}{\mathbb{P}(Z = i, Y = y)} \cdot \frac{\pi - 1}{|\mathcal{A}|\pi - 1} \frac{\mathbb{P}(Z = j, Y = y)}{\mathbb{P}(Y = y, A = k)} \\ &= \frac{\mathbb{P}(Z = j, Y = y)}{\mathbb{P}(Z = i, Y = y)} \cdot \frac{\pi(\pi - 1) + \frac{1-\pi}{|\mathcal{A}|-1}(\pi + |\mathcal{A}| - 2 + (|\mathcal{A}| - 2)(\pi - 1))}{|\mathcal{A}|\pi - 1} \\ &= 0 \end{aligned}$$

And now for $i \in \mathcal{A}$

$$\begin{aligned} G_i G^{-1}_{,i} &= \sum_k G_{i,k} G^{-1}_{k,i} \\ &= \pi \frac{\mathbb{P}(Y = y, A = i)}{\mathbb{P}(Z = i, Y = y)} \cdot \frac{\pi + |\mathcal{A}| - 2}{|\mathcal{A}|\pi - 1} \frac{\mathbb{P}(Z = i, Y = y)}{\mathbb{P}(Y = y, A = i)} + \sum_{k \setminus \{i\}} \bar{\pi} \frac{\mathbb{P}(Y = y, A = k)}{\mathbb{P}(Z = i, Y = y)} \cdot \frac{\pi - 1}{|\mathcal{A}|\pi - 1} \frac{\mathbb{P}(Z = i, Y = y)}{\mathbb{P}(Y = y, A = k)} \\ &= \frac{\pi(\pi + |\mathcal{A}| - 2) + \frac{1-\pi}{|\mathcal{A}|-1}(\pi - 1)(|\mathcal{A}| - 1)}{|\mathcal{A}|\pi - 1} \\ &= 1 \end{aligned}$$

Which proves that it is indeed the inverse.

The matrix G involves estimating the probabilities $\mathbb{P}(Y = y, A = a)$ which we do not have access to but can similarly recover by noting that:

$$\begin{aligned} \mathbf{Q}_{yz} &= \sum_{a \in \mathcal{A}} \mathbb{P}\left(Y = y, Z = z|A = a\right) \mathbb{P}\left(A = a\right) \\ &= \sum_{a \in \mathcal{A}} \mathbb{P}\left(Y = y|A = a\right) \mathbb{P}\left(Z = z|A = a\right) \mathbb{P}(A = a) \\ &= \pi \mathbb{P}\left(Y = y, A = z\right) + \sum_{a \neq z} \bar{\pi} \mathbb{P}\left(Y = y, A = a\right) \end{aligned} \tag{3}$$

Let the matrix $\Pi \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ be as follows $\Pi_{i,j} = \pi$ if $i = j$ and $\Pi_{i,j} = \bar{\pi}$ if $i \neq j$. We know from equation (18) that:

$$\begin{bmatrix} \mathbf{Q}_{y0} \\ \vdots \\ \mathbf{Q}_{y,|\mathcal{A}|-1} \end{bmatrix} = \Pi \begin{bmatrix} \mathbb{P}(Y = y, A = 0) \\ \vdots \\ \mathbb{P}(Y = y, A = |\mathcal{A}| - 1) \end{bmatrix}$$

$$\mathbf{Q}_{y,.} = \Pi \, \mathbb{P}\left(Y = y, A\right) \textit{ (notation)}$$

Therefore $\Pi_k^{-1}\mathbf{Q}_{y,.} = \mathbb{P}(Y = y, A = k)$ where Π_k^{-1} is the k'th row of Π^{-1}. Now Π^{-1} is as such: $\Pi^{-1}_{i,i} = \frac{\pi+|\mathcal{A}|-2}{|\mathcal{A}|\pi-1}$ and $\Pi^{-1}_{i,j} = \frac{\pi-1}{|\mathcal{A}|\pi-1}$ if $i \neq j$ with the same proof as for the inverse of G. Therefore our empirical estimator for $\mathbb{P}(\hat{Y} = 1|Y = y, A = a)$ is $\hat{G}_a^{-1}q^S_{y,.}$ where $\hat{G}^{-1}$ is defined with the empirical versions of the probabilities involved where $\mathbb{P}(Y = y, A = a)$ is estimated by $\Pi_a^{-1}\mathbf{Q}^S_{y,.}$. One issue that arises here is that while the sum of our estimator entries sum to 1, some entries might be in fact negative and therefore we need to project the derived estimator onto the simplex. We later discuss the implications of this step.

Step 2: Concentration of raw estimator

Let us first denote some things: $n^S_{y,z} = \sum_i \mathbf{1}(y_i = y, z_i = z)$, $\mathbf{Q}_{y,z} = \mathbb{P}(Y = y, Z = z)$, and the random variables $S_{y,z} = \{i : y_i = y, z_i = z\}$.
We have that $\mathbb{E}[\hat{G}_z^{-1}q^S_{y,.}|S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}] = G_z^{-1}q_{y,.} = \gamma_{y,z}$. Inspired by the proof of Lemma 2 in (Woodworth et al., 2017) we have:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{G}_z^{-1}q^S_{y,.} - \gamma_{yz}| > t\right) \\
&\overset{(a)}{=} \sum_{S_{y,0},\cdots,S_{y,|\mathcal{A}|-1}} \mathbb{P}\left(|\hat{G}_z^{-1}q^S_{y,.} - \gamma_{yz}| > t|S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}\right)\mathbb{P}\left(S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}\right) \\
&\overset{(b)}{\leq} \mathbb{P}\left(\cup_{a\in\mathcal{A}}\{n^S_{y,a} < \frac{n\mathbf{Q}_{y,a}}{2}\}\right) \\
&+ \sum_{\forall z, S_{yz}: n^S_{yz} \geq \frac{n\mathbf{Q}_{yz}}{2}} \mathbb{P}\left(|\hat{G}_z^{-1}q^S_{y,.} - \gamma_{yz}| > t|S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}\right)\mathbb{P}\left(S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}\right) \\
&\overset{(c)}{\leq} |A| \exp\left(-\frac{\min_a n\mathbf{Q}_{ya}}{8}\right) \\
&+ \sum_{\forall z, S_{yz}: n^S_{yz} \geq \frac{n\mathbf{Q}_{yz}}{2}} \mathbb{P}\left(|\hat{G}_z^{-1}q^S_{y,.} - \gamma_{yz}| > t|S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}\right)\mathbb{P}\left(S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}\right)
\end{aligned}$$

Step (a) follows by conditioning over over all $|\mathcal{A}|^n$ possible configurations of $S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1} \subset [n]$, step (b) comes by splitting over configurations where $\forall z, S_{yz} : n^S_{yz} \geq \frac{n\mathbf{Q}_{yz}}{2}$ and the complement of the previous event and upper bounding this complement by the probability that there $\exists z$ s.t. $n^S_{yz} < \frac{n\mathbf{Q}_{yz}}{2}$. Finally step (c) comes from a union bound and then a Chernoff bound on $n^S_{yz} \sim \text{Binomial}(n, \mathbf{Q}_{yz})$ and taking the minimum over $\mathbf{Q}_{ya}$.

