A Fairness Analysis on Private Aggregation of Teacher Ensembles

Cuong Tran, My H. Dinh, Kyle Beiter, Ferdinando Fioretto

Syracuse University
cutran@syr.edu, mydinh@syr.edu, kbeiter@syr.edu, ffiorett@syr.edu

Abstract

The Private Aggregation of Teacher Ensembles (PATE) (Papernot et al. 2018) is an important private machine learning framework. It combines multiple learning models used as teachers for a student model that learns to predict an output chosen by noisy voting among the teachers. The resulting model satisfies differential privacy and has been shown effective in learning high quality private models in semisupervised settings or when one wishes to protect the data labels.

This paper asks whether this privacy-preserving framework introduces or exacerbates bias and unfairness and shows that PATE can introduce accuracy disparity among individuals and groups of individuals. The paper analyzes which algorithmic and data properties are responsible for the disproportionate impacts, why these aspects are affecting different groups disproportionately, and proposes guidelines to mitigate these effects. The proposed approach is evaluated on several datasets and settings.

1 Introduction

The availability of large datasets and inexpensive computational resources has rendered the use of machine learning (ML) systems instrumental for many critical decisions involving individuals, including criminal assessment, landing, and hiring, all of which have a profound social impact. A key concern for the adoption of these system regards how they handle bias and discrimination and how much information they leak about the individuals whose data is used as input.

Differential Privacy (DP) (Dwork et al. 2006) is an algorithmic property that bounds the risks of disclosing sensitive information of individuals participating in a computation. It has become the paradigm of choice in privacy-preserving machine learning systems and its deployments are growing at a fast rate. However, it was recently observed that DP systems may induce biased and unfair outcomes for different groups of individuals (Bagdasaryan, Poursaeed, and Shmatikov 2019; Pujol et al. 2020; Xu, Du, and Wu 2021).

The resulting outcomes can have significant societal and economic impacts on the involved individuals: classification errors may penalize some groups over others in important determinations including criminal assessment, landing, and hiring (Bagdasaryan, Poursaeed, and Shmatikov 2019) or can result in disparities regarding the allocation of critical funds and benefits (Pujol et al. 2020). *While these surprising observations are becoming increasingly common, their causes are largely understudied and not fully understood.*

This paper makes a step toward this important quest, and studies the disparate impacts arising when training a model using *Private Aggregation of Teacher Ensembles* (PATE) (Papernot et al. 2018) an important and popular privacy-preserving machine learning framework. It combines multiple agnostic learning models used as teachers for a student model which learns to predict an output chosen by noisy voting among the teachers. The resulting model satisfies differential privacy and has been shown effective in learning high quality private models in semisupervised settings or when one wishes to protect the data labels.

The paper analyzes which properties of the algorithm and the data are responsible for the disproportionate impacts, why these aspects are affecting different individuals or groups of individuals disproportionately, and proposes a solution that may aid mitigating these effects.

In summary, the paper makes the following contributions:

1. It uses a fairness notion that relies on the concept of excessive risk, and measures the direct impact of privacy to the model accuracy for individuals or groups.
2. It analyzes this fairness notion in PATE, a state-of-the-art privacy-preserving ML framework.
3. It isolates key components of the model parameters and the data properties which are responsible for the observed disparate impacts.
4. It studies when and why these components affect different individuals or groups disproportionately.
5. Finally, based on these findings, it proposes a method that may aid in mitigating these unfairness effects while retaining high accuracy.

To the best of the authors knowledge, this work represents a first effort toward understanding the reasons of the disparate impacts in privacy-preserving ensemble models.

2 Related Work

The study of the disparate impacts caused by privacy-preserving algorithms has recently seen several important developments. Ekstrand, Joshaghani, and Mehrpouyan (2018) raise questions about the tradeoffs involved between

privacy and fairness. Cummings et al. (2019) study the tradeoffs arising between differential privacy and equal opportunity, a fairness notion requiring a classifier to produce equal true positive rates across different groups. They show that there exists no classifier that simultaneously achieves $(\epsilon, 0)$-DP, satisfies equal opportunity, and has accuracy better than a constant classifier. This development has risen the question of whether one can practically build fair models while retaining sensitive information private. Jagielski et al. (2018) presents two algorithms that satisfy (ϵ, δ)-differential privacy and equalized odds. Mozannar, Ohannessian, and Srebro (2020) develops methods to adapt a nondiscriminatory learner to work with privatized protected attributes and Tran, Fioretto, and Hentenryck (2021) proposes a differentially private learning approach to enforce several group fairness notions using a Lagrangian dual method.

Pujol et al. (2020) were seemingly the first to show, empirically, that resource allocation decisions made using DP datasets may disproportionately affect some groups of individuals over others. These studies were complemented theoretically by Tran et al. (2021). Similar observations were also made in the context of model learning. Bagdasaryan, Poursaeed, and Shmatikov (2019) empirically observed that the accuracy of a DP model trained using DP-Stochastic Gradient Descent (DP-SGD) decreases disproportionately across groups causing larger negative impacts to the underrepresented groups. Farrand et al. (2020); Uniyal et al. (2021) reaches similar conclusions and show that this disparate impact is not limited to highly imbalanced data.

This paper builds on this body of work and their important empirical observations. It provides an analysis for the reasons of unfairness in the context of semi-supervised private learning ensembles, a commonly adopted scheme in privacy-preserving ML systems as well as introduces mitigating guidelines.

3 Preliminaries: Differential Privacy

Differential privacy (DP) (Dwork et al. 2006) is a strong privacy notion used to quantify and bound the privacy loss of an individual's participation in a computation. Informally, it states that the probability of any output does not change much when a record is added or removed from a dataset, limiting the amount of information that the output reveals about any individual. The action of adding or removing a record from a dataset D, resulting in a new dataset D', defines the notion of *adjacency*, denoted $D \sim D'$.

