Fair Learning with Private Demographic Data

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

Abstract

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

1. 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 by potentially underestimating discrimination (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, ignoring legal feasibility, 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.

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

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

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

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 & 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; Wang et al., 2020).

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, let $\mathcal{E}_1, \mathcal{E}_2$ be two probability events defined with respect to $(X, Y, \hat{Y})$, then define $(\mathcal{E}_1, \mathcal{E}_2)$-non-discrimination with respect to A as having:

$$\mathbb{P}\left(\mathcal{E}_1 | \mathcal{E}_2, A = a\right) = \mathbb{P}\left(\mathcal{E}_1 | \mathcal{E}_2, A = a'\right) \quad \forall a, a' \in \mathcal{A} \quad (1)$$

All the notions considered in Definition 1 can be cast into the above formulation for one or more set of events $(\mathcal{E}_1, \mathcal{E}_2)$. Additionally, one can naturally define approximate versions of the above fairness constraints. As an example, for the notion of equalized odds, let $\mathcal{A} = \{0, 1, \cdots, |\mathcal{A}| - 1\}$ and define $\gamma_{y,a}(\hat{Y}) = \mathbb{P}(\hat{Y} = 1 | Y = y, A = a)$, 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 of 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 immediately 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

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

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

$$\begin{cases} P_{i,i} = \frac{\pi\mathbb{P}(A=i,\mathcal{E}_2)}{\mathbb{P}(Z=i,\mathcal{E}_2)} \text{ for } i \in \mathcal{A} \\ P_{i,j} = \frac{\bar{\pi}\mathbb{P}(A=j,\mathcal{E}_2)}{\mathbb{P}(Z=i,\mathcal{E}_2)} \text{ for } i, j \in \mathcal{A} \text{ s.t.} i \neq j \end{cases} \tag{2}$$

Then we have the following linear system of equations:

$$\begin{bmatrix} \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z = 0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, Z = |\mathcal{A}| - 1) \end{bmatrix} = P \begin{bmatrix} \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = 0) \\ \vdots \\ \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = |\mathcal{A}| - 1) \end{bmatrix} \tag{3}$$

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. □

Note that while $\hat{Y}$ not being a function of Z is a sufficient condition for the conclusion in Proposition 1 to hold, the

more general condition for EO is that $\hat{Y}$ is independent of Z given A and Y, however actualizing this condition beyond simply ignoring Z at test time is unclear. We next study how to measure non-discrimination from samples. Unfortunately, Proposition 1 applies only in the population-limit. For example for the EO notion, 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 (3). This is done by inverting the matrix P defined in (2), 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 $\Pi_k^{-1} \mathbf{Q}_{y,.} = \mathbb{P}(Y = y, A = k)$ where Π_k^{-1} is the k'th row of Π^{-1}. Hence we can plug this estimate in 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:*

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

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:

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

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

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 learning method, such as the reductions approach of (Agarwal et al., 2018), 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) which we detail in the next subsection. 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 by 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 $\widetilde{Y} = f(\hat{Y}, Z)$. The predictor in the second step *does use* Z, however with a careful analysis we are able to show that it indeed guarantees non-discrimination with respect to A; note that naively using the post-processing procedure of (Hardt et al., 2016) fails. 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 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:

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

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

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

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

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 strategy of (Freund & Schapire, 1996). The auditor follows the exponentiated gradient algorithm and the learner picks it's best response to the auditor. 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 η, rounds T
$\boldsymbol{\theta}_1 \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{K}|}$
for $t = 1, 2, \cdots, T$ **do**
  $\boldsymbol{\lambda}_{t,k} \leftarrow B \frac{\exp(\theta_{t,k})}{1 + \sum_{k'} \exp(\theta_{t,k})} \forall k \in \mathcal{K}$
  $h_t \leftarrow \mathrm{BEST}_h(\boldsymbol{\lambda}_t)$
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \eta(M\boldsymbol{\gamma}^S(h_t) - \alpha_n \mathbf{1})$
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, $\mathrm{BEST}_h(\boldsymbol{\lambda})$), 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 $\mathrm{disc}(\hat{Y}) = \max_{y \in \{0,1\}, a \in \mathcal{A}} \Gamma_{ya}$.

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:*

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

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

Proof of Lemma 2 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 $\mathfrak{R}_n(\mathcal{H})$ 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.

Comparison with Differentially Private predictor. 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 (Theorem 4.4 (Jagielski et al., 2018)) both scale with the privacy level ϵ and the complexity of $\mathcal{H}$, meaning the excess terms in the bounds of Lemma 2 are both in the order of $O(\mathfrak{R}_n(\mathcal{H})/\epsilon)$ in their work. Contrast this with our error bound that is independent of ϵ, the catch is that discrimination obtained with LDP is significantly more impacted by the privacy level ϵ. Thus, central differential privacy and local differential privacy in this context give rise to a very different set of trade-offs.