We now recall McDiarmid's inequality (McDiarmid, 1989). Let $W^n = (W_1, \cdots, W_n) \in \mathcal{W}^n$ be n independent random variables and $f : \mathcal{W}^n \to \mathbb{R}$, if there exists constants $c_1, \cdots, c_n$ such that for all $i \in [n]$:

$$\sup_{w_1,\cdots,w_i,w_i',\cdots,w_n} |f(w_1, \cdots, w_i, \cdots, w_n) - f(w_1, \cdots, w_i', \cdots, w_n)| \leq c_i$$

Then for all $\epsilon > 0$:

$$\mathbb{P}\left(f(W_1, \cdots, W_n) - \mathbb{E}[f(W_1, \cdots, W_n]\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

Now conditioned on $S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1}$, our estimator $\hat{G}_z^{-1}q^S_{y,.}$ is only a function of $\hat{Y}_1, \cdots, \hat{Y}_n$, we try to bound how much can our estimator change on two dataset S and S' differing by only one value of $\hat{Y}_i$:

For convenience denote by $C_1 = \frac{\pi+|\mathcal{A}|-2}{|\mathcal{A}|\pi-1}$ and $C_2 = \frac{\pi-1}{|\mathcal{A}|\pi-1}$:

$$\begin{aligned}
&\sup_{S,S'} |\hat{G}_z^{-1} q^S_{y,.} - \hat{G}_z^{-1} q^{S'}_{y,.}| \\
&= \left| C_1 \frac{\Pi_z^{-1}\mathbf{Q}^S_{y,.}}{\mathbf{Q}^S_{y,z}} q^S_{y,z} + \sum_{a\in\mathcal{A}\setminus z} C_2 \frac{\Pi_a^{-1}\mathbf{Q}^S_{y,.}}{\mathbf{Q}^S_{y,z}} q^S_{y,a} - C_1 \frac{\Pi_z^{-1}\mathbf{Q}^S_{y,.}}{\mathbf{Q}^S_{y,z}} q^{S'}_{y,z} - \sum_{a\setminus z} C_2 \frac{\Pi_a^{-1}\mathbf{Q}^S_{y,.}}{\mathbf{Q}^S_{y,z}} q^{S'}_{y,a} \right| \\
&= \left| C_1 \frac{\Pi_z^{-1}\mathbf{Q}^S_{y,.}}{\mathbf{Q}^S_{y,z}} \left(\frac{\sum_{i\in S} \hat{Y}_i \mathbb{I}(Y_i = y, Z_i = z)}{n^S_{yz}} - \frac{\sum_{i\in S'} \hat{Y}_i \mathbb{I}(Y_i = y, Z_i = z)}{n^S_{yz}} \right) \right. \\
&\left. + \sum_{a\in\mathcal{A}\setminus z} C_2 \frac{\Pi_a^{-1}\mathbf{Q}^S_{y,.}}{\mathbf{Q}^S_{y,z}} \left(\frac{\sum_{i\in S} \hat{Y}_i \mathbb{I}(Y_i = y, Z_i = a)}{n^S_{ya}} - \frac{\sum_{i\in S'} \hat{Y}_i \mathbb{I}(Y_i = y, Z_i = a)}{n^S_{ya}} \right) \right| \\
&\leq \left| C_1 \frac{\max_a \Pi_a^{-1}\mathbf{Q}^S_{y,.}}{\mathbf{Q}^S_{y,z}} \frac{1}{n^S_{yz}} \right| = \left| C_1 \frac{\max_a C_1 n^S_{ya} + C_2(n - n^S_{ya})}{n^S_{yz}} \frac{1}{n^S_{yz}} \right| \\
&\leq \left| \left(\frac{C_1}{n^S_{yz}} \right)^2 n \right|
\end{aligned}$$

Therefore by McDiarmid's inequality we have:

$$\begin{aligned}
&\sum_{\forall z, S_{yz}: n^S_{yz} \geq \frac{n\mathbf{Q}_{yz}}{2}} \mathbb{P}\left(|\hat{G}_z^{-1} q^S_{y,.} - \gamma_{yz}| > t | S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1} \right) \mathbb{P}\left(S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1} \right) \\
&\leq \sum_{\forall z, S_{yz}: n^S_{yz} \geq \frac{n\mathbf{Q}_{yz}}{2}} 2\exp\left(-\frac{2t^2}{\left(\frac{C_1}{n^S_{yz}}\right)^4 n^3} \right) \mathbb{P}\left(S_{y,0}, \cdots, S_{y,|\mathcal{A}|-1} \right) \\
&\overset{(a)}{\leq} 2\exp\left(-\frac{2t^2}{\left(\frac{2C_1}{n\mathbf{Q}_{yz}}\right)^4 n^3} \right) = 2\exp\left(-2t^2 n \left(\frac{\mathbf{Q}_{yz}}{2C_1} \right)^4 \right)
\end{aligned}$$

step (a) is by noting that the inner quantity is maximized when $n^S_{yz} = \frac{n\mathbf{Q}_{yz}}{2}$, combining things:

$$\mathbb{P}\left(|\hat{G}_z^{-1} q^S_{y,.} - \gamma_{yz}| > t \right) \leq |A| \exp\left(-\frac{\min_a n\mathbf{Q}_{ya}}{8} \right) + 2\exp\left(-2t^2 n \left(\frac{\mathbf{Q}_{yz}}{2C_1} \right)^4 \right)$$

Now if $n \geq \frac{8\log(|\mathcal{A}|/\delta)}{\min_{yz} n\mathbf{Q}_{yz}}$ and $t \geq \sqrt{\frac{\log(2/\delta)}{2n}} \frac{4C_1^2}{\min_{yz} \mathbf{Q}^2_{yz}}$ then we have:

$$\mathbb{P}\left(|\hat{G}_z^{-1} q^S_{y,.} - \gamma_{yz}| > t \right) \leq \delta + \delta$$

Step 3: *Projecting the estimator onto the simplex*

One issue that arises is that our estimator for $\gamma_{y,z}$ does not lie in the range $[0, 1]$, and hence we have to project the whole vector onto the simplex for it to be valid; note that this is not required if we are only interested in differences i.e. computing discrimination. Our estimator for the vector of conditional probabilities for $a \in \mathcal{A}$ of $\mathbb{P}(\hat{Y} = 1|Y = y, A = a)$ is $\text{Proj}_{\Delta}(\hat{G}^{-1} q_{y,.})$ where $\text{Proj}_{\Delta}(x)$ is the orthogonal projection of x onto the simplex defined as:

$$\begin{aligned}
\text{Proj}_{\Delta}(\mathbf{x}) := \quad & \arg\min_{\mathbf{y}} \frac{1}{2} ||\mathbf{y} - \mathbf{x}||_2^2 \\
& \text{s.t. } \mathbf{y}^T \mathbf{1} = 1, \ \mathbf{y} \geq 0
\end{aligned}$$

The above problem can be solved optimally in a non-iterative manner in time $\mathcal{O}\left(|\mathcal{A}| \log(|\mathcal{A}|\right)$ (Duchi et al. (2008)). Denote by $\mathbf{x}' = \text{Proj}_{\Delta}(\mathbf{x})$, then by the definition of the projection for any $\mathbf{y} \in \Delta^{|\mathcal{A}|}$:

$$|\mathbf{x}' - \mathbf{y}| \leq |\mathbf{x} - \mathbf{y}|$$

however it does not hold that $||\mathbf{x}' - \mathbf{y}||_\infty \leq ||\mathbf{x} - \mathbf{y}||_\infty$, but : $||\mathbf{x}' - \mathbf{y}||_\infty \leq |\mathcal{A}| \cdot ||\mathbf{x} - \mathbf{y}||_\infty$. Therefore:

$$\mathbb{P}\left(\left|\text{Proj}_\Delta(\hat{G}_k^{-1}q_{y,.}^S) - \mathbb{P}(\hat{Y}=1|Y=y,A=k)\right| > t\right) \leq \mathbb{P}\left(\max_k \left|\hat{G}_k^{-1}q_{y,.}^S - \mathbb{P}(\hat{Y}=1|Y=y,A=k)\right| > \frac{t}{|\mathcal{A}|}\right)$$

Step 4: *Difference of Equalized odds*

Let $h_{ya} = \text{Proj}_\Delta(\hat{G}_a^{-1}q_{y,.}^S)$, using a series of triangle inequality,

$$\left||h_{ya}^S - h_{y0}^S| - |h_{ya} - h_{y0}|\right| \leq |h_{ya}^S - h_{y0}^S - h_{ya} + h_{y0}| \leq |h_{ya}^S - h_{ya}| + |h_{y0}^S - h_{y0}|$$

hence

$$\begin{aligned}\mathbb{P}\left(\left||h_{ya}^S - h_{y0}^S| - |h_{ya} - h_{y0}|\right| > 2t\right) &\leq \mathbb{P}\left(|h_{ya}^S - h_{ya}| + |h_{y0}^S - h_{y0}| > 2t\right) \\ &\overset{(a)}{\leq} \mathbb{P}\left(|h_{ya}^S - h_{ya}| > t\right) + \mathbb{P}\left(|h_{y0}^S - h_{y0}| > t\right) \\ &\leq 4\delta\end{aligned}$$

where (a) follows from union bound, and (b) follows from above using $n \geq \frac{8\log(|\mathcal{A}|/\delta)}{\min_{yz} n\mathbf{Q}_{yz}}$ and $t \geq \sqrt{\frac{\log(2/\delta)}{2n}\frac{4|\mathcal{A}|C_1^2}{\min_{yz}\mathbf{Q}_{yz}^2}}$ The lemma follows from collecting the failure probabilities for $y = 0, 1$, re-scaling δ and noting that $\min_{yz} \mathbf{Q}_{yz} \geq \min_{ya} \mathbf{P}_{ya}$.