Definition 1. *A mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ϵ, δ)-differentially private, if, for any two adjacent inputs $D \sim D' \in \mathcal{D}$, and any subset of output responses $R \subseteq \mathcal{R}$:*

$$\Pr[\mathcal{M}(D) \in R] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in R] + \delta.$$

Parameter $\epsilon > 0$ describes the *privacy loss* of the algorithm, with values close to 0 denoting strong privacy, while parameter $\delta \in [0, 1)$ captures the probability of failure of the algorithm to satisfy ϵ-DP. The global sensitivity Δ_ℓ of a real-valued function $\ell : \mathcal{D} \to \mathbb{R}$ is defined as the maximum amount by which ℓ changes in two adjacent inputs: $\Delta_\ell = \max_{D \sim D'} \|\ell(D) - \ell(D')\|$. In particular, the Gaussian mechanism, defined by $\mathcal{M}(D) = \ell(D) + \mathcal{N}(0, \Delta_\ell^2 \sigma^2)$, where $\mathcal{N}(0, \Delta_\ell^2 \sigma^2)$ is the Gaussian distribution with 0 mean and standard deviation $\Delta_\ell^2 \sigma^2$, satisfies (ϵ, δ)-DP for $\delta > \frac{4}{5} \exp(-(\sigma\epsilon)^2/2)$ and $\epsilon < 1$ (Dwork, Roth et al. 2014).

4 Problem Settings and Goals

This paper considers a *private* dataset D consisting of n individuals' data points $(\boldsymbol{x}_i, y_i)$, with $i \in [n]$, drawn i.i.d. from an unknown distribution Π. Therein, $\boldsymbol{x}_i \in \mathcal{X}$ is a feature vector that *may* contain a protected group attribute $\boldsymbol{a}_i \in \mathcal{A} \subset \mathcal{X}$, and $y_i \in \mathcal{Y} = [C]$ is a C-class label. For example, consider a classifier that needs to predict criminal defendant's recidivism. The training example features $\boldsymbol{x}_i$ may describe the individual's demographics, education, occupation, and crime committed, the protected attribute $\boldsymbol{a}_i$, if available, may describe the individual's gender or ethnicity, and y_i represents whether or not the individual has high risk to reoffend.

This paper studies the fairness implications arising when training privacy-preserving semi-supervised transfer learning models. The setting is depicted in Figure 1. We are given an ensemble of *teacher* models $\boldsymbol{T} = \{f^i\}_{i=1}^k$, with each $f^i : \mathcal{X} \to \mathcal{Y}$ trained on a non-overlapping portion D_i of D. This ensemble is used to transfer knowledge to a *student* model $\bar{f}_\theta : \mathcal{X} \to \mathcal{Y}$, where $\boldsymbol{\theta}$ denotes a vector of real-valued parameters associated with model $\bar{f}$.

The student model $\bar{f}$ is trained using a *public* dataset $\bar{D} = \{\boldsymbol{x}_i\}_{i=1}^m$ with samples drawn i.i.d. from the same distribution Π considered above but whose labels are unrevealed. The paper focuses on learning classifier $\bar{f}_\theta$ using knowledge transfer from the teacher model ensemble $\boldsymbol{T}$ while guaranteeing the privacy of each individual's data $(\boldsymbol{x}_i, y_i) \in D$. The sought model is learned by minimizing the regularized empirical risk function

$$\begin{aligned} \overset{\star}{\boldsymbol{\theta}} &= \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \bar{D}, \boldsymbol{T}) \\ &= \sum_{\boldsymbol{x} \in \bar{D}} \ell\left(\bar{f}_\theta(\boldsymbol{x}), v\left(\boldsymbol{T}(\boldsymbol{x})\right)\right) + \lambda \|\theta\|^2, \end{aligned} \tag{1}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function and measures the performance of the model, $v : \mathcal{Y}^k \to \mathcal{Y}$ is a *voting scheme* used to decide the prediction label from the ensemble $\boldsymbol{T}$, with $\boldsymbol{T}(\boldsymbol{x})$ used as a shorthand for $\{f^i(\boldsymbol{x})\}_{i=1}^k$, and $\lambda > 0$ is a regularization parameter.

The paper focuses on learning classifiers that protect the disclosure of the individual's data using the notion of differnetial privacy and it analyzes the fairness impact (as defined below) of privacy on different groups and individuals.

Privacy *Privacy* is achieved by using a differentially private version $\tilde{v}$ of the voting function v, defined as

$$\tilde{v}(\boldsymbol{T}(\boldsymbol{x})) = \operatorname{argmax}_j \{\#_j(\boldsymbol{T}(\boldsymbol{x})) + \mathcal{N}(0, \sigma^2)\} \tag{2}$$

which perturbs the reported counts $\#_j(\boldsymbol{T}(\boldsymbol{x})) = |\{i : i \in [k], f^i(\boldsymbol{x}) = j\}|$ associated to label $j \in \mathcal{Y}$, via additive Gaussian noise of zero mean and standard deviation σ. The overall approach, called *PATE*, guarantees (ϵ, δ)-differential privacy, with privacy loss scaling with the magnitude of the standard deviation σ and the size of the public dataset

