arXiv:2002.11651v1 [cs.LG] 26 Feb 2020

Fair Learning with Private Demographic Data

Hussein Mozannar [*]　Mesrob I. Ohannessian [†]　Nathan Srebro [‡]

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 Introduction

As algorithmic systems driven by machine learning start to play an increasingly important role in society, concerns arise over their compliance with laws, regulations and societal norms. In particular, machine learning systems have been found to be discriminating against certain demographic groups in applications of criminal assessment, lending and facial recognition (Barocas et al. (2019)). To ensure non-discrimination in learning tasks, knowledge of the sensitive attributes is essential, however, laws and regulation often prohibit access and use of this sensitive data. As an example, credit card companies do not have the right to ask about an individual's race when applying for credit, while at the same time they have to prove that their decisions are non-discriminatory (Commission (2013); Chen et al. (2019)).

Apple Card, a credit card offered by Apple and Goldman Sachs, was recently accused of being discriminatory Vigdor (2019). Married couples rushed to Twitter to report that there were significant differences in the credit limit given individually to each of them even though they had shared finances and similar income levels. Supposing Apple was trying to make sure its learned model was non discriminatory, it would have been forced to use proxies for gender and recent work has shown that proxies can be problematic Kallus et al. (2019). We are then faced with what seems to be two opposing societal notions to satisfy: we want our system to be non-discriminatory while maintaining the privacy of our sensitive attributes. Note that even if the features that our model uses are independent of the sensitive attributes, it is not enough to guarantee notions of non-discrimination that further condition on the truth, e.g. equalized odds. One potential workaround to this problem, is to allow the individuals to release their data in a locally differentially private manner (Dwork et al. (2006)) and then try to learn from this privatized data a non-discriminatory predictor. This allows us to guarantee that our decisions are fair while maintaining a degree of individual privacy to each user. Related work are briefly surveyed in Section 2

[*] Massachusetts Institute of Technology. Email: mozannar@mit.edu

[†] University of Illinois at Chicago. Email: mesrob@uic.edu

[‡] Toyota Technological Institute at Chicago. Email: nati@ttic.edu

In this work, we consider a binary classification framework where we have access to non-sensitive features X and locally-private versions of the sensitive attributes A denoted by Z. The details of the problem formulation are given in Section 3. Our contributions are as follows:

- We first give sufficient conditions on our predictor for non-discrimination to be equivalent under A and Z and derive estimators to measure discrimination using the private attributes Z. (Section 4)
- We give a learning algorithm based on the two-step procedure of Woodworth et al. (2017) and provide statistical guarantees for both the error and discrimination of the resulting predictor. The main innovation in terms of both the algorithm and its analysis is in accessing properties of the sensitive attribute A by carefully inverting the sample statistics of the private attributes Z. (Section 5)
- We highlight how some of the same approach can handle other forms of deficiency in demographic information, by giving an auditing algorithm with guarantees, when protected attributes are available only for a subset of the data. (Section 6)

Beyond the original motivation, this work conveys additional insight on the subtle trade-offs between error and discrimination. In this perspective, privacy is not in itself a requirement, but rather an analytic tool. We give some experimental illustrations of these trade-offs.

2 Related Work

Enforcing non-discrimination constraints in supervised learning has been extensively explored with many algorithms proposed to learn fair predictors with methods that fall generally in one category among pre-processing (Zemel et al. (2013)), in-processing (Cotter et al. (2018); Agarwal et al. (2018)), or post-processing (Hardt et al. (2016)). In this work we focus on group-wise statistical notions of discrimination, setting aside critical concerns of individual fairness (Dwork et al. (2012)).

Kilbertus et al. (2018) were the first to propose to learn a fair predictor without disclosing information about protected attributes, using secure multi-party computation (MPC). However, as Jagielski et al. (2018) noted, MPC does not guarantee that the predictor cannot leak individual information. In response, Jagielski et al. (2018) proposed differentially private (DP) (Dwork et al. (2006)) variants of fair learning algorithms. More recent work have similarly explored learning fair and DP predictors (Cummings et al. (2019); Xu et al. (2019); Alabi (2019); Bagdasaryan and Shmatikov (2019)). In our setting *local* privacy maintains all the guarantees of DP in addition to not allowing the learner to know for certain any sensitive information about a particular data point. Related work has also considered fair learning when the protected attribute is missing or noisy (Hashimoto et al. (2018); Gupta et al. (2018); Lamy et al. (2019); Awasthi et al. (2019); Kallus et al. (2019)).

Among these, the most related setting is that of Lamy et al. (2019), but it has several critical contrasting points with the present work. The simplest difference is the generalization here to non-binary groups, and the corresponding precise characterization of the equivalence between exact non-discrimination with respect to the original and private attributes. More importantly, their approach is only the *first* step of our algorithm. As we show in Lemma 2, the first step makes the non-discrimination guarantee depend on both the privacy level and the complexity of the hypothesis class, which could be very costly. We remedy this using the *second* step of our algorithm. Awasthi et al. (2019) consider a more general noise model for the protected attributes in the training data, but assume access to the actual protected attributes at test time. The fact that at test time A is provided guarantees that the predictor is not a function of Z and hence for the LDP noise mechanism by Proposition 1, we know that it is enough to guarantee non-discrimination with respect to Z to be non-discriminatory with respect to A, which considerably simplifies the problem.

3 Problem Formulation

A predictor $\hat{Y}$ of a binary target $Y \in \{0,1\}$ is a function of non-sensitive attributes $X \in \mathcal{X}$ and possibly also of a sensitive (or protected) attribute $A \in \mathcal{A}$ denoted as $\hat{Y} := h(X)$ or $\hat{Y} := h(X, A)$. We consider a binary classification task where the goal is to learn such a predictor, while ensuring a specified notion of non-discrimination with respect to A. As an example, when deciding to extend credit to a given individual, the protected attribute could denote someone's race and sex and the features X could contain the person's financial history, level of education and housing information. Note that X could very well include proxies for A such as zip code which could reliably infer race (Bureau (2014)).

Our focus here is on statistical notions of group-wise non-discrimination amongst which are the following:

Definition 1 (Fairness Definitions)**.** A classifier $\hat{Y}$ satisfies:

- Equalized odds (EO) if $\forall a \in \mathcal{A}$

$$\mathbb{P}(\hat{Y} = 1 | A = a, Y = y) = \mathbb{P}(\hat{Y} = 1 | Y = y) \;\; \forall y \in \{0,1\},$$

- Demographic parity (DP) if $\forall a \in \mathcal{A}$

$$\mathbb{P}(\hat{Y} = 1 | A = a) = \mathbb{P}(\hat{Y} = 1),$$

- Accuracy parity (AP) if $\forall a \in \mathcal{A}$

$$\mathbb{P}(\hat{Y} \neq Y | A = a) = \mathbb{P}(\hat{Y} \neq Y),$$

- False discovery ($\hat{y} = 1$) / omission ($\hat{y} = 0$) rates parity if $\forall a \in \mathcal{A}$

$$\mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}, A = a) = \mathbb{P}(\hat{Y} \neq Y | \hat{Y} = \hat{y}).$$