Now let us write t in terms of ϵ, we write each of the factors involving π in terms of ϵ:

$$C_1 = \frac{\pi + |\mathcal{A}| - 2}{|\mathcal{A}|\pi - 1} = \frac{|\mathcal{A}| - 2 + e^\epsilon}{e^\epsilon - 1}$$

and:

$$C_1^2 = \frac{e^{2\epsilon} + 2(|\mathcal{A}| - 2)e^\epsilon + (|\mathcal{A}| - 2)^2}{e^{2\epsilon} - 2e^\epsilon + 1} \leq \frac{2|\mathcal{A}|^2 e^{2\epsilon}}{e^{2\epsilon} - 2e^\epsilon + 1}$$

□

A.2. Section 5

A.2.1. First Step Algorithm Details

Recall that in Algorithm 1, the learner's best response gaced a given vector $\boldsymbol{\lambda}$ ($\text{BEST}_h(\boldsymbol{\lambda})$) puts all the mass on a single predictor $h \in \mathcal{H}$ as the langragian L is linear in Q. (Agarwal et al., 2018) shows that finding the learner's best response amounts to solving a cost sensitive classification problem. We now re-establish this reduction:

$$\begin{aligned}L(h, \boldsymbol{\lambda}) &= \hat{\text{err}}(h) + \boldsymbol{\lambda}^\top(M\boldsymbol{\gamma}(h) - \alpha_n \mathbf{1}) \\ &= \frac{1}{n}\sum_{i \in S}\mathbb{I}_{h(x_i) \neq y_i} - \alpha_n\boldsymbol{\lambda}^\top\mathbf{1} + \sum_{k,j} M_{k,j}\lambda_k\gamma_j^S(h) \\ &= -\alpha_n\boldsymbol{\lambda}^\top\mathbf{1} + \frac{1}{n}\sum_{i \in S}\mathbb{I}_{h(x_i) \neq y_i} + \sum_{k,j} M_{k,j}\lambda_k\frac{1/n \cdot h(x_i)\mathbb{I}_{(y_i,a_i)=j}}{1/n\sum_{s \in S}\mathbb{I}_{(y_s,a_s)=j}} \qquad (4)\end{aligned}$$

Thus from equation (19) and expanding the form of the matrix M we have that minimizing $L(h, \lambda)$ over $h \in \mathcal{H}$ is equivalent to solving a cost sensitive classification problem on $\{(x_i, c_i^0, c_i^1)\}_{i=1}^n$ where the costs are:

$$\begin{aligned}c_i^0 &= \mathbb{I}_{y_i \neq 0} \\ c_i^1 &= \mathbb{I}_{y_i \neq 1} + \frac{\lambda_{(a_i,y_i,+)} - \lambda_{(a_i,y_i,-)}}{p_{a_i,y_i}^S}\mathbb{I}_{a_i \neq 0} - \sum_{a \in \mathcal{A}\setminus\{0\}}\frac{\lambda_{(a,y_i,+)} - \lambda_{(a,y_i,-)}}{p_{0,y_i}^S}\end{aligned}$$

where $p^S_{a,y} = \frac{1}{n}\sum_{s\in S}\mathbb{I}_{(y_s=y,a_s=a)}$.

The goal of Algorithm 1 is to return for any degree of approximation $\vartheta \in \mathbb{R}^+$ a ϑ-approximate saddle point $(\hat{Q},\hat{\lambda})$ defined as:

$$L(\hat{Q},\hat{\boldsymbol{\lambda}}) \leq L(Q,\hat{\boldsymbol{\lambda}}) + \vartheta \;\; \forall Q \in \Delta_{\mathcal{H}} \tag{5}$$

$$L(\hat{Q},\hat{\boldsymbol{\lambda}}) \geq L(\hat{Q},\boldsymbol{\lambda}) - \vartheta \;\; \forall \boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{K}|}_+, ||\boldsymbol{\lambda}||_1 \leq B \tag{6}$$

From Theorem 1 in (Agarwal et al., 2018), if we run the algorithm for at least $\frac{16\log(4|\mathcal{A}|+1)}{\vartheta^2}$ iterations with learning rate $\eta = \frac{\vartheta}{8B}$ it returns a ϑ-approximate saddle point.

A.2.2. First Step Guarantees

Lemma 1. *Denote by $\mathbf{Q}_{yz} = \mathbb{P}(Y=y, Z=z)$, $q_{yz}(\hat{Y}) = \mathbb{P}(\hat{Y}=1|Y=y,Z=z)$, for $\delta \in (0,1/2)$ and h any binary predictor, if $n \geq \frac{8\log 8|\mathcal{A}|/\delta}{\min_{yz} Q_{yz}}$, then:*

$$\mathbb{P}\left(\max_{ya}\left||q^S_{ya}-q^S_{y0}| - |q_{ya}-q_{y0}|\right| > 2\sqrt{\frac{\log 16|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}\right) \leq \delta$$

Proof. Let $a \in \mathcal{A}$, denote $\mathbf{Q}_{yz} = \mathbb{P}(Y=y,Z=z)$, $q_{yz}(\hat{Y}) = \mathbb{P}(\hat{Y}=1|Y=y,Z=z)$, then by (Woodworth et al., 2017) or step 1 of Lemma 1:

$$\mathbb{P}\left(|q^S_{yz}-q_{yz}| > t\right) \leq \exp\left(-\frac{n\mathbf{Q}_{yz}}{8}\right) + 2\exp\left(-t^2 n\mathbf{Q}_{yz}\right)$$

Now using a series of triangle inequality identical to step 4 of Lemma 1,

$$\left||q^S_{ya}-q^S_{y0}| - |q_{ya}-q_{y0}|\right| \leq |q^S_{ya}-q^S_{y0}-q_{ya}+q_{y0}| \leq |q^S_{ya}-q_{y0}| + |q^S_{y0}-q_{y0}|$$

hence

$$\begin{aligned}\mathbb{P}\left(\left||q^S_{ya}-q^S_{y0}| - |q_{ya}-q_{y0}|\right| > 2t\right) &\leq \mathbb{P}\left(|q^S_{ya}-q_{y0}| + |q^S_{y0}-q_{y0}| > 2t\right)\\ &\overset{(a)}{\leq} \mathbb{P}\left(|q^S_{ya}-q_{y0}| > t\right) + \mathbb{P}\left(|q^S_{y0}-q_{y0}| > t\right)\\ &\leq 2\exp\left(-\frac{n\min_{yz}\mathbf{Q}_{yz}}{8}\right) + 4\exp\left(-t^2 n\min_{yz}\mathbf{Q}_{yz}\right)\\ &\overset{(b)}{\leq} \frac{\delta}{2|\mathcal{A}|}\end{aligned}$$

where (a) follows from union bound, and (b) follows if $n \geq \frac{8\log 8|\mathcal{A}|/\delta}{\min_{yz} Q_{yz}}$ and $t = \sqrt{\frac{\log 16|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$

The lemma follows from collecting the failure probabilities for $y=0,1$ and $\forall a \in \mathcal{A}$. □

Lemma 2. *If a binary predictor $\hat{Y}$ is independent of Z given A, then if the groups are binary it holds that:*

$$q_{y1}(\hat{Y}) - q_{y0}(\hat{Y}) = \left(\gamma_{y1}(\hat{Y}) - \gamma_{y0}(\hat{Y})\right)\frac{(2\pi-1)\mathbf{P}_{y1}\mathbf{P}_{y0}}{\mathbf{Q}_{y1}\mathbf{Q}_{y0}} \tag{7}$$

For general $|\mathcal{A}|$ different groups, we have $\forall k,j \in \mathcal{A}$ the following relation:

$$|\gamma_{y,k}-\gamma_{y,j}| \leq 5C\frac{\max_i \mathbb{P}(Z=i,Y=y)}{\min_j \mathbb{P}(A=j,Y=y)^2}\left|\max_z q_{y,z} - \min_{z'} q_{y,z'}\right|$$

where $C = \frac{\pi+|\mathcal{A}|-2}{|\mathcal{A}|\pi-1}$.