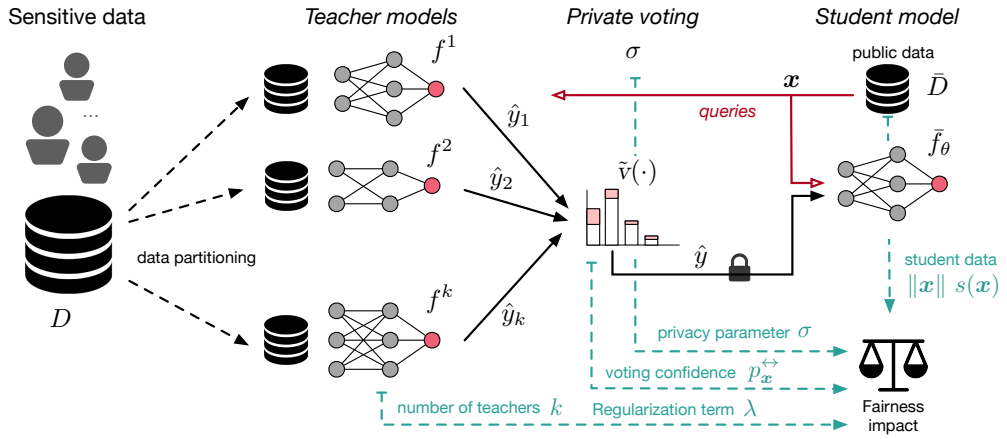

Figure 1: Illustration of PATE and aspects contributing to fairness.

$\bar{D}$ (Papernot et al. 2018). A detailed discussion reviewing the privacy analysis of PATE is reported in Appendix 11. Throughout the paper, the privacy-preserving parameters of the model $\bar{f}$ are denoted with $\tilde{\boldsymbol{\theta}}$.

Fairness The fairness analysis focuses on the notion of *excessive risk* (Wang, Ye, and Xu 2017; Zhang et al. 2017). It defines the difference between the private and non private risk functions:

$$R(S, \boldsymbol{T}) \stackrel{\text{def}}{=} \mathbb{E}_{\tilde{\boldsymbol{\theta}}}\left[\mathcal{L}(\tilde{\boldsymbol{\theta}}; S, \boldsymbol{T})\right] - \mathcal{L}(\hat{\boldsymbol{\theta}}; S, \boldsymbol{T}), \tag{3}$$

where the expectation is defined over the randomness of the private mechanism, S is a subset of $\bar{D}$, and $\tilde{\boldsymbol{\theta}}$ denotes the private student's model parameters while $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \bar{D}, \boldsymbol{T})$. The above definition captures both individual $R(\{\boldsymbol{x}\}, \boldsymbol{T})$ excessive risk for a sample $\boldsymbol{x}$ and group $R(\bar{D}_{\leftarrow a}, \boldsymbol{T})$ excessive risk for a group a, where $\bar{D}_{\leftarrow a}$ denotes the subset of $\bar{D}$ containing exclusively samples whose group attribute is $a \in \mathcal{A}$. This paper uses shorthands $R(\boldsymbol{x})$ and $R(\bar{D}_{\leftarrow a})$ to denote $R(\boldsymbol{x}, \boldsymbol{T})$ and $R(\bar{D}_{\leftarrow a}, \boldsymbol{T})$.

Finally, this paper assumes that the private mechanisms are non-trivial, i.e., they minimize the population-level excessive risk $R(\bar{D})$ and the fairness goal is to minimize excessive risk difference among all individuals and/or groups.

5 PATE Fairness Analysis: Roadmap

The next sections focus on two orthogonal aspects of PATE: the *algorithm's parameters* and the *public student data distribution characteristics* and analyze their fairness impact.

Within the algorithm's parameters, in addition to the privacy variable σ, the paper reveals two surprising aspects which have a direct impact on fairness: The size k of the teacher ensemble and the regularization parameter λ associated with the student risk function. Regarding the public student data's characteristics, the paper shows that the magnitude of the sample input norms $\|\boldsymbol{x}\|$ and the distance of a sample to the decision boundary (denoted $s(\boldsymbol{x})$) play decisive roles to exacerbate the excessive risk induced by the student model. These aspects are illustrated schematically with green dotted lines in Figure 1.

Several aspects of the analysis in this paper rely on the following definition.

Definition 2 (Flipping probability)**.** *Given a data sample $(\boldsymbol{x}, y) \in D$, for an ensemble model $\boldsymbol{T}$ and voting scheme v, the* flipping probability *of $\boldsymbol{T}$ is defined as:*

$$p_{\boldsymbol{x}}^{\leftrightarrow} \stackrel{def}{=} \Pr\left[\tilde{v}(\boldsymbol{T}(\boldsymbol{x})) \neq v(\boldsymbol{T}(\boldsymbol{x}))\right]. \tag{4}$$

It connects the *voting confidence* of the teacher ensemble with the perturbation induced by the privacy-preserving voting scheme, and will be instrumental in the fairness analysis introduced below.

The following sections use several standard datasets including UCI Adults, Credit card, Bank, and Parkinsons (Blake and Merz 1988; Little et al. 2007; Moro, Cortez, and Rita 2014) to support the theoretical claims. The results use feed-forward networks with two hidden layers and nonlinear ReLU activations for both the ensemble and student models. All reported metrics are average of 100 repetitions, used to compute the empirical expectations. When not otherwise stated, the experiments refer to the *Credit card* dataset.

The main paper reports a glimpse of the empirical results, which appears in an extended form in the Appendix (13). Additional description of the dataset and proofs of all theorems are reported in the Appendix.