Our treatment extends to a very broad family of demographic fairness constraints. Additionally, one can naturally define approximate versions of the above fairness constraints. As an example, for the notion of equalized odds, let $\gamma_{y,a}(\hat{Y}) = \mathbb{P}(\hat{Y} = 1 | Y = y, A = a)$, then $\hat{Y}$ satisfies α-EO if:

$$\max_{y \in \{0,1\}, a \in \mathcal{A}} \Gamma_{ya} := \left| \gamma_{y,a}(\hat{Y}) - \gamma_{y,0}(\hat{Y}) \right| \leq \alpha$$

While it is clear that learning or auditing fair predictors requires knowledge of the protected attributes, laws and regulations often restrict the use and the collection of this data (Jagielski et al. (2018)). Moreover, even if there are no restrictions on the usage of the protected attribute, it is desirable that this information is not leaked by (1) the algorithm's output and (2) the data collected. Local differential privacy (LDP) guarantees that the entity holding the data does not know for certain the protected attribute of any data point, which in turn makes sure that any algorithm built on this data is differentially private. Formally a locally $\epsilon-$differentially private mechanism Q is defined as follows:

Definition 2. Q is $\epsilon-$differentially private if (Duchi et al. (2013)):

$$\max_{z,a,a'} \frac{Q(Z=z|a)}{Q\left(Z=z|a'\right)} \leq e^{\epsilon}$$

The mechanism we employ is the randomized response mechanism (Warner (1965); Kairouz et al. (2014)):

$$Q(z|a) = \begin{cases} \frac{e^{\varepsilon}}{|\mathcal{A}|-1+e^{\varepsilon}} := \pi & \text{if } z = a \\ \frac{1}{|\mathcal{A}|-1+e^{\varepsilon}} := \bar{\pi} & \text{if } z \neq a \end{cases}$$

The choice of the randomized response mechanism is motivated by its optimality for distribution estimation under LDP constraints Kairouz et al. (2014, 2016)

The hope is that LDP samples of A are sufficient to ensure non-discrimination, allowing us to refrain from the problematic use proxies for A. For the remainder of this paper, we assume that we have access to n samples $S = \{(x_i, y_i, z_i)\}_{i=1}^{n}$ which are the result of an *i.i.d* draw from an unknown distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{A}$ where $\mathcal{A} = \{0, 1, \cdots, |\mathcal{A}| - 1\}$ and $\mathcal{Y} = \{0, 1\}$, but where A is not observed and instead Z is sampled from $Q(.|A)$ independently from X and Y. We call Z the *privatized protected attribute*. To emphasize the difference between A and Z with respect to fairness, let $q_{y,a}(\hat{Y}) = \mathbb{P}(\hat{Y} = 1|Y = y, Z = a)$, note that $\hat{Y}$ satisfies α-EO *with respect to* Z if:

$$\max_{y \in \{0,1\}, a \in \mathcal{Z}} \left| q_{y,a}(\hat{Y}) - q_{y,0}(\hat{Y}) \right| \leq \alpha.$$

4 Auditing for Discrimination

The two main questions we answer in this section is whether non-discrimination with respect to A and Z are equivalent and how to estimate the non-discrimination of a given predictor.

First, note that if a certain predictor $\hat{Y} = h(X, Z)$ uses Z for predictions and is non-discriminatory with respect to Z, then it is possible for it to in fact be discriminatory with respect to A. In Appendix A, we give an explicit example of such a predictor, that violates the equivalence for EO. This illustrates that naïve implementations of fair learning methods can be more discriminatory than perceived. Any method that naïvely uses the attribute Z for its final predictions cannot guarantee any level of non-discrimination with respect to A especially post-processing methods.

This however is not the case when predictors do not avail themselves of the privatized protected attribute Z. Namely, let's consider $\hat{Y}$ that are only a function of X. Since the randomness in the privatization mechanism is independent of X, this implies in particular that $\hat{Y}$ is independent of Z given A. Our first result is that exact non-discrimination is invariant under local privacy:

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 Sketch. We consider 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 define non-discrimination with respect to A as having:

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

Define this notion similarly with respect to Z. We can obtain the following relation for the conditional probabilities

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

Let P be the following $|\mathcal{A}| \times |\mathcal{A}|$ matrix:

$$\begin{cases} P_{i,i} = \frac{\pi \mathbb{P}(A=i, \mathcal{E}_2)}{\mathbb{P}(Z=i, \mathcal{E}_2)} \text{ for } i \in \mathcal{A} \\ P_{i,j} = \frac{\bar{\pi} \mathbb{P}(A=j, \mathcal{E}_2)}{\mathbb{P}(Z=i, \mathcal{E}_2)} \text{ for } i, j \in \mathcal{A} \text{ s.t.} i \neq j \end{cases} \tag{1}$$

Then we have the following linear system of equations:

$$\begin{bmatrix} \mathbb{P}(\mathcal{E}_1 | \mathcal{E}_2, Z = 0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1 | \mathcal{E}_2, Z = |\mathcal{A}| - 1) \end{bmatrix} = P \begin{bmatrix} \mathbb{P}(\mathcal{E}_1 | \mathcal{E}_2, A = 0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1 | \mathcal{E}_2, A = |\mathcal{A}| - 1) \end{bmatrix} \tag{2}$$

The matrix P is row-stochastic and invertible, from this linear system we can deduce that non-discrimination with respect to Z and A are equivalent; details are left to Appendix A. □

We next study how to measure non-discrimination from samples. Unfortunately, Proposition 1 applies only in the population-limit. For example for the notion EO, despite what it seems to suggest, naïve sample α-discrimination relative to Z underestimates discrimination relative to A. Interestingly however, for any of the considered fairness notions, we can recover the statistics of the population with respect to A via a linear system of equations relating them to those of Z as in (2). This is done by inverting the matrix P defined in (1), however more care is needed: to compute the matrix P one needs to compute quantities involving the attribute A, which then all have to be related back to Z. Using this relation, we derive an estimator for the discrimination of a predictor that does not suffer from the bias of the naïve approach. First we set key notations for the rest of the paper: $\mathbf{P}_{ya} := \mathbb{P}(Y = y, A = a)$, $\mathbf{Q}_{ya} := \mathbb{P}(Y = y, Z = a)$ and $C = \frac{|\mathcal{A}| - 2 + e^{\epsilon}}{e^{\epsilon} - 1}$. The latter captures the scale of privatization: $C \approx O(\epsilon^{-1})$ if $\epsilon \ll 1$.