Proof. We begin by noting the following relationship established in step 4 of Lemma 1:

$$\begin{aligned}\mathbb{P}(\hat{Y}=1|Y=y,Z=a) = \pi\mathbb{P}(\hat{Y}=1|Y=y,A=a)&\frac{\mathbb{P}(Y=y,A=a)}{\mathbb{P}(Z=a,Y=y)}\\ +\sum_{a'\setminus a}\bar{\pi}\mathbb{P}(\hat{Y}=1|Y=y,A=a')&\frac{\mathbb{P}(Y=y,A=a')}{\mathbb{P}(Z=a,Y=y)}\end{aligned}$$

From the above equation, we can evaluate for any $a,b\in\mathcal{A}$ the difference between q_{ya} and q_{yb} in terms of $\gamma_{y.}$, denoting $\mathbf{P}_{ya}=\mathbb{P}(Y=y,A=a)$:

$$\begin{aligned}
q_{ya}-q_{yb} &= \gamma_{ya}\frac{\pi\mathbf{P}_{ya}}{\mathbf{Q}_{ya}}+\sum_{a'\setminus a}\gamma_{ya'}\frac{\bar{\pi}\mathbf{P}_{ya'}}{\mathbf{Q}_{ya}}-\gamma_{yb}\frac{\pi\mathbf{P}_{yb}}{\mathbf{Q}_{yb}}-\sum_{b'\setminus b}\gamma_{yb'}\frac{\bar{\pi}\mathbf{P}_{yb'}}{\mathbf{Q}_{yb}}\\
&= \frac{\gamma_{ya}\pi\mathbf{P}_{ya}\mathbf{Q}_{yb}+\gamma_{yb}\bar{\pi}\mathbf{P}_{yb}\mathbf{Q}_{yb}+\sum_{a'\setminus\{a,b\}}\bar{\pi}\gamma_{ya'}\mathbf{P}_{ya'}\mathbf{Q}_{yb}}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&- \frac{\gamma_{yb}\pi\mathbf{P}_{yb}\mathbf{Q}_{ya}+\gamma_{ya}\bar{\pi}\mathbf{P}_{ya}\mathbf{Q}_{ya}+\sum_{b'\setminus\{a,b\}}\bar{\pi}\gamma_{yb'}\mathbf{P}_{yb'}\mathbf{Q}_{ya}}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&= \frac{\gamma_{ya}\mathbf{P}_{ya}(\pi\mathbf{Q}_{yb}-\bar{\pi}\mathbf{Q}_{ya})-\gamma_{yb}\mathbf{P}_{yb}(\pi\mathbf{Q}_{ya}-\bar{\pi}\mathbf{Q}_{yb})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}+\frac{\sum_{c\setminus\{a,b\}}\bar{\pi}\gamma_{yc}\mathbf{P}_{yc}(\mathbf{Q}_{yb}-\mathbf{Q}_{ya})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&\overset{(a)}{=} \frac{\gamma_{ya}\mathbf{P}_{ya}(\pi(\pi\mathbf{P}_{yb}+\bar{\pi}\mathbf{P}_{ya}+\bar{\pi}\sum_{c\setminus\{a,b\}}\mathbf{P}_{yc})-\bar{\pi}(\pi\mathbf{P}_{ya}+\bar{\pi}\mathbf{P}_{yb}+\bar{\pi}\sum_{c\setminus\{a,b\}}\mathbf{P}_{yc})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&- \frac{\gamma_{yb}\mathbf{P}_{yb}(\pi(\pi\mathbf{P}_{ya}+\bar{\pi}\mathbf{P}_{yb}+\bar{\pi}\sum_{c\setminus\{a,b\}}\mathbf{P}_{yc})-\bar{\pi}(\pi\mathbf{P}_{yb}+\bar{\pi}\mathbf{P}_{ya}+\bar{\pi}\sum_{c\setminus\{a,b\}}\mathbf{P}_{yc}))}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&+ \frac{\sum_{c\setminus\{a,b\}}\bar{\pi}\gamma_{yc}\mathbf{P}_{yc}(\mathbf{Q}_{yb}-\mathbf{Q}_{ya})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&= \frac{\gamma_{ya}\mathbf{P}_{ya}(\pi^2\mathbf{P}_{yb}-\bar{\pi}^2\mathbf{P}_{yb}+(\pi-\bar{\pi})\bar{\pi}\sum_{c\setminus\{a,b\}}\mathbf{P}_{yc})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&- \frac{\gamma_{yb}\mathbf{P}_{yb}(\pi^2\mathbf{P}_{ya}-\bar{\pi}^2\mathbf{P}_{ya}+(\pi-\bar{\pi})\bar{\pi}\sum_{c\setminus\{a,b\}}\mathbf{P}_{yc})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}+\frac{\sum_{c\setminus\{a,b\}}\bar{\pi}\gamma_{yc}\mathbf{P}_{yc}(\mathbf{Q}_{yb}-\mathbf{Q}_{ya})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&= \frac{(\gamma_{ya}-\gamma_{yb})\mathbf{P}_{ya}\mathbf{P}_{yb}(\pi^2-\bar{\pi}^2)}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}\\
&+ \frac{(\gamma_{ya}\mathbf{P}_{ya}-\gamma_{yb}\mathbf{P}_{yb})(\pi-\bar{\pi})\bar{\pi}\sum_{c\setminus\{a,b\}}\mathbf{P}_{yc}}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}+\frac{\sum_{c\setminus\{a,b\}}\bar{\pi}\gamma_{yc}\mathbf{P}_{yc}(\mathbf{Q}_{yb}-\mathbf{Q}_{ya})}{\mathbf{Q}_{ya}\mathbf{Q}_{yb}}
\end{aligned}$$

where step (a) follows from expanding by equation (18). If $\mathcal{A}=\{0,1\}$ then the above reduces to:

$$q_{y1}-q_{y0}=(\gamma_{y1}-\gamma_{y0})\frac{(2\pi-1)\mathbf{P}_{y1}\mathbf{P}_{y0}}{\mathbf{Q}_{y1}\mathbf{Q}_{y0}}$$

Now when the groups are not binary we instead rely on an upper bound.

Let $q_y=[\mathbb{P}(\hat{Y}=1|Y=y,Z=0),\cdots,\mathbb{P}(\hat{Y}=1|Y=y,Z=|\mathcal{A}|-1)]^\top$, $\gamma=[\mathbb{P}(\hat{Y}=1|Y=y,A=0),\cdots,\mathbb{P}(\hat{Y}=$