6 Algorithm's Parameters

This section focuses on analyzing the algorithm's parameters that affect the disparate impact of the student model outputs. In more details, it shows that, in addition to the privacy parameter σ, the regularization term λ of the empirical risk function $\mathcal{L}(\boldsymbol{\theta}, \bar{D}, \boldsymbol{T})$ (see Equation (1)) and the size k of the teacher ensemble $\boldsymbol{T}$ largely control the difference between model learned with noisy and clean labels. The fairness analysis reported in this section assumes that the student model loss $\ell(\cdot)$ is convex and *decomposable*:

Definition 3 (Decomposable function)**.** *A function $\ell(\cdot)$ is* decomposable *if there exists a parametric function $h_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}$, a constant real number c, and a function $z : \mathbb{R} \to \mathbb{R}$, such*

that, for $\boldsymbol{x} \in \mathcal{X}$, and $y \in \mathcal{Y}$:

$$\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = z(h_{\boldsymbol{\theta}}(\boldsymbol{x})) + c\, y\, h_{\theta}(\boldsymbol{x}). \tag{5}$$

Note that a number of loss functions commonly adopted in machine learning, including the logistic loss or the least square loss function, are decomposable (Gao et al. 2016; Patrini et al. 2014). Additionally, while it is common to impose restrictions on the nature of the loss function to render the analysis tractable, our findings are empirically validated on non-linear models.

The following theorem sheds light on the unfairness induced by PATE and the dependency with its parameters. It provides an upper bound on the expected difference between the non-private and private student model parameters. As the paper will show in Theorem 3, this quantity is closely related with the excessive risk. Therein, $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ represent the parameters of student model $\bar{f}$ which are learned as a result of training, respectively, with a clean or noisy voting scheme.

Theorem 1. *Consider a student model $\bar{f}_{\boldsymbol{\theta}}$ trained with a convex and decomposable loss function $\ell(\cdot)$. Then, the expected difference between the private and non-private model parameters is upper bounded as follows:*

$$\mathbb{E}\left[\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|\right] \leq \frac{|c|}{m\lambda}\left[\sum_{\boldsymbol{x}\in\bar{D}} p^{\leftrightarrow}_{\boldsymbol{x}} \|g_{\boldsymbol{x}}\|\right], \tag{6}$$

where c is a real constant and $g_{\boldsymbol{x}} = \max_{\boldsymbol{\theta}} \|\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}(\boldsymbol{x})\|$ represents the maximum gradient norm distortion introduced by a sample $\boldsymbol{x}$. Both c and h are defined as in Equation (5).

The proof relies on λ-strong convexity of the loss function $\mathcal{L}(\cdot)$ (see Appendix 12). Theorem 1 relates the difference in the expected private and non-private student parameters with three key factors: **(1)** the regularization term λ, **(2)** the flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$, and **(3)** the the maximum gradient norm distortion $g_{\boldsymbol{x}}$ induced by a sample $\boldsymbol{x}$. The former two factors are mechanisms-dependent components and the subject of study of this section. As it will be shown next, they are controlled by the size k of the teacher ensemble and the noise parameter σ. The discussion about data dependent components, including those related with the gradient norms is delegated to Section 7.

Throughout the paper, the quantity $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|$ is referred to as *model sensitivity to privacy*, or simply *model sensitivity*, as it captures the effect of the private teacher voting on the student learned model.

The impact of the regularization term λ The first immediate observation of Theorem 1 is that variations of the regularization term λ can reduce or magnify the difference between the private and non-private student model parameters. Since the model sensitivity $\mathbb{E}\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|$ relates directly to the excessive risk (see Theorem 3), the regularization term affects the disparate impact of the privacy-preserving student model. These effects are further illustrated in Figure 2. The figure shows how increasing λ reduces the empirical expected difference between the privacy-preserving and original model parameters $\mathbb{E}\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|$ (left), as well as the excessive risk $R(\bar{D}_{\leftarrow a})$ difference between groups $a = 0$ and $a = 1$ (middle). Note, however, that while larger λ values may reduce the model unfairness, they can hurt the resulting model accuracy, as shown in the right plot. The latter is an intuitive and recognized effect of large regularizers factors.

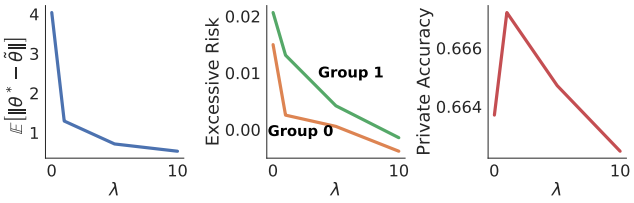

Figure 2: Credit-card dataset with $\sigma = 50, k = 150$. Model sensitivity (left), empirical risk (middle), and model accuracy (right) as a function of the regularization term.

The impact of the teachers ensemble size k The second aspect considered in this section is the relation between the ensemble size k and the resulting private model fairness. The following result relates the size of the ensemble with its voting confidence.

Theorem 2. *For a sample $\boldsymbol{x} \in \bar{D}$ assume that the teacher models outputs $f^i(\boldsymbol{x})$ are in agreement for all $i \in [k]$. Then, the flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ is given by:*

$$p^{\leftrightarrow}_{\boldsymbol{x}} = 1 - \Phi\left(\frac{k}{\sqrt{2}\sigma}\right), \tag{7}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, and σ is the standard deviation in the Gaussian mechanism.

The proof is based on the properties of independent Gaussian random variables.

The analysis above sheds light on the outcome of the teachers voting scheme and its relation with the ensemble size k (as well as the privacy parameter σ). It shows that larger k values correspond to smaller flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$. Combined with Theorem 1, the result suggests that the difference between the private and non-private model parameters is inversely proportional to k.

While for simplicity of analysis Theorem 2 requires the decision of all teachers to agree on a given sample $\boldsymbol{x}$, our empirical analysis supports this result for the more general scenario where different teachers have different agreements on a sample. Figure 3 (left) illustrates the relation between the number k of teachers and the flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ of the ensemble. The plot shows a clear trend indicating that larger ensembles result in smaller flipping probabilities.