Let P be the $\mathcal{A} \times \mathcal{A}$ matrix as such: $\begin{cases} P_{i,i} = \pi \frac{\mathbf{P}_{yi}}{\mathbf{Q}_{yi}} \text{ for } i \in \mathcal{A} \\ P_{i,j} = \bar{\pi} \frac{\mathbf{P}_{yj}}{\mathbf{Q}_{yi}} \text{ for } i, j \in \mathcal{A} \text{ s.t.} i \neq j \end{cases}$

Then one can relate $q_{y,.}$ and $\gamma_{y,.}$ via:

$$\begin{bmatrix} q_{y0} \\ \vdots \\ q_{y,|\mathcal{A}|-1} \end{bmatrix} = P \begin{bmatrix} \gamma_{y,0} \\ \vdots \\ \gamma_{y,|\mathcal{A}|-1} \end{bmatrix}$$

And thus by inverting P we can recover $\gamma_{y,a}$, however, the matrix P involves estimating the probabilities $\mathbb{P}(Y=y, A=a)$ which we do not have access to but can similarly recover by noting that: $\mathbf{Q}_{yz} = \pi \mathbf{P}_{yz} + \sum_{a \neq z} \bar{\pi} \mathbf{P}_{ya}$

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 $i \neq j$. Therefore $\Pi_k^{-1} \mathbf{Q}_{y,\cdot} = \mathbb{P}(Y = y, A = k)$ where Π_k^{-1} is the k'th row of Π^{-1}. Hence we can plug this estimates to compute P and invert the linear system to measure our discrimination In Lemma 1, we characterize the sample complexity needed by our estimator to bound the violation in discrimination, specifically for the EO constraint. The privacy penalty C arises from $||P||_\infty$.

Lemma 1. *For any $\delta \in (0, 1/2)$, any binary predictor $\hat{Y} := h(X)$, denote by $\widetilde{\Gamma}^S_{ya}$ our proposed estimator for Γ_{ya} based on S, if $n \geq \frac{8 \log(8|\mathcal{A}|/\delta)}{\min_{ya} \mathbf{P}_{ya}}$, we have:*

$$\mathbb{P}\left(\max_{ya} |\widetilde{\Gamma}^S_{ya} - \Gamma_{ya}| > \sqrt{\frac{\log(16/\delta)}{2n}} \frac{4C^2}{\min_{ya} \mathbf{P}^2_{ya}} \right) \leq \delta \tag{3}$$

5 Learning Fair Predictors

In this section, we give a strategy to learn a non-discriminatory predictor with respect to A from the data S, which only contains the privatized attribute Z. As in Lemma 1, for concreteness and clarity we restrict the analysis to the notion of equalized odds (EO) — most of the analysis extends directly to other constraints. In light of the limitation identified by Proposition 1, let $\mathcal{H}$ be a hypothesis class of functions that depend only on X. Instead of a single predictor in the class, we exhibit a distribution over hypotheses, which we interpret as a randomized predictor. Let $\Delta_{\mathcal{H}}$ be the set of all distributions over $\mathcal{H}$, and denote such a randomized predictor by $Q \in \Delta_{\mathcal{H}}$. The goal is to learn a predictor that approximates the performance of the optimal non-discriminatory distribution:

$$Y^* = \arg \min_{Q \in \Delta_{\mathcal{H}}} \mathbb{P}(Q(X) \neq Y) \tag{4}$$

$$s.t. \quad \gamma_{y,a}(Q) = \gamma_{y,0}(Q) \; \forall y \in \{0,1\}, \forall a \in \mathcal{A} \tag{5}$$

A first natural approach would be to treat the private attribute Z as if it were A and ensure on S that the learned predictor is non-discriminatory. Since the hypothesis class $\mathcal{H}$ consists of functions that depend only on X, Proposition 1 applies and offers hope that, if we are able to achieve exact non discriminatory with respect to Z, we would be in fact non-discriminatory with respect to A. There are two problems with the above approach. First, exact non-discrimination is computationally hard to achieve and approximate non-discrimination underestimates the discrimination by the privacy penalty C. And second, using an in-processing method such as the reductions approach of Agarwal et al. (2018) to learn results in a discrimination guarantee that scales with the complexity of $\mathcal{H}$.

Our approach is to adapt the two-step procedure of Woodworth et al. (2017) to our setting. We start by dividing our data set S into two equal parts S_1 and S_2. The first step is to learn an approximately non-discriminatory predictor $\hat{Y} = Q(X)$ with respect to Z on S_1 via the reductions approach of Agarwal et al. (2018). This predictor has low error, but may be highly discriminatory due to the complexity of the class affecting the generalization of non-discrimination of $\hat{Y}$. The aim of the second step is to produce a final predictor $\widetilde{Y}$ that corrects for this discrimination, without increasing its error much. We modify the post-processing procedure of Hardt et al. (2016) to give us non-discrimination with respect to A directly for the derived predictor $\widetilde{Y} = f(\hat{Y}, Z)$. Two relationships link the first step to the second: how discrimination with respect to Z and with respect to A relate and how the discrimination from the first step affects the error of the derived predictor. In the following subsections we describe each of the steps, along with the statistical guarantees on their performance.

5.1 Step 1: Approximate Non-Discrimination with respect to Z

The first step aims to learn a predictor $\hat{Y}$ that is approximately α_n-discriminatory with respect to Z defined as:

$$\hat{Y} = \arg \min_{Q \in \Delta_{\mathcal{H}}} \mathrm{err}^{S_1}(Q(X)) \tag{6}$$

$$\text{s.t.} \max_{y \in \{0,1\}} |q^{S_1}_{y,a}(Q) - q^{S_1}_{y,a}(Q)| \leq \alpha_n \tag{7}$$

where for $Q \in \Delta_{\mathcal{H}}$, we use the shorthand $\mathrm{err}(Q) = \mathbb{P}(Q(X) \neq Y)$ and quantities with a superscript S_1 indicate their empirical counterparts. To solve the optimization problem defined in (6), we reduce the constrained optimization problem to a weighted unconstrained problem, following the approach of Agarwal et al. (2018). As is typical with the family of fairness criteria considered, the constraint in (7) can be rewritten as a linear constraint on $\hat{Y}$ explicitly. Let $\mathcal{J} = \mathcal{Y} \times \mathcal{A}$, $\mathcal{K} = \mathcal{Y} \times \mathcal{A} \setminus \{0\} \times \{-,+\}$ and define $\boldsymbol{\gamma}(Q) \in \mathbb{R}^{|\mathcal{J}|}$ with $\boldsymbol{\gamma}(Q)_{(y,a)} = \gamma_{y,a}(Q)$, with the matrix $M \in \mathbb{R}^{|\mathcal{K}| \times |\mathcal{J}|}$ having entries: $M_{(y,a,+),(a',y')} = \mathbb{I}(a = a', y = y'), M_{(y,a,-),(a',y')} = -\mathbb{I}(a = a', y = y'), M_{(y,a,+),(0,y')} = \mathbb{I}(y = y'), M_{(y,a,-),(0,y')} = -\mathbb{I}(y = y')$. With this reparametrization, we can write α_n-EO as:

$$M\boldsymbol{\gamma}(Q) \leq \alpha_n \mathbf{1} \tag{8}$$

Let us introduce the Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{K}|}_{+}$ and define the Lagrangian:

$$L(Q, \lambda) = \mathrm{err}(Q) + \lambda^{\top}(M\boldsymbol{\gamma}(Q) - \alpha \mathbf{1}) \tag{9}$$

We constrain the norm of λ with $B \in \mathbb{R}^{+}$ and consider the following two dual problems:

$$\min_{Q \in \Delta_{\mathcal{H}}} \max_{\boldsymbol{\lambda} \in \mathcal{R}^{|\mathcal{K}|}_{+}, ||\boldsymbol{\lambda}||_1 \leq B} L(Q, \boldsymbol{\lambda}) \tag{10}$$

$$\max_{\boldsymbol{\lambda} \in \mathcal{R}^{|\mathcal{K}|}_{+}, ||\boldsymbol{\lambda}||_1 \leq B} \min_{Q \in \Delta_{\mathcal{H}}} L(Q, \boldsymbol{\lambda}) \tag{11}$$

Note that L is linear in both Q and $\boldsymbol{\lambda}$ and their domains are convex and compact, hence the respective solution of both problems form a saddle point of L (Agarwal et al. (2018)). To find the saddle point, we treat our problem as a zero-sum game between two players: the Q-player "learner" and the λ-player "auditor" and use the method of Freund and Schapire (1996). The auditor follows the exponentiated gradient algorithm and the learner picks his best response. The approach is fully described in Algorithm 1.