$1|Y=y, A=|\mathcal{A}|-1)]^\top$, in the proof of Lemma 1 we established that $G^{-1}q_y=\gamma_y$, now let $k, j\in\mathcal{A}$ then we have:

$$\begin{aligned}
|\gamma_{y,k}-\gamma_{y,j}| &= |G_k^{-1}q_y - G_j^{-1}q_y| \\
&\overset{(a)}{=} |G_k^{-1}(q_y-q') - G_j^{-1}(q_y-q')| \\
&= |(q_y-q')(G_k^{-1}-G_j^{-1})| \\
&\leq |q_y-q'|_\infty |G_k^{-1}-G_j^{-1}|_1 \text{ (Holder's Inequality)} \\
&= |\max_z q_{y,z} - \min_{z'} q_{y,z'}| \cdot |G_k^{-1}-G_j^{-1}|_1
\end{aligned} \tag{8}$$

in step (a) we introduce $q'=[\min_z q_{y,z},\cdots,\min_z q_{y,z}]^\top$, and note that $G_k^{-1}q'=\min_z q_{y,z}$ as the rows of G sum to 1 by the proof of Proposition 1, therefore $G_k^{-1}q'=G_j^{-1}q'$ and the difference in the previous step is unchanged. Now let us take expand the right most term in equation (23), for ease of notation let $\mathbb{P}(Z=i,Y=y)=z_i$ and $\mathbb{P}(A=i,Y=y)=a_i$:

$$\begin{aligned}
&|G_k^{-1}-G_j^{-1}|_1 \\
&= \sum_{a\in\mathcal{A}\setminus\{k,j\}} \left|C_2(\frac{z_a}{a_k}-\frac{z_a}{a_i})\right| + \left|C_1\frac{z_k}{a_k}-C_2\frac{z_k}{a_i}\right| + \left|C_2\frac{z_i}{a_k}-C_1\frac{z_i}{a_i}\right| \\
&= \sum_{a\in\mathcal{A}\setminus\{k,j\}} \left|C_2(\frac{z_a a_i - z_a a_k}{a_k a_i})\right| + \left|C_1\frac{z_k a_i}{a_k a_i}-C_2\frac{z_k a_k}{a_k a_i}\right| + \left|C_2\frac{z_i a_i}{a_k a_i}-C_1\frac{z_i a_k}{a_k a_i}\right| \\
&= \sum_{a\in\mathcal{A}\setminus\{k,j\}} \left|C_2\frac{z_a(a_i-a_k)}{a_k a_i}\right| + \left|\frac{z_k(C_1 a_i - C_2 a_k)}{a_k a_i}\right| + \left|\frac{z_i(C_2 a_i - C_1 a_k)}{a_k a_i}\right| \\
&\leq \max_a z_a \cdot \left((|\mathcal{A}|-2)\left|C_2\frac{(a_i-a_k)}{a_k a_i}\right| + \left|\frac{(C_1 a_i - C_2 a_k)}{a_k a_i}\right| + \left|\frac{(C_2 a_i - C_1 a_k)}{a_k a_i}\right|\right) \\
&\leq \max_a z_a \cdot \left((|\mathcal{A}|-2)\left|C_2\frac{1}{\min_z a_z^2}\right| + \left|2C_1\frac{1}{\min_z a_z^2}\right| + \left|2C_1\frac{1}{\min_z a_z^2}\right|\right) \\
&\leq \frac{\max_a z_a}{\min_z a_z^2}\cdot((|\mathcal{A}|-2)\,|C_2|+4C_1) \leq \frac{\max_i \mathbb{P}(Z=i,Y=y)}{\min_j \mathbb{P}(A=j,Y=y)^2}5C_1
\end{aligned}$$

Hence we have the following inequality:

$$|\gamma_{y,k}-\gamma_{y,j}| \leq 5C_1\frac{\max_i \mathbb{P}(Z=i,Y=y)}{\min_j \mathbb{P}(A=j,Y=y)^2}\left|\max_z q_{y,z} - \min_{z'} q_{y,z'}\right|$$

where $C_1=\frac{\pi+|\mathcal{A}|-2}{|\mathcal{A}|\pi-1}$.

□

We now recall some helper lemmas from (Agarwal et al., 2018).

Lemma 3 (Lemma 2 (Agarwal et al., 2018))**.** *For any distribution Q satisfying the empirical constraints on dataset S: $M\gamma^S(Q)\leq\alpha_n\mathbf{1}$, $\hat{Q}$ the output $\hat{Y}$ of Algorithm 1 satisfies:*

$$\text{err}^S(\hat{Y}) \leq \text{err}(Q) + 2\vartheta \tag{9}$$

Lemma 4 (Lemma 3 (Agarwal et al., 2018))**.** *The discrimination of $\hat{Y}$, output of Algorithm 1, satisfies:*

$$\max_{y,a} |q^S_{y,a}(\hat{Y}) - q^S_{y,0}(\hat{Y})| \leq 2\alpha_n + 2\frac{1+2\vartheta}{B} \tag{10}$$

Lemma 2 [Guarantees for Step 1] *Given a hypothesis class $\mathcal{H}$, a distribution over (X,A,Y), $B\in\mathbb{R}^+$ and any $\delta\in(0,1/2)$, then with probability greater than $1-\delta$, if $n\geq\frac{16\log 8|\mathcal{A}|/\delta}{\min_{ya}\mathbf{P}_{ya}}$, $\alpha_n=2\sqrt{\frac{\log|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}$ and we let $\vartheta=\mathfrak{R}_{n/2}(\mathcal{H})+\sqrt{\frac{\log 8/\delta}{n}}$,* *then running Algorithm 1 on dataset S with $T \geq \frac{16\log(4|\mathcal{A}|+1)}{\vartheta^2}$ and learning rate $\eta = \frac{\vartheta}{8B}$ returns a predictor $\hat{Y}$ satisfying the following:*

$$\mathrm{err}(\hat{Y}) \leq_{\delta/2} \mathrm{err}(Y^*) + 4\mathfrak{R}_{n/2}(\mathcal{H}) + 4\sqrt{\frac{\log 1/\delta}{n}}$$

$$\mathrm{disc}(\hat{Y}) \leq_{\delta/2} \frac{5C}{\min_{ya}\mathbf{P}_{ya}^2}\left(\frac{2}{B} + 6\mathfrak{R}_{\frac{\min_{ya} n\mathbf{P}_{ya}}{4}}(\mathcal{H}) + 10\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}\right)$$

Proof. From Theorem 1 in (Agarwal et al., 2018), if we run the algorithm for at least $\frac{16\log(4|\mathcal{A}|+1)}{\vartheta^2}$ iterations with learning rate $\eta = \frac{\vartheta}{8B}$ it returns a ϑ-approximate saddle point. We set ϑ at the end of the proof to balance the bounds.

For step 1 we have access to $S_1 = \{(x_i, y_i, z_i)\}_{i=1}^{n/2}$, denote by $\mathrm{err}(\hat{Y}) = \mathbb{P}(\hat{Y} \neq Y)$, using the Rademacher complexity bound (Theorem 3.5 (Mohri et al., 2018)) and the fact that $\mathfrak{R}_n(\Delta_{\mathcal{H}}) = \mathfrak{R}_n(\mathcal{H})$ we have:

$$\mathrm{err}(\hat{Y}) \leq_{\delta/4} \mathrm{err}^S(\hat{Y}) + \mathfrak{R}_{n/2}(\mathcal{H}) + \sqrt{\frac{\log 8/\delta}{n}} \tag{11}$$

Now from Lemma 5 of (Woodworth et al., 2017), with probability greater than $1-\delta/4$, Y^* is in the feasible set of step 1 if $\alpha_n \geq 2\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$, hence we can apply Lemma 7 with Y^* and the concentration bound (26) :

$$\mathrm{err}(\hat{Y}) \leq_{\delta/2} \mathrm{err}(Y^*) + 2\vartheta + 2\mathfrak{R}_{n/2}(\mathcal{H}) + 2\sqrt{\frac{\log 8/\delta}{n}}$$

For the constraint, from Lemma 5, if $n \geq \frac{16\log 8|\mathcal{A}|/\delta}{\min_{yz}\mathbf{Q}_{yz}}$, then

$$\max_{ya}\left||q_{ya}^S - q_{y0}^S| - |q_{ya} - q_{y0}|\right| \leq_{\delta/4} 2\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$$

Similarly from the standard Rademacher complexity bound (Theorem 3.3 (Mohri et al., 2018)) and since our function class for the constraint is $\mathcal{H}$ it holds that (by Lemma 6 (Agarwal et al., 2018)):

$$\max_{ya}|q_{ya} - q_{y0}| \leq_{\delta/4} |q_{ya}^S - q_{y0}^S| + 2\mathfrak{R}_{\frac{\min_{yz} n\mathbf{Q}_{yz}}{4}}(\mathcal{H}) + 2\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$$

Applying Lemma 8:

$$|q_{ya}^S - q_{y0}^S| \leq 2\alpha_n + 2\frac{1+2\vartheta}{B} \tag{12}$$

Combining things with $\alpha_n = 2\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$:

$$\max_{ya}|q_{ya} - q_{y0}| \leq_{\delta/4} \frac{2+4\vartheta}{B} + 2\mathfrak{R}_{\frac{\min_{yz} n\mathbf{Q}_{yz}}{4}}(\mathcal{H}) + 6\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$$

Now by Lemma 6 we can re-state the above in terms of A:

$$\max_{a}|\gamma_{y,a} - \gamma_{y,0}| \leq_{\delta/4} \frac{5C}{\min_{ya}\mathbf{P}_{ya}^2}\left(\frac{2+4\vartheta}{B} + 2\mathfrak{R}_{\frac{\min_{yz} n\mathbf{Q}_{yz}}{4}}(\mathcal{H}) + 6\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}\right)$$

For simplicity, we can thus set $\vartheta = \mathfrak{R}_{n/2}(\mathcal{H}) + \sqrt{\frac{\log 8/\delta}{n}}$, by noting that $\min_{yz}\mathbf{Q}_{yz} \geq \min_{ya}\mathbf{P}_{ya}$ we obtain the lemma statement.