Next, analogously to what is reported in Figure 2, Figure 4 shows that increasing k reduces the difference in the expected private and non-private model parameters (left), reduces the group excessive risk difference (middle), and increases the model $\bar{f}$ accuracy (right). However, similarly as for the regularization term λ, there is also a downside of using very large ensembles: large values k can reduce the accuracy of the (private and non-private) models. While studying these tradeoffs goes beyond the scope of this work, we believe this behavior is related with the bias-variance tradeoff imposed on the growing ensemble: The larger the ensemble

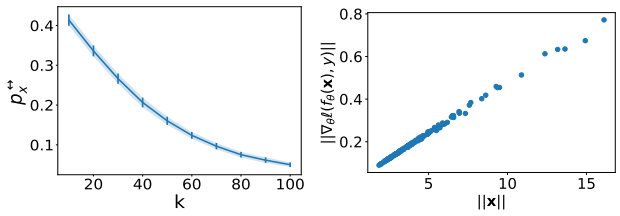

Figure 3: Credit-card dataset: Average flipping probability $p_{\boldsymbol{x}}^{\leftrightarrow}$ for samples $\boldsymbol{x} \in \bar{D}$ as a function of the ensemble size k (left) and relation between gradient and input norms (right).

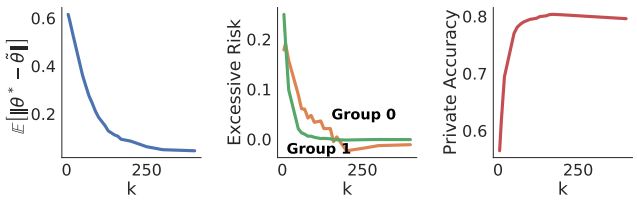

Figure 4: Credit card dataset with $\sigma = 50$, $\lambda = 100$. Expected model sensitivity (left), empirical risk (middle), and model accuracy (right) as a function of the ensemble size.

the less data each teacher is given to train their models, thus affecting their voting accuracy. We believe this is an interesting and important direction for future work.
This section concludes with a useful corollary of Theorem 1.

Corollary 1 (Theorem 1)**.** *Let $\bar{f}_\theta$ be a* logistic regression *classifier. Its expected model sensitivity is upper bounded as:*

$$\mathbb{E}\left[\|\overset{\star}{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|\right] \leq \frac{1}{m\lambda}\left[\sum_{\boldsymbol{x}\in\bar{D}} p_{\boldsymbol{x}}^{\leftrightarrow}\|\boldsymbol{x}\|\right]. \tag{8}$$

The result above highlights several interesting points. First, in logistic regression, samples with large input norms can have a non negligible impact on fairness. This place emphasis on an nontrivial aspect of the student data properties which may affect fairness and is subject of study of the next section. Second, it indicates the presence of a relation between gradient norms and input norms, which is further highlighted in Figure 3 (right). The plot illustrates the strong correlation between input norms and their associated gradient norms.

7 Student's Data Properties

Having examined the algorithmic properties of PATE affecting fairness, this section turns on analyzing a set of properties concerning the student data which regulate the disproportionate impacts of the algorithm. The subsequent set of results shows that the norms of the student's data samples and their distance to the decision boundary are two key factor tied to the exacerbation of excessive risk in PATE.

The following is a corollary of Theorem 1 and bounds the second order statistics of the model sensitivity to privacy.

Corollary 2 (Theorem 1)**.** *Given the same settings and assumption of Theorem 1, it follows:*

$$\mathbb{E}\left[\|\overset{\star}{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|^2\right] \leq \frac{|c|^2}{m\lambda^2}\left[\sum_{\boldsymbol{x}\in\bar{D}} p_{\boldsymbol{x}}^{\leftrightarrow 2}\|g_{\boldsymbol{x}}\|^2\right]. \tag{9}$$

Note that as similarly shown by Corollary 1, when $\bar{f}_\theta$ is a logistic regression model, the gradient norm $\|g_{\boldsymbol{x}}\|$ in Equation (9) can be substituted with the input norm $\|\boldsymbol{x}\|$.

The result above is useful to derive an upper bound on the excessive risk, as illustrated in the following theorem.

Theorem 3. *Let $\ell(\cdot)$ be a $\beta_{\boldsymbol{x}}$-smooth loss function. The excessive risk $R(\boldsymbol{x})$ of a sample $\boldsymbol{x}$ is upper bounded as:*

$$R(\boldsymbol{x}) \leq \|\nabla_{\overset{\star}{\boldsymbol{\theta}}}\ell(\bar{f}_{\overset{\star}{\boldsymbol{\theta}}}(\boldsymbol{x}), y)\|U_1 + \frac{1}{2}\beta_x U_2, \tag{10}$$

where, $U_1 = \mathbb{E}\left[\|\overset{\star}{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|\right]$ and $U_2 = \mathbb{E}\left[\|\overset{\star}{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|^2\right]$ capture the first and second order statistics of the model sensitivity.

The proof of the above theorem relies on Theorem 1 and Corollary 2, which provide bounds for the first and second order statistics of the model sensitivity, and on the properties of smooth functions.

Theorem 3 provides an upper bound on the (individual) excessive risk. It shows the presence of three central factors controlling this excessive risk: **(1)** *the gradient norm* $\|\nabla_{\overset{\star}{\boldsymbol{\theta}}}\ell(\bar{f}_{\overset{\star}{\boldsymbol{\theta}}}(\boldsymbol{x}), y)\|$ for a sample $\boldsymbol{x}$, **(2)** *the smoothness parameter* $\beta_{\boldsymbol{x}}$ associated with a sample $\boldsymbol{x}$, and **(3)** the *model sensitivity* (captured by terms U_1 and U_2). As the paper shows next, these seemingly unrelated factors are controlled *indirectly* by two key data aspects: the samples *input norms* and their *distance to the decision boundary*.