Algorithm 1: Exp. gradient reduction for fair classification Agarwal et al. (2018)

```
Input: training data $(X_i, Y_i, Z_i)_{i=1}^{n/2}$, bound $B$, learning rate $\eta$, rounds $T$
$\boldsymbol{\theta}_1 \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{K}|}$
for $t = 1, 2, \cdots, T$ do
    $\lambda_{t,k} \leftarrow B \frac{\exp(\theta_{t,k})}{1+\sum_{k'} \exp(\theta_{t,k})} \forall k \in \mathcal{K}$
    $h_t \leftarrow \text{BEST}_h(\boldsymbol{\lambda}_t)$
    $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \eta(M\boldsymbol{\gamma}^S(h_t) - \alpha_n \mathbf{1})$
end
$\hat{Y} \leftarrow \frac{1}{T}\sum_{t=1}^{T} h_t, \hat{\boldsymbol{\lambda}} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\lambda}_t$
Return $(\hat{Y}, \hat{\boldsymbol{\lambda}})$
```

Faced with a given vector $\boldsymbol{\lambda}$ the learner's best response, $\text{BEST}_h(\boldsymbol{\lambda})$), puts all the mass on a single predictor $h \in \mathcal{H}$ as the Lagrangian 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 reestablish the reduction in detail in Appendix A, as there are slight differences with our setup. In particular, in Lemma 2, we establish a generalization bound on the error of the first step predictor $\hat{Y}$ and on its discrimination, defined as the maximum violation in the EO constraint. To denote the latter similarly to the error, we use the shorthand $\text{disc}(\hat{Y}) = \max_{y\in\{0,1\}, a\in\mathcal{A}} \Gamma_{ya}$.

Lemma 2. *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 $n \geq \frac{16 \log 8|\mathcal{A}|/\delta}{\min_{ya} \mathbf{P}_{ya}}$, $\alpha_n = 2\sqrt{\frac{\log 64|\mathcal{A}|/\delta}{n \min_{ya} \mathbf{P}_{ya}}}$ and we let $\vartheta = \mathfrak{R}_{n/2}(\mathcal{H}) + \sqrt{\frac{\log 8/\delta}{n}}$, then running Algorithm 1 on 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:*

$$err(\hat{Y}) \leq_{\delta/2} err(Y^*) + 4\mathfrak{R}_{n/2}(\mathcal{H}) + 4\sqrt{\frac{\log 8/\delta}{n}} \tag{12}$$

$$disc(\hat{Y}) \leq_{\delta/2} \frac{5C}{\min_{ya} \mathbf{P}_{ya}^2} \left(\frac{2}{B} + 6\mathfrak{R}_{\frac{\min_{ya} n\mathbf{P}_{ya}}{4}}(\mathcal{H}) + 10\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n \min_{ya} \mathbf{P}_{ya}}} \right) \tag{13}$$

Proof of Lemma 3 can be found in Appendix A. Note that the error bound in Lemma 2 does not scale with the privacy level, however the discrimination bound is not only hit by the privacy, through C, but is further multiplied by the Rademacher complexity of $\mathcal{H}$. Our goal in the next step is to reduce the sample complexity required to achieve low discrimination by removing the dependence on the complexity of the model class in the discrimination bound.

Jagielski et al. (2018) modifies Algorithm 1 to ensure that the model is differentially private with respect to A assuming access to data with the non-private attribute A. The error and discrimination generalization bounds obtained by Jagielski et al. (2018) both scale with the privacy level ϵ and the complexity of $\mathcal{H}$ as opposed to our error bound that is privacy independent, hence surprisingly LDP can give better error guarantees than DP.

5.2 Step 2: Post-hoc correction to achieve non-discrimination for A

We correct the predictor we learned in step 1 using a modified version of the post-processing procedure of Hardt et al. (2016) on the data set S_2. The derived second step predictor $\widetilde{Y}$ is fully characterized by $2|\mathcal{A}|$ probabilities $\mathbb{P}(\widetilde{Y}=1|\hat{Y}=\hat{y}, Z=a) := p_{\hat{y},z}$. If we naïvely derive the predictor applying the post-processing procedure of Hardt et al. (2016) on S_2 then this *does not* imply that the predictor satisfies EO as the derived predictor is an explicit function of Z, cf. the discussion in Section 4. Our approach is to directly ensure non-discrimination with respect to A and hence achieve our goal. Two facts make this possible. First, the base predictor of step 1 is not a function of Z and hence we can measure its false negative and positive rates using the estimator from Lemma 1. And second, to compute these rates for $\widetilde{Y}$, we can exploit its special structure. In particular, note the following decomposition:

$$\begin{aligned}\mathbb{P}(\widetilde{Y}=1|Y=y, A=a) &= \mathbb{P}(\widetilde{Y}=1|\hat{Y}=0, A=a)\mathbb{P}(\hat{Y}=0|Y=y, A=a)\\ &\quad + \mathbb{P}(\widetilde{Y}=1|\hat{Y}=1, A=a)\mathbb{P}(\hat{Y}=1|Y=y, A=a)\end{aligned} \tag{14}$$

we have that:

$$\mathbb{P}(\widetilde{Y}=1|\hat{Y}=\hat{y}, A=a) = \pi p_{\hat{y},a} + \bar{\pi} \sum_{a' \in \mathcal{A}\backslash a} p_{\hat{y},a'} := \widetilde{p}_{\hat{y},a}$$

and $\mathbb{P}(\hat{Y}|Y=y, A=a)$ can be recovered by Lemma 1, denote $\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=\hat{y}|Y=y, A=a)$ our estimator based on the empirical $\mathbb{P}^{S_2}(\hat{Y}|Y,Z)$. Therefore we can compute sample versions of the conditional probabilities (14).