□

A.2.3. Second Step Algorithm Details

Given a predictor $\hat{Y}$, Hardt et al. give a simple procedure to obtain a derived predictor $\widetilde{Y}$ that is non-discriminatory (Hardt et al., 2016) by solving a constrained linear program (LP). One of the caveats of the approach is that it requires the use of the protected attribute at test time, and in our setting we do not have access to A but Z. We have seen in section 4 that predictors that rely on Z cannot be trusted even if they are completely non-discriminatory with respect to the privatized attribute. Despite this difficulty, it turns out if the base predictor $\hat{Y}$ is independent of Z given A, then we can re-write the LP to obtain a derived predictor $\widetilde{Y} = h(\hat{Y}, Z)$ that minimizes the error while being non-discriminatory with respect to A.

The approach boils down to solving the following linear program (LP):

$$\begin{aligned}
\min \quad & \mathbb{P}(\widetilde{Y} \neq Y) \\
s.t. \quad & \mathbb{P}(\widetilde{Y} = 1 | A = a, Y = y) = \mathbb{P}(\widetilde{Y} = 1 | Y = y, A = 0) \\
& \forall y \in \{0,1\}, \forall a \in \mathcal{A}
\end{aligned}$$

We can write this objective by optimizing over $2|\mathcal{A}|$ probabilities $p_{\hat{y},a} := \mathbb{P}(\widetilde{Y} = 1 | \hat{Y} = \hat{y}, A = a)$ that completely specify the behavior of $\widetilde{Y}$:

$$\widetilde{Y} = \arg\min_{p_{\cdot,\cdot}} \quad \sum_{\hat{y},a} (\mathbb{P}(\hat{Y} = \hat{y}, A = a, Y = 0) - \mathbb{P}(\hat{Y} = \hat{y}, A = a, Y = 1)) \cdot p_{\hat{y},a} \tag{13}$$

$$s.t. \quad p_{0,a}\mathbb{P}(\hat{Y} = 0 | Y = y, A = a) + p_{1,a}\mathbb{P}(\hat{Y} = 1 | Y = y, A = a) \tag{14}$$

$$= p_{0,0}\mathbb{P}(\hat{Y} = 0 | Y = y, A = 0) + p_{1,0}\mathbb{P}(\hat{Y} = 1 | Y = y, A = 0), \quad \forall y \in \{0,1\}, \forall a \in \mathcal{A} \tag{15}$$

$$0 \leq p_{\hat{y},a} \leq 1 \quad \forall \hat{y} \in \{0,1\}, \forall a \in \mathcal{A}$$

Unfortunately we cannot directly solve the above program as we do not have access to A, however we can solve the problem with Z replacing A; we denote this as the naïve program and as we have previously mentioned it cannot assure any degree of non-discrimination with respect to A. Now let us see how we can transform this naïve program to satisfy equalized odds. We optimize over the set of variables that denote $p_{\hat{y},z} := \mathbb{P}(\widetilde{Y} = 1 | \hat{Y} = \hat{y}, Z = z)$. Now for the constraint note that $\mathbb{P}(\widetilde{Y} = 1 | \hat{Y} = \hat{y}, A = a)$ can be expressed as a mixture of our decision variables:

$$\begin{aligned}
\mathbb{P}(\widetilde{Y} = 1 | \hat{Y} = \hat{y}, A = a) &= \sum_{a'} \mathbb{P}(\widetilde{Y} = 1 | \hat{Y} = \hat{y}, Z = a', A = a) \mathbb{P}(Z = a' | A = a, \hat{Y} = \hat{y}) \\
&= \pi \mathbb{P}(\widetilde{Y} = 1 | \hat{Y} = \hat{y}, Z = a) + \sum_{a' \setminus a} \bar{\pi} \mathbb{P}(\widetilde{Y} = 1 | \hat{Y} = \hat{y}, Z = a')
\end{aligned}$$

Since we assumed the base predictor $\hat{Y}$ is independent of Z given A then $\mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = a)$ can be recovered from the following linear system by using the same estimator we developed previously in Lemma 1:

$$\begin{aligned}
\mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = a) &= \pi \mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = a) \frac{\mathbb{P}(A = a, Y = y)}{\mathbb{P}(Z = a, Y = y)} \\
&+ \sum_{a' \neq a} \bar{\pi} \mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = a) \frac{\mathbb{P}(A = a', Y = y)}{\mathbb{P}(Z = a, Y = y)}
\end{aligned}$$

On the other hand for the objective we have:

$$\mathbb{P}(\hat{Y} = \hat{y}, Z = a, Y = y) = \pi \mathbb{P}(\hat{Y} = \hat{y}, A = a, Y = y) + \bar{\pi} \sum_{a' \neq a} \mathbb{P}(\hat{Y} = \hat{y}, A = a', Y = y)$$

And hence our estimator for $\mathbb{P}(\hat{Y} = \hat{y}, A = a, Y = y)$ is constructed by multiplying by the inverse of Π and projecting onto the simplex.

Denote by $\widetilde{p}_{\hat{y},a} = \pi p_{\hat{y},a} + \sum_{a'\backslash a} \hat{\pi} p_{\hat{y},a'}$ and $\widetilde{\mathbb{P}}^S(\hat{Y} = \hat{y}|Y = y, A = a)$ our estimator for $\mathbb{P}(\hat{Y} = \hat{y}|Y = y, A = a)$ and similarly $\widetilde{\mathbb{P}}^S(\hat{Y} = \hat{y}, Y = y, A = a)$. We propose to solve the following optimization problem:

$$\widetilde{Y} = \arg\min_{p_{\cdot,\cdot}} \quad \sum_{\hat{y},a} (\widetilde{\mathbb{P}}^S(\hat{Y} = \hat{y}, Z = a, Y = 0) - \widetilde{\mathbb{P}}^S(\hat{Y} = \hat{y}, Z = a, Y = 1)) \cdot \widetilde{p}_{\hat{y},a} \tag{16}$$

$$\begin{aligned} s.t. \quad & \widetilde{p}_{0,a}\widetilde{\mathbb{P}}^S(\hat{Y} = 0|Y = y, A = a) + \widetilde{p}_{1,a}\widetilde{\mathbb{P}}^S(\hat{Y} = 1|Y = y, A = a) \\ & = \widetilde{p}_{0,0}\widetilde{\mathbb{P}}^S(\hat{Y} = 0|Y = y, A = 0) + \widetilde{p}_{1,0}\widetilde{\mathbb{P}}^S(\hat{Y} = 1|Y = y, A = 0), \quad \forall y \in \{0,1\}, \forall a \in \mathcal{A} \qquad (17) \\ & 0 \le p_{\hat{y},a} \le 1 \quad \forall \hat{y} \in \{0,1\}, \forall a \in \mathcal{A} \end{aligned}$$