The rest of the section focuses on logistic regression models, however, as our experimental results illustrate, the observations extend to complex nonlinear models as well.

The impact of the data input norms First notice that the norm $\|\boldsymbol{x}\|$ of a sample $\boldsymbol{x}$ strongly influences the quantities U_1 and U_2, as already observed by Corollary 1. This aspect is further illustrated in Figure 5 (left), which shows a strong correlation between the input norms and the expected model sensitivity. Thus, samples with higher input norms may have a nontrivial impact to the model sensitivity and, in turn, to its unfairness.

Next, the following proposition sheds light on the relation between the norm of a sample $\boldsymbol{x}$ and its associated gradient norm $\|\nabla_{\overset{\star}{\boldsymbol{\theta}}}\ell(\bar{f}_{\overset{\star}{\boldsymbol{\theta}}}(\boldsymbol{x}), y)\|$.

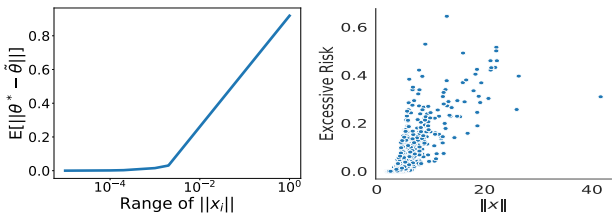

Figure 5: Credit-card data: Relation between input norms and model sensitivity (left) and Spearman correlation between input norms and excessive risk (right).

Proposition 1. *Let $\bar{f}_{\boldsymbol{\theta}}$ be a logistic regression binary classifier with cross entropy loss function. For a given sample $(\boldsymbol{x}, y) \in \bar{D}$, the gradient $\nabla^{\star}_{\boldsymbol{\theta}} \ell(\bar{f}^{\star}_{\boldsymbol{\theta}}(\boldsymbol{x}), y)$ is given by:*

$$\nabla^{\star}_{\boldsymbol{\theta}} \ell(\bar{f}^{\star}_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = \left(\bar{f}^{\star}_{\boldsymbol{\theta}}(\boldsymbol{x}) - y\right) \otimes \boldsymbol{x},$$

where $\otimes$ expresses the Kronecker product.

Recall that gradient norms have a proportional effect on the upper bound of the excessive risk (Equation (10)). Thus, the relation above sheds further light on the weight that samples with large norms may have in controlling their associated excessive risk, as shown in Figure 5 (right), which shows the Spearman correlation between these two quantities.

Finally, the discussion notes that the smoothness parameter $\beta_{\boldsymbol{x}}$ captures the local flatness of the loss function at a point $\boldsymbol{x}$. A derivation of $\beta_{\boldsymbol{x}}$ for logistic regression classifier is provided below.

Proposition 2. *Consider again a binary logistic regression as in Proposition 1. The smoothness parameter $\beta_{\boldsymbol{x}}$ for a sample $\boldsymbol{x}$ is given by (Shi et al. 2021): $\beta_{\boldsymbol{x}} = 0.25\|\boldsymbol{x}\|^2$.*

The above clearly illustrates the relationship between input norms $\|\boldsymbol{x}\|$ and the smoothness parameters $\beta_{\boldsymbol{x}}$.

To summarize, propositions 1 and 2 illustrate that individuals $\boldsymbol{x}$ with large (small) input norms tends to have large (small) gradient norm and smoothness parameters, thus controlling the model sensitivity and, in turn, the excessive risk $R(\boldsymbol{x})$. An extended analysis of the above claim is provided in Appendix 13.

The impact of the distance to decision boundary As mentioned in the previous section, the flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$associated with a sample $\boldsymbol{x} \in \bar{D}$ directly controls the model sensitivity $\mathbb{E}[\|\overset{\star}{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|]$. Beside the discussed factors, this section further studies which characteristics of sample $\boldsymbol{x}$ can causes it to have a high flipping probability.

Intuitively, samples close to the decision boundary are associated to small ensemble voting confidence and vice-versa. To illustrate this intuition the paper borrows the concept of *closeness to the decision boundary* from Tran, Dinh, and Fioretto (2021).

Definition 4 (Closeness to decision boundary)**.** *Let $f_{\boldsymbol{\theta}}$ be a C-classes classifier trained using data $\bar{D}$ with its true labels. The closeness to the decision boundary $s(\boldsymbol{x})$ is defined as:*

$$s(\boldsymbol{x}) \overset{def}{=} 1 - \sum_{c=1}^{C} f_{\boldsymbol{\theta},c}(\boldsymbol{x})^2,$$

where $f_{\boldsymbol{\theta},c}$ denotes the softmax probability for class c.

The above, (together with Theorem 5 of (Tran, Dinh, and Fioretto 2021)) illustrate that large (small) $s(\boldsymbol{x})$ values are associated to close (distant) projections of point $\boldsymbol{x}$ to the model decision boundary. The concept of closeness to decision boundary gives a way to indirectly quantify the flipping probability of a sample. Empirically, the correlation between the distance to decision boundary of sample $\boldsymbol{x}$ and its flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ is illustrated in Figure 6 (left). The plots are once again generated using a neural network with nonlinear objective and the relation holds for all datasets analyzed. Notice the strong positive correlation between these two quantities. The plot indicates that the samples that are close to the decision boundary will have a higher probability of "flipping" their label, thus resulting in a worse excessive risk. Finally, the proportional effect of the flipping probability on the excessive risks is illustrated in Figure 6 (right).