Our modified post-hoc correction reduces to solving the following constrained linear program:

$$\begin{aligned}\widetilde{Y} = \arg\min_{p_{\cdot,\cdot}} \quad & \sum_{\hat{y},a} \Big(\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=\hat{y}, Z=a, Y=0) \; - \widetilde{\mathbb{P}}^{S_2}(\hat{Y}=\hat{y}, Z=a, Y=1))\Big) \cdot \widetilde{p}_{\hat{y},a}\\ s.t. \quad & \Big| \widetilde{p}_{0,a}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=0|Y=y, A=a) + \widetilde{p}_{1,a}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=1|Y=y, A=a)\\ & - \widetilde{p}_{0,0}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=0|Y=y, A=0) \; - \widetilde{p}_{1,0}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=1|Y=y, A=0)\Big| \le \widetilde{\alpha}_n, \; \forall y, a\\ & 0 \le p_{\hat{y},a} \le 1 \quad \forall \hat{y} \in \{0,1\}, \forall a \in \mathcal{A}\end{aligned} \tag{15}$$

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 $n \ge \frac{16\log(8|\mathcal{A}|/\delta)}{\min_{ya}\mathbf{P}_{ya}}$, $\alpha_n = \sqrt{\frac{8\log 64/\delta}{n\min_{yz}\mathbf{Q}_{yz}}}$ and $\widetilde{\alpha}_n = \sqrt{\frac{\log(64/\delta)}{2n}\frac{4|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}_{ya}^2}}$, the predictor resulting from the two-step procedure satisfies:*

$$\begin{aligned} err(\widetilde{Y}) &\le_\delta err(Y^*) + \frac{5C}{\min_{ya}\mathbf{P}_{ya}^2}\left(\frac{2}{B} + 10\mathfrak{R}_{\frac{\min_{ya} n\mathbf{P}_{ya}}{4}}(\mathcal{H}) \; + 18|\mathcal{A}|\sqrt{\frac{2\log 64|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}\;\right)\\ disc(\widetilde{Y}) &\le_\delta \sqrt{\frac{\log(\frac{64}{\delta})}{2n}\frac{8|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}_{ya}^2}}\end{aligned} \tag{16}$$

Proof Sketch. Since the predictor obtained in step 1 is only a function of X, we can prove the following guarantees on its performance with $\widetilde{Y}^*$ being an optimal non-discriminatory derived

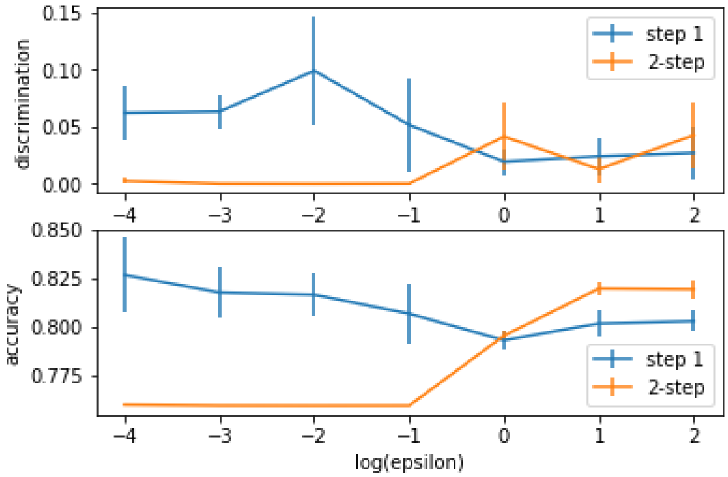

Figure 1: Plots of discrimination violation and accuracy of the step 1 predictor $\hat{Y}$ and the two-step predictor $\widetilde{Y}$ versus the privacy level ϵ on the Adult Income dataset Kohavi (1996). Error bars show 95% confidence interval for the average.

predictor from $\hat{Y}$:

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

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

We next 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 it s discrimination:

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

Plugging in the error and discriminating proved in Lemma 2 we obtain the theorem statement. A detailed proof is given in Appendix A.2.4. □

Our final predictor $\widetilde{Y}$ has a discrimination guarantee that is independent of the model complexity, however this comes at a cost of a privacy penalty entering the error bound. This creates a new set of trade-offs that do not appear in the absence of the privacy constraint, fairness and error start to trade-off more severely with increasing levels of privacy.

5.3 Experimental Illustration

Data We use the adult income data set Kohavi (1996) containing 48,842 examples. The task is to predict whether a person's income is higher than 50 thousand dollars. Each data point has 14 features including education and occupation, the protected attribute A we use is gender: male or female.

Approach We use a logistic regression model for classification. For the reductions approach, we use the fairlearn package available at `https://github.com/fairlearn/fairlearn`. We set $T = 50$, $\eta = 2.0$ and $B = 100$ for all experiments. We split the data into 75% for training and 25% for testing. We repeat the splitting over 10 trials.

Effect of privacy We plot in Figure 1 the resulting discrimination violation and model accuracy against increasing privacy levels ϵ for the predictor $\hat{Y}$ resulting from step 1 , trained on all the training data, and the two-step predictor $\widetilde{Y}$ trained on S_1 and S_2. We observe that $\widetilde{Y}$ achieves lower discrimination than $\hat{Y}$ across the different privacy levels. This comes at a cost of lower accuracy, which improves at lower privacy regimes (large epsilon). The predictor of step 1 only begins to suffer on accuracy when the privacy level is low enough as the fairness constraint is void at high levels of privacy (small epsilon).

6 Discussion and Extensions

Could this approach for private demographic data be used to learn non-discriminatory predictors under other forms of deficiency in demographic information? In this section, we consider another case of interest: when individuals retain the choice of whether to release their sensitive information or not, as in the example of credit card companies. Practically, this means that the learner's data contains one part that has protected attribute labels and another that doesn't.

Privacy as effective number of labeled samples Suppose we are given n_ℓ fully labeled samples: $S_\ell = \{(x_1, a_1, y_1), \cdots, (x_{n_\ell}, a_{n_\ell}, y_{n_\ell})\}$ drawn *i.i.d* from an unknown distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ where $\mathcal{A} = \{0, 1, \cdots, |\mathcal{A}| - 1\}$ and $\mathcal{Y} = \{0, 1, \cdots, |\mathcal{Y}| - 1\}$, and n_u samples that are missing the protected attribute: $S_u = \{(x_1, y_1), \cdots, (x_{n_u}, y_{n_i})\}$ drawn *i.i.d* from the marginal of P over $\mathcal{X} \times \mathcal{Y}$. Define $n := n_\ell + n_u$, $S = S_\ell \cup S_u$ and let $\beta > 0$ be such that $n_\ell := \beta n$ and $n_u = (1 - \beta)n$. The objective is to learn a non-discriminatory predictor $\hat{Y}$ from the data S.

To mimic step 1 of our methodology, we propose to modify the reductions approach, so as to allow the learner, Q-player, to learn on the entirety of S while the auditor, $\boldsymbol{\lambda}$-player, uses only S_ℓ. We do this by first defining a two data set version of the Lagrangian, as such:

$$L^{S,S_\ell}(Q, \lambda) = \mathrm{err}^S(Q) + \lambda^\top (M\boldsymbol{\gamma}^{S_\ell}(Q) - \alpha \mathbf{1}). \tag{17}$$

This changes Algorithm 1 in two key ways: first, the update of $\boldsymbol{\theta}$ now only relies on S_ℓ and, second, the best response of the learner is still a cost-sensitive learning problem, however now the cost depends on whether sample i is in S_ℓ or S_u. If it is in S_u, i.e. it does not have a group label, then the instance loss is the misclassification loss, while if it is in S_ℓ its loss is defined as before. Lemma 3 characterizes the performance of the learned predictor $\hat{Y}$ using the approach just described.