A.2.4. Second Step Guarantees

Lemma 5 (Step 2 guarantees)**.** *Let $\hat{Y}$ be a binary predictor that is independent of Z given A, for any $\delta \in (0, 1/2)$, if $n \ge \frac{32\log(8|\mathcal{A}|/\delta)}{\min_{ya}\mathbf{P}_{ya}}$, $\widetilde{\alpha}_n = \sqrt{\frac{\log(64/\delta)}{2n}\frac{4|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya*}}}$ and with $\widetilde{Y}*$ an optimal 0-discriminatory predictor derived from $\hat{Y}$, then with probability greater than $1-\delta/2$ we have:*

$$\text{err}(\widetilde{Y}) \le \text{err}(\widetilde{Y}^*) + 4|\mathcal{A}|C\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}$$

$$\text{disc}(\widetilde{Y}) \le \sqrt{\frac{\log(\frac{64}{\delta})}{2n}\frac{8|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya}}}$$

Proof. Denote $\text{err}(\widetilde{Y}) = \mathbb{P}(\widetilde{Y} \neq Y)$ and $q_{\hat{y},a,y} := \mathbb{P}(\hat{Y} = \hat{y}, Z = a, Y = 1)$, then for any $\widetilde{Y}$ in the derived set of $\hat{Y}$ (also by Lemma B.2 Jagielski et al. (2018)):

$$\begin{aligned} \left|\text{err}^S(\widetilde{Y}) - \text{err}(\widetilde{Y})\right| &= \left|\sum_{\hat{y},a} \widetilde{p}_{\hat{y},a} \cdot \left((\widetilde{q}^S_{\hat{y},a,0} - q_{\hat{y},a,0}) + (q_{\hat{y},a,1} - \widetilde{q}^S_{\hat{y},a,1})\right)\right| \\ &\le |\sum_{\hat{y},a} \widetilde{q}^S_{\hat{y},a,0} - q_{\hat{y},a,0}| + |\sum_{\hat{y},a} q_{\hat{y},a,1} - \widetilde{q}^S_{\hat{y},a,1}| \\ &\le \sum_{\hat{y},a} |\widetilde{q}^S_{\hat{y},a,0} - q_{\hat{y},a,0}| + \sum_{\hat{y},a} | q_{\hat{y},a,1} - \widetilde{q}^S_{\hat{y},a,1}| \end{aligned} \tag{18}$$

Now our estimator $\widetilde{q}^S_{\hat{y},a,y}$ for $q_{\hat{y},a,y}$ is obtained by multiplying by the inverse of the matrix Π and projecting onto the simplex, as was done in Lemma 1. Using the same arguments of step 2 of the proof of Lemma 1 using Mcdirmid's inequality we have:

$$\mathbb{P}\left(|\widetilde{q}^S_{\hat{y},a,y} - q_{\hat{y},a,y}| > t\right) \le 2\exp(-\frac{2t^2n}{C^2})$$

Hence

$$\begin{aligned} \mathbb{P}(\left|\text{err}^S(\widetilde{Y}) - \text{err}(\widetilde{Y})\right| > t) &\le \mathbb{P}(\sum_{\hat{y},a} |q^S_{\hat{y},a,0} - q_{\hat{y},a,0}| + \sum_{\hat{y},a} | q_{\hat{y},a,1} - q^S_{\hat{y},a,1}| > t) \\ &\le 8|\mathcal{A}|\exp(-2n\left(\frac{t}{4|\mathcal{A}|}\frac{|\mathcal{A}|\pi - 1}{\pi + |\mathcal{A}| - 2}\right)^2) \end{aligned} \tag{19}$$

Thus if $t \ge \frac{4|\mathcal{A}|(\pi+|\mathcal{A}|-2)}{|\mathcal{A}|\pi-1}\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}$:

$$\mathbb{P}\left(\left|\text{err}^S(\widetilde{Y}) - \text{err}(\widetilde{Y})\right| > \frac{4|\mathcal{A}|(\pi + |\mathcal{A}| - 2)}{|\mathcal{A}|\pi - 1}\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}\right) \le \delta/4$$

Now for the fairness constraint, denote $\Gamma_{y,a}(\widetilde{Y}) = |\mathbb{P}(\widetilde{Y}=1|Y=\hat{y}, A=a) - \mathbb{P}(\widetilde{Y}=1|Y=y, A=0)|$, then:

$$\begin{aligned}
&|\widetilde{\Gamma}^S_{y,a}(\widetilde{Y}) - \Gamma_{y,a}(\widetilde{Y})| \ = \\
&|\widetilde{p}_{0,a}(1-\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=a)) + \widetilde{p}_{1,a}\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=a) \\
&-\widetilde{p}_{0,0}(1-\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=0)) - \widetilde{p}_{1,0}\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=0) \\
&-\widetilde{p}_{0,a}(1-\widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=a)) - \widetilde{p}_{1,a}\widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=a) \\
&+\widetilde{p}_{0,0}(1-\widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=0)) + \widetilde{p}_{1,0}\widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=0)| \\
&\leq |\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=a) - \widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=a)| \cdot |\widetilde{p}_{1,a} - \widetilde{p}_{0,a}| \\
&+|\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=0) - \widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=0)| \cdot |\widetilde{p}_{1,0} - \widetilde{p}_{0,0}| \\
&\leq |\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=a) - \widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=a)| \\
&+|\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=0) - \widetilde{\mathbb{P}}(\hat{Y}=1|Y=y, A=0)|
\end{aligned}$$

From the proof of Lemma 1, let $C = \frac{\pi+|\mathcal{A}|-2}{|\mathcal{A}|\pi-1}$, then if $n \geq \frac{32\log(8|\mathcal{A}|/\delta)}{\min_{ya}\mathbf{P}_{ya}}$, we have:

$$\mathbb{P}\left(\max_{ya}|\widetilde{\Gamma}^S_{ya} - \Gamma_{ya}| > \sqrt{\frac{\log(64/\delta)}{2n}\frac{4|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya}}}\right) \leq \delta/4$$

Now if $\widetilde{\alpha}_n \geq \sqrt{\frac{\log(64/\delta)}{2n}\frac{4|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya}}}$, then by the same argument of Lemma 5 in Woodworth et al. (2017), any 0-discriminatory $\widetilde{Y}^*$ derived from $\hat{Y}$ is in the feasible set of step 2 with probability greater than $1-\delta/4$, hence by the optimality of $\widetilde{Y}$ on S_2:

$$\text{err}(\widetilde{Y}) \leq_{\delta/2} \text{err}(\widetilde{Y}^*) + \frac{4|\mathcal{A}|(\pi+|\mathcal{A}|-2)}{|\mathcal{A}|\pi-1}\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}$$

□

We are now ready for the proof of Theorem 1.

Theorem 1 *For any hypothesis class $\mathcal{H}$, any distribution over (X, A, Y) and any $\delta \in (0, 1/2)$, then with probability $1-\delta$,* if $n \geq \frac{16\log(8|\mathcal{A}|/\delta)}{\min_{ya}\mathbf{P}_{ya}}$, $\alpha_n = \sqrt{\frac{8\log 64/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$ and $\widetilde{\alpha}_n = \sqrt{\frac{\log(64/\delta)}{2n}\frac{4|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya}}}$ then the predictor resulting from the two-step procedure satisfies:

$$\text{err}(\widetilde{Y}) \leq_\delta \text{err}(Y^*) + \frac{5C}{\min_{ya}\mathbf{P}^2_{ya}}\left(\frac{2}{B} + 10\mathfrak{R}_{\frac{\min_{ya} n\mathbf{P}_{ya}}{4}}(\mathcal{H}) + 18|\mathcal{A}|\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}\right)$$

$$\text{disc}(\widetilde{Y}) \leq_\delta \sqrt{\frac{\log(\frac{64}{\delta})}{2n}\frac{8|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya}}}$$

Proof. Since the predictor obtained in step 1 is only a function of X, then the guarantees of step 2 immediately apply by Lemma 9:

$$\text{err}(\widetilde{Y}) \leq_{\delta/2} \text{err}(\widetilde{Y}^*) + 4|\mathcal{A}|C\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}$$

$$\text{disc}(\widetilde{Y}) \leq_{\delta/2} \sqrt{\frac{\log(\frac{64}{\delta})}{2n}\frac{8|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya}}}$$