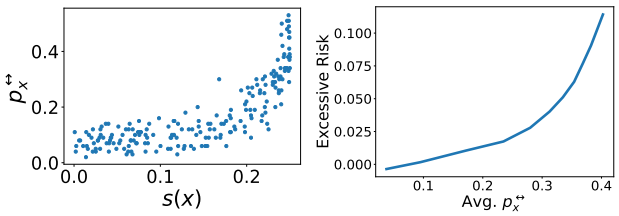

Figure 6: Credit-card data: Spearman correlation between the closeness to decision boundary $s(\boldsymbol{x})$ and the flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ (left) and relation between input norms and excessive risk (right).

8 Mitigation solution

The previous sections highlighted the presence of several algorithmic and data-related factors which affect the disparate impact of the student model. A common role of these factors was their effects on the model sensitivity $\mathbb{E}\|\overset{\star}{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|$ which, in turn, is related with the excessive risk of different groups, whose difference we would like to minimize.

Motivated by these observations, this section proposes a mitigating strategy that aims at reducing the sensitivity of the private model parameters. To do so, the paper exploits the idea of using *soft labels* (as defined below). When using the traditional voting process (denoted *hard labels* in this section), in low voting confidence regimes a small noise perturbation may significantly affect the result of the voting scheme. Consider, for example the case of a binary classifier where for a sample $\boldsymbol{x}$, $k/2 + 1$ teachers vote for label 0 and $k/2 - 1$ for label 1 for some even ensemble size k. When a perturbations are induced to these counts to guarantee privacy, the process can report the incorrect label ($\hat{y} = 1$) with high probability. As a result, the private student model parameters obtained from private training with hard labels can be sensitive to the noisy voting, and may deviate significantly from the non-private one. This issue can be partially addressed by the introduction of soft labels:

Definition 5 (Soft label)**.** *The* soft label *of a sample $\boldsymbol{x}$ is:*

$$\boldsymbol{\alpha}(\boldsymbol{x}) = \left(\frac{\#_c(\boldsymbol{T}(\boldsymbol{x}))}{k}\right)_{c=1}^{C}$$

and their privacy-preserving counterparts:

$$\tilde{\boldsymbol{\alpha}}(\boldsymbol{x}) = \left(\frac{\#_c(\boldsymbol{T}(\boldsymbol{x})) + \mathcal{N}(0, \sigma^2)}{k}\right)_{c=1}^{C}.$$

To exploit soft labels, the training step of the student model is altered to use the following loss function:

$$\ell'(\hat{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \tilde{\boldsymbol{\alpha}}) = \sum_{c=1}^{C} \tilde{\alpha}_c \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), c), \tag{11}$$

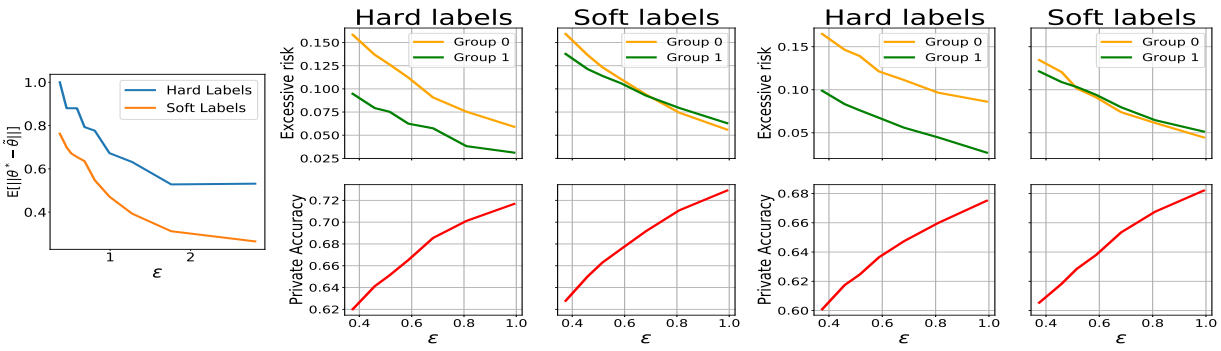

Figure 7: Training privately PATE with hard and soft labels: Model sensitivity at varying of the privacy loss (left) on Parkinson dataset and excessive risk at varying of the privacy loss for Bank (middle) and Parkinson (right) datasets.

which can be considered as a weighted version of the original loss function $\ell(\hat{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), c)$ on class label c, whose weight is its confidence $\tilde{\alpha}_c$. Note that $\ell'(\hat{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \tilde{\boldsymbol{\alpha}}) = \ell(\hat{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))$ when all teachers in the ensemble chose the same label. The privacy analysis for this model is similar that of classical PATE and is reported in Appendix 11.

The effectiveness of this scheme is demonstrated in Figure 7. The experiment settings are reported in details in the Appendix and reflect those described at the end of Section 5. The left subplot shows the relation between the model sensitivity $\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}^{\star} - \tilde{\boldsymbol{\theta}}\|\right]$ at varying levels of the privacy loss ϵ (dictated by the noise level σ). Notice how the student models trained using soft labels reduce their sensitivity to privacy when compared to the counterparts that use hard labels.

The middle and right plots of Figure 7 illustrate the effects of the proposed mitigating solution in terms of utility/fairness tradeoff on the private student model. The top subplots illustrate the group excessive risks $R(\bar{D}_{\leftarrow 0})$ and $R(\bar{D}_{\leftarrow 1})$ associated with minority (0) and majority (1) groups while the bottom subplot illustrate the accuracy of the model, at increasing of the privacy loss ϵ. Notice how soft labels can reduce the disparate impacts in private training (top), which consistently reduces the difference in excessive risks between two groups, suggesting an improvement in fairness. Finally, notice that while fairness is improved there is seemingly no cost in accuracy. On the contrary, using soft labels produces comparable or better models to the counterparts produced with the hard labels.