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:

$$err(\hat{Y}) \leq_\delta err(Y^*) + 4\mathfrak{R}_n(\mathcal{H}) + 4\sqrt{\frac{\log 4/\delta}{n}} \tag{18}$$

$$disc(\hat{Y}) \leq_\delta \frac{2}{B} + 6\mathfrak{R}_{\frac{\min_{ya} n_\ell \mathbf{P}_{ya}}{2}}(\mathcal{H}) + 10\sqrt{\frac{2\log 32|\mathcal{A}|/\delta}{n_\ell \min_{ya} \mathbf{P}_{ya}}} \tag{19}$$

Notice the similarities between Lemma 2 and 3. The error bound we obtain depends on the entire number of samples n as in the privacy case and the discrimination guarantee is forcibly controlled by the number of labeled group samples n_ℓ. We can thus interpret the discrimination bound in Lemma 2 as having an effective number of samples controlled by the privacy level ϵ.

Trade-offs and proxies To complete the parallel with the proposed methodology, what remains is to mimic step 2, to devise ways to have lower sample complexities to achieve non-discrimination. Clearly the dependence on n_ℓ is statistically necessary and the only area of improvement is to remove the dependence on the complexity of the model class. If the sensitive attribute is never available at test time, we cannot apply the post-processing procedure of Hardt et al. (2016) in a two-stage fashion Woodworth et al. (2017).

In practice, to compensate for the missing direct information, if legally permitted, the learner may leverage multiple sources of data and combine them to obtain indirect access to the sensitive information Kallus et al. (2019) of individuals. The way this is modeled mathematically is by having recourse to proxies. One of the most widely used proxies is the Bayesian Improved Surname Geocoding (BISG) method, BISG is used to estimate race membership given the last name and geolocation of an individual Adjaye-Gbewonyo et al. (2014); Fiscella and Fremont (2006). Using this proxy, one can impute the missing membership labels and then proceed to audit or learn a predictor. But a big issue with proxies is that they may lead to biased estimators for discrimination Kallus et al. (2019). In order to avoid these pitfalls, one promising line of investigation is to learn it simultaneously with the predictor.

What form of proxies can help us measure the discrimination of a certain predictor $\hat{Y} : \mathcal{X} \to \mathcal{Y}$? Some of the aforementioned issues are due to the fact that features X are in general insufficient to estimate group membership, even through the complete probabilistic proxy $\mathbb{P}(A|X)$. In particular for EO, if A is not completely identifiable from X then using this proxy leads to inconsistent estimates. In contrast, if we have access to the probabilistic proxy $\mathbb{P}(A|X,Y)$, we then propose the following estimator (see also Chen et al. (2019))

$$\widetilde{\gamma}^S_{ya}(\hat{Y}) = \frac{\sum_{i=1}^n \hat{Y}(x_i)\mathbf{1}(y_i = y)\mathbb{P}(A = a|x_i, y_i)}{\sum_{i=1}^n \mathbf{1}(y_i = y)\mathbb{P}(A = a|x_i, y_i)}, \tag{20}$$

which enjoys consistency, via a relatively straightforward proof that we omit:

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 $\widetilde{\gamma}^S_{ya}$ is consistent. As $n \to \infty$*

$$\widetilde{\gamma}^S_{ya} \to_p \gamma_{ya}.$$

We end our discussion here by pointing out that if such a proxy can be efficiently learned from samples, then it can reduce a missing attribute problem effectively to a private attribute problem, allowing us to use much of the same machinery presented in this paper.

7 Conclusion

We studied learning non-discriminatory predictors when the protected attributes are privatized or noisy. We observed that, in the population limit, non-discrimination against noisy attributes is equivalent to that against original attributes. We showed this to hold for various fairness criteria. We then characterized the amount of difficulty, in sample complexity, that privacy adds to testing non-discrimination. Using this relationship, we proposed how to carefully adapt existing non-discriminatory learners to work with privatized protected attributes. Care is crucial, as naively using these learners may create the illusion of non-discrimination, while continuing to be highly discriminatory. We ended by highlighting future work on how to learn predictors in the absence of any demographic data or prior proxy information.

References

Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., and Omer, S. B. (2014). 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.

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

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

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

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

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

Bureau, C. F. P. (2014). Using publicly available information to proxy for unidentified race and ethnicity.

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

Commission, F. T. (2013). Your equal credit opportunity rights.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kohavi, R. (1996). 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*, pages 202–207.

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

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

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

Vigdor, N. (2019). Apple card investigated after gender discrimination complaints.

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

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

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

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

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 2. 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 $h(x) = \begin{cases} 0 \text{ if } x \leq 1/2 \\ \frac{1}{2x} \text{ if } x > 1/2 \end{cases}$, note that $h(x) \in [0, 1]$ so that the predictor $\hat{Y}_2$ is valid.

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

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:

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

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

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

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

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

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

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:

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

$$+ \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{P}\left(\mathcal{E}_1 | \mathcal{E}_2, A = a'\right)) \frac{\bar{\pi}\mathbb{P}(A = a', \mathcal{E}_2)}{\mathbb{P}(Z = a', \mathcal{E}_2)} \tag{21}$$

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 (21) 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 (21) 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 $\mathcal{E}_{2_y} = \{Y = y\}$:

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

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 $p_a = \mathbb{P}(A = a)$:

$$
\begin{aligned}
&\mathbb{P}(\hat{Y} = 1|Z = a)\\
&= \sum_{a'} \mathbb{P}(\hat{Y} = 1|Z = a, A = a')\mathbb{P}(A = a'|Z = a)\\
&= \sum_{a'} \mathbb{P}(\hat{Y} = 1|A = a')\mathbb{P}(A = a'|Z = a)
\end{aligned}
$$

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

Now for equal accuracy among groups:

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

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

$$
\begin{aligned}
&\mathbb{P}(\hat{Y} = 1|Y = y, Z = a) = \\
&\sum_{a'} \mathbb{P}(\hat{Y} = 1|Y = y, Z = a, A = a')\mathbb{P}(A = a'|Z = a, Y = y) \\
&\overset{(a)}{=} \sum_{a'} \mathbb{P}(\hat{Y} = 1|Y = y, A = a')\mathbb{P}(A = a'|Z = a, Y = y) \\
&= \sum_{a'} \mathbb{P}(\hat{Y} = 1|Y = y, A = a')\frac{\mathbb{P}(Z = a, Y = y|A = a')\mathbb{P}(A = a')}{\mathbb{P}(Z = a, Y = y)} \\
&\overset{(b)}{=} \sum_{a'} \mathbb{P}(\hat{Y} = 1|Y = y, A = a')\frac{\mathbb{P}(Z = a|A = a')\mathbb{P}(Y = y|A = a')\mathbb{P}(A = a')}{\mathbb{P}(Z = a, Y = y)} \\
&= \pi\mathbb{P}(\hat{Y} = 1|Y = y, A = a)\frac{\mathbb{P}(Y = y, A = a)}{\mathbb{P}(Z = a, Y = y)} \qquad (22) \\
&+ \sum_{a'\backslash a} \bar{\pi}\mathbb{P}(\hat{Y} = 1|Y = y, A = a')\frac{\mathbb{P}(Y = y, A = a')}{\mathbb{P}(Z = a, Y = y)}
\end{aligned}
$$