Now we have to relate the loss of the optimal derived predictor from $\hat{Y}$, denoted by $\widetilde{Y}^*$, to the loss of the optimal non-discriminatory predictor in $\mathcal{H}$. We can apply Lemma 4 in Woodworth et al. (2017) as the solution of our derived LP is in expectation equal to that in terms of A. Lemma 4 in Woodworth et al. (2017) tells us that the optimal derived predictor has a loss that is less or equal than the sum of the loss of the base predictor and its discrimination:

$$\mathrm{err}(\widetilde{Y}^*) \leq \mathrm{err}(\hat{Y}) + \mathrm{disc}(\hat{Y}) \tag{20}$$

We have then by Lemma 9 the loss of the optimal derived predictor:

$$\begin{aligned}\mathrm{err}(\widetilde{Y}^*) &\leq_\delta \mathrm{err}(Y^*) + 4\sqrt{\frac{\log 1/\delta}{n}} + 4\mathfrak{R}_{n/2}(\mathcal{H}) + \frac{5C}{\min_{ya}\mathbf{P}_{ya}^2}\left(\frac{2}{B} + 6\mathfrak{R}_{\frac{\min_{ya} n\mathbf{P}_{ya}}{4}}(\mathcal{H}) + 10\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}\right)\\ &\leq_\delta \mathrm{err}(Y^*) + \frac{5C}{\min_{ya}\mathbf{P}_{ya}^2}\left(\frac{2}{B} + 10\mathfrak{R}_{\frac{\min_{ya} n\mathbf{P}_{ya}}{4}}(\mathcal{H}) + 14\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}\right)\end{aligned}$$

Hence our derived predictor satisfies:

$$\begin{aligned}\mathrm{err}(\widetilde{Y}) &\leq_{\delta/2} \mathrm{err}(\widetilde{Y}^*) + 4|\mathcal{A}|C\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}\\ &\leq_\delta \mathrm{err}(Y^*) + \frac{5C}{\min_{ya}\mathbf{P}_{ya}^2}\left(\frac{2}{B} + 10\mathfrak{R}_{\frac{\min_{ya} n\mathbf{P}_{ya}}{4}}(\mathcal{H}) + 18|\mathcal{A}|\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}\right)\end{aligned}$$

□

A.3. Section 6

Lemma 3 *Given a hypothesis class $\mathcal{H}$, a distribution over (X, A, Y), $B \in \mathbb{R}^+$ and any $\delta \in (0, 1/2)$, then with probability greater than $1-\delta$, if $n_\ell \geq \frac{8\log 4|\mathcal{A}|/\delta}{\min_{ya}\mathbf{P}_{ya}}$, $\alpha_n = 2\sqrt{\frac{\log 32|\mathcal{A}|/\delta}{n_\ell\min_{ya}\mathbf{P}_{ya}}}$ and we let $\vartheta = \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log 4/\delta}{n}}$, then running Algorithm 1 on data set S with $T \geq \frac{16\log(4|\mathcal{A}|+1)}{\vartheta^2}$ and learning rate $\eta = \frac{\vartheta}{8B}$ returns a predictor $\hat{Y}$ satisfying the following:*

$$\begin{aligned}\mathrm{err}(\hat{Y}) &\leq_\delta \mathrm{err}(Y^*) + 4\mathfrak{R}_n(\mathcal{H}) + 4\sqrt{\frac{\log 4/\delta}{n}}\\ disc(\hat{Y}) &\leq_\delta \frac{2}{B} + 6\mathfrak{R}_{\frac{\min_{ya} n_\ell\mathbf{P}_{ya}}{2}}(\mathcal{H}) + 10\sqrt{\frac{2\log 32|\mathcal{A}|/\delta}{n_\ell\min_{ya}\mathbf{P}_{ya}}}\end{aligned}$$

Proof. The proof follows immediately from Lemma 2 with the identical error bound and replacing n by n_l in the discrimination bound. The two dataset langragian does not impact Theorem 1 in (Agarwal et al., 2018) and the definition of an approximate saddle point remains the same as both players have the same objective. □

Lemma 4 *Let $S = \{(x_i, a_i, y_i)\}_{i=1}^n$ i.i.d. $\sim \mathbb{P}^n(A, X, Y)$, the estimator $\widetilde{\gamma}_{ya}^S$ is consistent. As $n \to \infty$*

$$\widetilde{\gamma}_{ya}^S \to_p \gamma_{ya}.$$

Proof.

$$
\begin{aligned}
\widetilde{\gamma}_{ya}(\hat{Y}) &= \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{Y}(x_i)\mathbf{1}(y_i=y)\mathbb{P}(A=a|x_i,y_i)}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(y_i=y)\mathbb{P}(A=a|x_i,y_i)} \\
&\to \frac{\mathbb{E}[\hat{Y}(X)\mathbb{I}(Y=y)\mathbb{P}(A=a|X,Y)]}{\mathbb{E}[\mathbb{I}(Y=y)\mathbb{P}(A=a|X,Y)]} \\
&= \frac{\mathbb{E}[\hat{Y}(X)\mathbb{I}(Y=y)\mathbb{P}(A=a|X,Y)]}{\int_x \mathbb{P}(X=x,Y=y)\mathbb{P}(A=a|X=x,Y=y)dx} \\
&= \frac{\int_x \mathbb{P}(X=x,Y=y)\hat{Y}(x)\mathbb{P}(A=a|X=x,Y=y)dx}{\mathbb{P}(Y=y,A=a)} \\
&= \frac{\int_x \mathbb{P}(X=x|Y=y,A=a)\mathbb{P}(Y=y,A=a)\hat{Y}(x)dx}{\mathbb{P}(Y=y,A=a)} \\
&= \mathbb{E}_{X|Y=y,A=a}\hat{Y}(X) = \mathbb{P}(\hat{Y}=1|Y=y,A=a) = \gamma_{ya}
\end{aligned}
$$

□

References

Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., and Omer, S. B. Using the bayesian improved surname geocoding method (bisg) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. *Health services research*, 49(1):268–283, 2014.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Alabi, D. The cost of a reductions approach to private fair optimization. *arXiv preprint arXiv:1906.09613*, 2019.

Awasthi, P., Kleindessner, M., and Morgenstern, J. Effectiveness of equalized odds for fair classification under imperfect group information. *arXiv preprint arXiv:1906.03284*, 2019.

Bagdasaryan, E. and Shmatikov, V. Differential privacy has disparate impact on model accuracy. *arXiv preprint arXiv:1905.12101*, 2019.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Bureau, C. F. P. Using publicly available information to proxy for unidentified race and ethnicity, June 2014. URL files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348. ACM, 2019.

Commission, F. T. Your equal credit opportunity rights, January 2013. URL www.consumer.ftc.gov/articles/0347-your-equal-credit-opportunity-rights.

Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. *arXiv preprint arXiv:1807.00028*, 2018.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. On the compatibility of privacy and fairness. ., 2019.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279. ACM, 2008.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Fiscella, K. and Fremont, A. M. Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research*, 41(4p1):1482–1500, 2006.

Freund, Y. and Schapire, R. E. Game theory, on-line prediction and boosting. In *COLT*, volume 96, pp. 325–332. Citeseer, 1996.

Gupta, M., Cotter, A., Fard, M. M., and Wang, S. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1934–1943, 2018.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. Differentially private fair learning. *arXiv preprint arXiv:1812.02696*, 2018.

Kairouz, P., Oh, S., and Viswanath, P. Extremal mechanisms for local differential privacy. In *Advances in neural information processing systems*, pp. 2879–2887, 2014.

Kairouz, P., Bonawitz, K., and Ramage, D. Discrete distribution estimation under local privacy. *arXiv preprint arXiv:1602.07387*, 2016.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019.

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., and Weller, A. Blind justice: Fairness with encrypted sensitive attributes. *arXiv preprint arXiv:1806.03281*, 2018.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996.

Lamy, A. L., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. *arXiv preprint arXiv:1901.10837*, 2019.

McDiarmid, C. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Vigdor, N. Apple card investigated after gender discrimination complaints, November 2019. URL `https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html`.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M. I. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pp. 1920–1953, 2017.

Xu, D., Yuan, S., and Wu, X. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 594–599. ACM, 2019.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.