5.2. Step 2: Post-hoc correction to achieve non-discrimination with respect to 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 to 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:

$$\begin{aligned}&\mathbb{P}(\widetilde{Y}=1|Y=y,A=a)= \\ &\mathbb{P}(\widetilde{Y}=1|\hat{Y}=0,A=a)\mathbb{P}(\hat{Y}=0|Y=y,A=a)\\ &+\mathbb{P}(\widetilde{Y}=1|\hat{Y}=1,A=a)\mathbb{P}(\hat{Y}=1|Y=y,A=a)\end{aligned} \tag{12}$$

and we have that:

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

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 (12).

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

$$\begin{aligned}\widetilde{Y}=\arg\min_{p_{\cdot,\cdot}}\quad &\sum_{\hat{y},a}\Big(\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=\hat{y},Z=a,Y=0)\\ &-\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=\hat{y},Z=a,Y=1)\Big)\cdot\widetilde{p}_{\hat{y},a}\end{aligned}$$

$$\begin{aligned}s.t.\quad &\Big|\widetilde{p}_{0,a}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=0|Y=y,A=a)\\ &+\widetilde{p}_{1,a}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=1|Y=y,A=a)\\ &-\widetilde{p}_{0,0}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=0|Y=y,A=0)\\ &-\widetilde{p}_{1,0}\widetilde{\mathbb{P}}^{S_2}(\hat{Y}=1|Y=y,A=0)\Big|\le\widetilde{\alpha}_n,\ \forall y,a\\ &0\le p_{\hat{y},a}\le 1\quad\forall\hat{y}\in\{0,1\},\forall a\in\mathcal{A}\end{aligned} \tag{13}$$

The following Theorem illustrates the performance of our proposed estimator $\widetilde{Y}$.

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

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

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 predictor from $\hat{Y}$:

$$\begin{aligned}\mathrm{err}(\widetilde{Y})\le_{\delta/2}\ &\mathrm{err}(\widetilde{Y}^*)+4|\mathcal{A}|C\sqrt{\frac{\log(32|\mathcal{A}|/\delta)}{2n}}\\ \mathrm{disc}(\widetilde{Y})\le_{\delta/2}\ &\sqrt{\frac{\log(\frac{64}{\delta})}{2n}}\frac{8|\mathcal{A}|C^2}{\min_{ya}\mathbf{P}^2_{ya}}\end{aligned}$$

Next, 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 no greater than the sum of the loss of the base predictor and its discrimination:

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

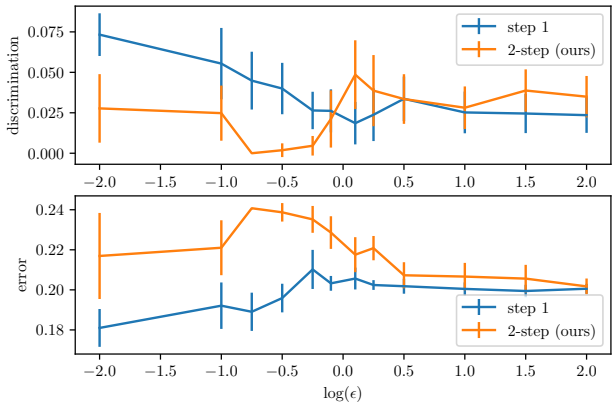

Figure 1. Plots of discrimination violation and accuracy of the step 1 predictor $\hat{Y}$ and the two-step predictor $\widetilde{Y}$ versus the privacy level ϵ on the Adult Income dataset (Kohavi, 1996). Error bars show 95% confidence interval for the average.

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 $\$50k$. 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 implementation in the fairlearn package [1]. 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).

Code to reproduce Figure 1 is publicly available [2].

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. As a first step towards understanding this setting, 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: $S_u = \{(x_1, y_1), \cdots, (x_{n_u}, y_{n_u})\}$ drawn $i.i.d$ from the marginal of $\mathcal{P}$ 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 $n_u = (1 - \beta)n$. This data assumption is equivalent to having individuals not reporting their attributes uniformly at random with probability $1 - \beta$. 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:

$$L^{S,S_\ell}(Q, \boldsymbol{\lambda}) = \mathrm{err}^S(Q) + \boldsymbol{\lambda}^\top (M\boldsymbol{\gamma}^{S_\ell}(Q) - \alpha \mathbf{1}). \quad (14)$$

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 $n_\ell \geq \frac{8 \log 4|\mathcal{A}|/\delta}{\min_{ya} \mathbf{P}_{ya}}$,*

[1] https://github.com/fairlearn/fairlearn

[2] https://github.com/husseinmozannar/fairlearn_private_data

$\alpha_n = 2\sqrt{\frac{\log 32|\mathcal{A}|/\delta}{n_\ell \min_{ya} \mathbf{P}_{ya}}}$ *and we let* $\vartheta = \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{\log 4/\delta}{n}}$, *then running the modified Algorithm 1 on data set* S *and* S_l *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:*

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

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

A short proof of Lemma 3 can be found in Appendix A. 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.