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 (22) 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 $i \neq j \in \mathcal{A}$:

$$
\begin{aligned}
&G_i G^{-1}_{,j} \\
&= \sum_k G_{i,k} G^{-1}_{k,j}
\end{aligned}
$$

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

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

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

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:

$$
\begin{aligned}
\mathbf{Q}_{yz}&=\sum_{a\in\mathcal{A}}\mathbb{P}\left(Y=y,Z=z|A=a\right)\mathbb{P}\left(A=a\right)\\
&=\sum_{a\in\mathcal{A}}\mathbb{P}\left(Y=y|A=a\right)\mathbb{P}\left(Z=z|A=a\right)\mathbb{P}(A=a)\\
&=\pi\mathbb{P}\left(Y=y,A=z\right)+\sum_{a\neq z}\bar{\pi}\mathbb{P}\left(Y=y,A=a\right) \qquad (23)
\end{aligned}
$$

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

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

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 $\forall z, S_{yz} : n_{yz}^S \geq \frac{n\mathbf{Q}_{yz}}{2}$ and the complement of the previous event and upper bounding this complement by the probability that there $\exists z$ s.t. $n_{yz}^S < \frac{n\mathbf{Q}_{yz}}{2}$. Finally step (c) comes from a union bound and then a Chernoff bound on $n_{yz}^S \sim \text{Binomial}(n, \mathbf{Q}_{yz})$ 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 $\hat{G}_z^{-1} q_{y,.}^S$ 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 $C_2 = \frac{\pi - 1}{|\mathcal{A}|\pi - 1}$:

$$
\sup_{S,S'} |\hat{G}_z^{-1} q_{y,.}^S - \hat{G}_z^{-1} q_{y,.}^{S'}|
$$

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

Therefore by McDiarmid's inequality we have:

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

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

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

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:

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

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 $\text{Proj}_{\Delta}(\hat{G}^{-1} q_{y,\cdot})$ where $\text{Proj}_{\Delta}(x)$ is the orthogonal projection of x onto the simplex defined as:

$$
\text{Proj}_{\Delta}(\mathbf{x}) := \quad \arg\min_{\mathbf{y}} \frac{1}{2} \left|\left| \mathbf{y} - \mathbf{x} \right|\right|_2^2
$$

$$\text{s.t. } \mathbf{y}^T\mathbf{1} = 1,\ \mathbf{y} \geq 0$$

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:

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

Step 4: *Difference of Equalized odds*
Let $h_{ya} = \text{Proj}_\Delta(\hat{G}_a^{-1}q^S_{y,.})$, using a series of triangle inequality,

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

hence

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

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:

$$\begin{aligned}
L(h,\lambda) &= \hat{\text{err}}(h) + \lambda^\top (M\boldsymbol{\gamma}(h) - \alpha_n \mathbf{1}) \\
&= \frac{1}{n}\sum_{i\in S} \mathbb{I}_{h(x_i)\neq y_i} - \alpha_n \lambda^\top \mathbf{1} + \sum_{k,j} M_{k,j}\lambda_k \gamma_j^S(h) \\
&= -\alpha_n \lambda^\top \mathbf{1} + \frac{1}{n}\sum_{i\in S} \mathbb{I}_{h(x_i)\neq y_i} + \sum_{k,j} M_{k,j}\lambda_k \frac{1/n\cdot h(x_i)\mathbb{I}_{(y_i,a_i)=j}}{1/n\sum_{s\in S}\mathbb{I}_{(y_s,a_s)=j}} \qquad (24)
\end{aligned}$$

Thus from equation (24) 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:

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

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:

$$L(\hat{Q},\hat{\lambda}) \leq L(Q,\hat{\lambda}) + \vartheta \quad \forall Q \in \Delta_{\mathcal{H}} \qquad (25)$$

$$L(\hat{Q},\hat{\lambda}) \geq L(\hat{Q},\boldsymbol{\lambda}) - \vartheta \quad \forall \boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{K}|}, ||\boldsymbol{\lambda}||_1 \leq B \qquad (26)$$

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 5. *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

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

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

$$
q_{y1}(\hat{Y})-q_{y0}(\hat{Y})=\Big(\gamma_{y1}(\hat{Y})-\gamma_{y0}(\hat{Y})\Big)\frac{(2\pi-1)\mathbf{P}_{y1}\mathbf{P}_{y0}}{\mathbf{Q}_{y1}\mathbf{Q}_{y0}} \tag{27}
$$

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:

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

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 $\gamma_{y\cdot}$, denoting $\mathbf{P}_{ya}=\mathbb{P}(Y=y,A=a)$:

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

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

where step (a) follows from expanding by equation (23). 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:

$$
\begin{aligned}
|\gamma_{y,k} - \gamma_{y,j}| &= |G_k^{-1}q_y - G_j^{-1}q_y| \\
&\overset{(a)}{=} |G_k^{-1}(q_y - q') - G_j^{-1}(q_y - q')| \\
&= |(q_y - q')(G_k^{-1} - G_j^{-1})| \\
&\leq |q_y - q'|_\infty |G_k^{-1} - G_j^{-1}|_1 \text{ (Holder's Inequality)} \\
&= |\max_z q_{y,z} - \min_{z'} q_{y,z'}| \cdot |G_k^{-1} - G_j^{-1}|_1
\end{aligned} \tag{28}
$$

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 (28), for ease of notation let $\mathbb{P}(Z = i, Y = y) = z_i$ and $\mathbb{P}(A = i, Y = y) = a_i$:

$$
\begin{aligned}
&|G_k^{-1} - G_j^{-1}|_1 \\
&= \sum_{a \in \mathcal{A}\backslash\{k,j\}} \left| C_2(\frac{z_a}{a_k} - \frac{z_a}{a_i}) \right| + \left| C_1\frac{z_k}{a_k} - C_2\frac{z_k}{a_i} \right| + \left| C_2\frac{z_i}{a_k} - C_1\frac{z_i}{a_i} \right| \\
&= \sum_{a \in \mathcal{A}\backslash\{k,j\}} \left| C_2(\frac{z_a a_i - z_a a_k}{a_k a_i}) \right| + \left| C_1\frac{z_k a_i}{a_k a_i} - C_2\frac{z_k a_k}{a_k a_i} \right| + \left| C_2\frac{z_i a_i}{a_k a_i} - C_1\frac{z_i a_k}{a_k a_i} \right|
\end{aligned}
$$

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

Hence we have the following inequality:

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

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

□

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

Lemma 7 (Lemma 2 Agarwal et al. (2018)). *For any distribution Q satisfying the empirical constraints on dataset S: $M\gamma^S(Q) \le \alpha_n\mathbf{1}$, $\hat{Q}$ the output $\hat{Y}$ of Algorithm 1 satisfies:*

$$
err^S(\hat{Y}) \le err(Q) + 2\vartheta \tag{29}
$$

Lemma 8 (Lemma 3 Agarwal et al. (2018)). *The discrimination of $\hat{Y}$, output of Algorithm 1, satisfies:*

$$
\max_{y,a} |q^S_{y,a}(\hat{Y}) - q^S_{y,0}(\hat{Y})| \le 2\alpha_n + 2\frac{1+2\vartheta}{B} \tag{30}
$$