Additional experiments, including illustrating the behavior of the mitigating solution at varying of the number k of teachers are reported in the appendix and the general message is consistent with what described above. Finally, an important benefit about the proposed solution is that it *does not* require the protected group information ($a \in \mathcal{A}$) to be part of the training data. Thus, it is applicable in challenging situations when it is not feasible to collect or use protected features (e.g., under GDPR (Lahoti et al. 2020)).

These results are significant. They suggest that this mitigating solution can be an effective strategy for improving the disparate impact of private model ensembles without sacrificing accuracy.

9 Discussion

We note that the proposed mitigating solution relates to concepts explored in robust machine learning. In particular, Papernot et al. (2016) noted that training a classifiers with soft labels can increase its robustness against adversarial samples. This connection is not coincidental. Indeed, the model sensitivity is affected by the voting outcomes of the teacher ensemble (Theorems 1 and 3). Similarly to robust ML models being insensitive to input perturbations, strongly agreeing ensemble will be less sensitive to noise and viceversa. This observation raises a question about the connection of robustness and fairness in private models (Yurochkin, Bower, and Sun 2019). We believe that this connection is an important direction for the private ML community.

Finally, we notice that the use of more advanced voting schemes, such as the interactive GNMAX (Papernot et al. 2018), may produce different fairness results. While this is an interesting avenue for extending our analysis, sophisticated voting schemes may introduce sampling bias (e.g., interactive GNMAX may exclude samples with low ensemble voting agreement). Such bias may trigger some nontrivial unfairness issues on its own.

10 Conclusions

This work was motivated by the recent observations regarding the effects of differential privacy to the disparate impacts of machine learning models. The paper introduced a notion of fairness that relies on the concept of excessive risk and analyzed this notion in the Private Aggregation of Teacher Ensembles (PATE) (Papernot et al. 2018), an important privacy-preserving machine learning framework used in semisupervised settings or when one wishes to protect the data labels. This paper isolated key components related with the algorithms parameters and the public training data characteristics which are responsible for exacerbating the disparate impacts, it studied the factors affecting these components, and introduced a mitigation solution.

Given the increasing presence of privacy-preserving data-driven algorithms in consequential decisions, we believe that this work may represents an important and broadly applicable step toward understanding the sources of disparate impacts observed in differentially private learning systems.

References

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, 15479–15488.

Blake, C.; and Merz, C. 1988. UCI repository of machine learning databases.

Carcillo, F.; Le Borgne, Y.-A.; Caelen, O.; Kessaci, Y.; Oblé, F.; and Bontempi, G. 2019. Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection.

Cummings, R.; Gupta, V.; Kimpara, D.; and Morgenstern, J. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 309–315.

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

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Ekstrand, M. D.; Joshaghani, R.; and Mehrpouyan, H. 2018. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, 35–47.

Farrand, T.; Mireshghallah, F.; Singh, S.; and Trask, A. 2020. Neither Private Nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 15–19.

Gao, W.; Wang, L.; Zhou, Z.-H.; et al. 2016. Risk minimization in the presence of label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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*.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. H. 2020. Fairness without Demographics through Adversarially Reweighted Learning. arXiv:2006.13114.

Little, M.; Mcsharry, P.; Roberts, S.; Costello, D.; and Moroz, I. 2007. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *Biomedical engineering online*, 6: 23.

Mironov, I. 2017. Rényi Differential Privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*.

Moro, S.; Cortez, P.; and Rita, P. 2014. A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.*, 62: 22–31.

Mozannar, H.; Ohannessian, M. I.; and Srebro, N. 2020. Fair Learning with Private Demographic Data. In *Proceedings of the 37th International Conference on Machine Learning*.

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; and Swami, A. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, 582–597. IEEE.

Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Erlingsson, U. 2018. Scalable Private Learning with PATE.

Patrini, G.; Nock, R.; Rivera, P.; and Caetano, T. 2014. (Almost) no label no cry. *Advances in Neural Information Processing Systems*, 27: 190–198.

Pujol, D.; McKenna, R.; Kuppam, S.; Hay, M.; Machanavajjhala, A.; and Miklau, G. 2020. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 189–199.

Sadowski, P. 2021. Lecture Notes: Notes on Backpropagation. URL: https://www.ics.uci.edu/~pjsadows/notes.pdf. Last visited on 2021/05/01.

Shalev-Shwartz, S. 2007. Online Learning: Theory, Algorithms, and Applications.

Shi, Z.; Loizou, N.; Richtárik, P.; and Takáč, M. 2021. AI-SARAH: Adaptive and Implicit Stochastic Recursive Gradient Methods. arXiv:2102.09700.

Tran, C.; Dinh, M. H.; and Fioretto, F. 2021. Differentially Private Deep Learning under the Fairness Lens. arXiv:2106.02674.

Tran, C.; Fioretto, F.; and Hentenryck, P. V. 2021. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 9932–9939. AAAI Press.

Tran, C.; Fioretto, F.; Hentenryck, P. V.; and Yao, Z. 2021. Decision Making with Differential Privacy under a Fairness Lens. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 560–566.

Uniyal, A.; Naidu, R.; Kotti, S.; Singh, S.; Kenfack, P.; Mireshghallah, F.; and Trask, A. 2021. DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy?

Wang, D.; Ye, M.; and Xu, J. 2017. Differentially Private Empirical Risk Minimization Revisited: Faster and More General. In *Advances in Neural Information Processing Systems*.

Xu, D.; Du, W.; and Wu, X. 2021. Removing Disparate Impact on Model Accuracy in Differentially Private Stochastic Gradient Descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, 1924–1932. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383325.

Yurochkin, M.; Bower, A.; and Sun, Y. 2019. Training individually fair ML models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*.

Zhang, J.; Zheng, K.; Mou, W.; and Wang, L. 2017. Efficient Private ERM for Smooth Objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 3922–3928.