Individual choice of reporting It may be more realistic to assume that the choice of individuals to report their protected attributes may depend on their own characteristics, let $t(x, y, a) \in (0, 1]$ (reporting probability function) be the probability that an individual (x, y, a) chooses to report their protected attributes. This can be codified using a binary reporting random variable T where $\mathbb{P}(T = 1|X = x, Y = y, A = a) = t(x, y, a)$. Note that in the setting of Lemma 3, $t(x, y, a) = \beta$, a constant. Starting from a dataset S of n individuals sampled $i.i.d.$ from $\mathbb{P}$, each individual i flips a coin with bias $t(x_i, y_i, a_i)$ and accordingly chooses to include their attribute a_i in S. The result of this process is a splitting of S into $S_\ell = \{(x_1, a_1, y_1), \cdots, (x_{n_\ell}, a_{n_\ell}, y_{n_\ell})\}$ (individuals who report their attributes) and $S_u = \{(x_1, y_1), \cdots, (x_{n_u}, y_{n_u})\}$ (individuals who do not report). The goal again is to learn a non discriminatory predictor $\hat{Y}$.

The immediate question is whether we can use our modified algorithm with the two-dataset Langragian (14) and obtain similar guarantees to those in Lemma 3 in this more general setting. This question boils down to asking if the naïve empirical estimate of discrimination is consistent and the answer depends both on the reporting probability function t and the notion of discrimination considered as illustrated in the following proposition.

Proposition 2. *Consider* $(\mathcal{E}_1, \mathcal{E}_2)$*-non-discrimination with respect to* A*. Fix a reporting probability function* $t : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to (0, 1]$*. If the resulting* T *and* $\mathcal{E}_1$ *are conditionally independent given* $\{A, \mathcal{E}_2\}$*, then for all* $a \in \mathcal{A}$ *we have*

$$\mathbb{P}^{S_\ell}(\mathcal{E}_1|\mathcal{E}_2, A = a) \to_p \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = a), \quad \text{as } n \to \infty.$$

where, for each n*,* S_ℓ *is generated via individual random reporting through* t*.*

Proof. Given $a \in \mathcal{A}$, our estimate $\mathbb{P}^{S_\ell}(\mathcal{E}_1|\mathcal{E}_2, A = a)$ is nothing but the empirical estimator of $\mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = a, T = 1)$ where $\{T = 1\}$ denotes the event that an individual does report their attributes and are thus included in S_ℓ. As an immediate consequence we have:

$$\mathbb{P}^{S_\ell}(\mathcal{E}_1|\mathcal{E}_2, A = a) \to_p \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = a, T = 1)$$

Now, by assumption, T and $\mathcal{E}_1$ are conditionally independent given $\{A, \mathcal{E}_2\}$ and thus:

$$\mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = a, T = 1) = \mathbb{P}(\mathcal{E}_1|\mathcal{E}_2, A = a)$$

which completes the proof. Note that the event $\mathcal{E}_1$ has strictly positive probability given $\{\mathcal{E}_2, A = a, T = 1\}$, as the reporting probability function is strictly positive. □

If the independence condition in Proposition 2 is satisfied, then this immediately suggests that we can run Algorithm (1) and obtain learning guarantees.
To illustrate this concretely, suppose the notion of non-discrimination is EO. Consider any reporting probability function of the form $t_1 : \mathcal{Y} \times \mathcal{A} \to (0, 1]$ (does not depend on the non-sensitive attributes). Furthermore suppose our hypothesis class consists of functions that depend only on X. The conditional independence condition in Proposition 2 thus holds and we can estimate the discrimination of any predictor in our class. The only change to Lemma 3 in this setup is that the effective number of samples in the discrimination bound is now: $n \min_{ya} \mathbf{P}_{ya} \cdot \mathbf{T}_{ya}$ where $\mathbf{T}_{ya} = \mathbb{P}(T = 1|Y = y, A = a)$ (T is the r.v. that denotes reporting); the proof of this observation is immediate.

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 in Lemma 3 is statistically necessary without any assumptions on the data generating process 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 & 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))

$$\widetilde{\gamma}^S_{ya}(\hat{Y}) = \frac{\sum_{i=1}^n \hat{Y}(x_i)\mathbf{1}(y_i = y)\mathbb{P}(A = a|x_i, y_i)}{\sum_{i=1}^n \mathbf{1}(y_i = y)\mathbb{P}(A = a|x_i, y_i)}, \quad (15)$$

which enjoys consistency, via a relatively straightforward proof found in Appendix A.

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. With the same approach and without recourse to proxy information, we ended by highlighting when and how we can learn predictors when individuals can choose to withhold their protected attributes.

Acknowledgements

This paper is based upon work supported in part by the NSF Program on Fairness in AI in Collaboration with Amazon under Award No. IIS-1939743, titled FAI: Addressing the 3D Challenges for Data-Driven Fairness: Deficiency, Dynamics, and Disagreement. Any opinion, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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