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 $n \ge \frac{16\log 8|\mathcal{A}|/\delta}{\min_{ya}\mathbf{P}_{ya}}$, $\alpha_n = 2\sqrt{\frac{\log|\mathcal{A}|/\delta}{n\min_{ya}\mathbf{P}_{ya}}}$ and we let $\vartheta = \mathfrak{R}_{n/2}(\mathcal{H}) + \sqrt{\frac{\log 8/\delta}{n}}$, then running Algorithm 1 on dataset S with $T \ge \frac{16\log(4|\mathcal{A}|+1)}{\vartheta^2}$ and learning rate $\eta = \frac{\vartheta}{8B}$ returns a predictor $\hat{Y}$ satisfying the following:*

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

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 $\text{err}(\hat{Y}) = \mathbb{P}(\hat{Y} \neq Y)$, 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:

$$\text{err}(\hat{Y}) \leq_{\delta/4} \text{err}^S(\hat{Y}) + \mathfrak{R}_{n/2}(\mathcal{H}) + \sqrt{\frac{\log 8/\delta}{n}} \tag{31}$$

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 (31) :

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

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

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

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:

$$|q_{ya}^S - q_{y0}^S| \leq 2\alpha_n + 2\frac{1 + 2\vartheta}{B} \tag{32}$$

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):

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

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 $\widetilde{Y}$:

$$\widetilde{Y} = \arg\min_{p_{\cdot,\cdot}} \quad \sum_{\hat{y},a} (\mathbb{P}(\hat{Y} = \hat{y}, A = a, Y = 0) - \mathbb{P}(\hat{Y} = \hat{y}, A = a, Y = 1)) \cdot p_{\hat{y},a} \tag{33}$$

$$s.t. \quad p_{0,a} \mathbb{P}(\hat{Y} = 0 | Y = y, A = a) + p_{1,a} \mathbb{P}(\hat{Y} = 1 | Y = y, A = a) \tag{34}$$

$$= p_{0,0} \mathbb{P}(\hat{Y} = 0 | Y = y, A = 0) + p_{1,0} \mathbb{P}(\hat{Y} = 1 | Y = y, A = 0), \quad \forall y \in \{0, 1\}, \forall a \in \mathcal{A} \tag{35}$$

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

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 nav̈e 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 nav̈e 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:

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

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 $\widetilde{p}_{\hat{y},a} = \pi p_{\hat{y},a} + \sum_{a' \setminus a} \bar{\pi} p_{\hat{y},a'}$ 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:

$$\widetilde{Y} = \arg\min_{p_{\cdot,\cdot}} \quad \sum_{\hat{y},a} (\widetilde{\mathbb{P}}^S(\hat{Y}=\hat{y}, Z=a, Y=0) - \widetilde{\mathbb{P}}^S(\hat{Y}=\hat{y}, Z=a, Y=1)) \cdot \widetilde{p}_{\hat{y},a} \tag{36}$$

$$\begin{aligned} s.t. \quad & \widetilde{p}_{0,a}\widetilde{\mathbb{P}}^S(\hat{Y}=0|Y=y, A=a) + \widetilde{p}_{1,a}\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=a) \\ & = \widetilde{p}_{0,0}\widetilde{\mathbb{P}}^S(\hat{Y}=0|Y=y, A=0) + \widetilde{p}_{1,0}\widetilde{\mathbb{P}}^S(\hat{Y}=1|Y=y, A=0), \quad \forall y \in \{0,1\}, \forall a \in \mathcal{A} \qquad (37) \\ & 0 \leq p_{\hat{y},a} \leq 1 \quad \forall \hat{y} \in \{0,1\}, \forall a \in \mathcal{A} \end{aligned}$$

A.2.4 Second Step Guarantees

Lemma 9 (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 $n \geq \frac{32\log(8|\mathcal{A}|/\delta)}{\min_{ya} \mathbf{P}_{ya}}$, $\widetilde{\alpha}_n = \sqrt{\frac{\log(64/\delta)}{2n}\frac{4|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya^*}}}$ and with $\widetilde{Y}*$ an optimal 0-discriminatory predictor derived from $\hat{Y}$, then with probability greater than $1-\delta/2$ we have:*

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

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

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)):

$$\begin{aligned} \left|\text{err}^S(\widetilde{Y}) - \text{err}(\widetilde{Y})\right| &= \left|\sum_{\hat{y},a} \widetilde{p}_{\hat{y},a} \cdot \left((\widetilde{q}^S_{\hat{y},a,0} - q_{\hat{y},a,0}) + (q_{\hat{y},a,1} - \widetilde{q}^S_{\hat{y},a,1})\right)\right| \\ &\leq |\sum_{\hat{y},a} \widetilde{q}^S_{\hat{y},a,0} - q_{\hat{y},a,0}| + |\sum_{\hat{y},a} \ q_{\hat{y},a,1} - \widetilde{q}^S_{\hat{y},a,1}| \\ &\leq \sum_{\hat{y},a} |\widetilde{q}^S_{\hat{y},a,0} - q_{\hat{y},a,0}| + \sum_{\hat{y},a} | \ q_{\hat{y},a,1} - \widetilde{q}^S_{\hat{y},a,1}| \end{aligned} \tag{38}$$

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:

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

Hence

$$\begin{aligned}\mathbb{P}(\left|\mathrm{err}^S(\widetilde{Y}) - \mathrm{err}(\widetilde{Y})\right| > t) &\leq \mathbb{P}(\sum_{\hat{y},a} |q^S_{\hat{y},a,0} - q_{\hat{y},a,0}| + \sum_{\hat{y},a} |\ q_{\hat{y},a,1} - q^S_{\hat{y},a,1}| > t) \\ &\leq 8|\mathcal{A}|\exp(-2n\left(\frac{t}{4|\mathcal{A}|}\frac{|\mathcal{A}|\pi - 1}{\pi + |\mathcal{A}| - 2}\right)^2) \qquad (39)\end{aligned}$$

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

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

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:

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

□

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 $\widetilde{\alpha}_n = \sqrt{\frac{\log(64/\delta)}{2n}\frac{4|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}_{ya}^2}}$ then the predictor resulting from the two-step procedure satisfies:

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

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

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:

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

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

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:

$$\operatorname{err}(\widetilde{Y}^*) \leq \operatorname{err}(\hat{Y}) + \operatorname{disc}(\hat{Y}) \tag{40}$$

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

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

Hence our derived predictor satisfies:

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

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

□

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:*

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

$$\mathrm{disc}(\hat{Y}) \leq_\delta \frac{2}{B} + 6\mathfrak{R}_{\frac{\min_{ya} n_\ell \mathbf{P}_{ya}}{2}}(\mathcal{H}) + 10\sqrt{\frac{2\log 32|\mathcal{A}|/\delta}{n_\ell \min_{ya}\mathbf{P}_{ya}}}$$

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.

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

□