A FAIRNESS ANALYSIS ON DIFFERENTIALLY PRIVATE AGGREGATION OF TEACHER ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Private Aggregation of Teacher Ensembles (PATE) 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 semi-supervised settings or when one wishes to protect the data labels.

This paper asks whether this privacy-preserving framework introduces or exacerbates 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 observed disproportionate impacts, why these aspects are affecting different groups disproportionately, and proposes guidelines to largely mitigate these effects.

Code, additional experiments, and theorems' proofs are reported in the Appendix.

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. Two key concerns for the adoption of these systems regard 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. In the context of machine learning, DP ensures that algorithms can learn the relations between data and predictions while preventing them from memorizing sensitive information about any specific individual in the training data. While this property is appealing, it was recently observed that DP systems may induce biased and unfair outcomes for different groups of individuals (Bagdasaryan et al., 2019; Xu et al., 2021). The resulting outcomes can have significant societal impacts on the involved individuals: classification errors may penalize some groups over others in important decisions including criminal assessment, landing, and hiring or can result in disparities regarding the allocation of critical funds and benefits (Pujol et al., 2020; Fioretto et al., 2021). *While these surprising observations have become apparent in several contexts, 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 a state-of-the-art privacy-preserving ML framework called *Private Aggregation of Teacher Ensembles* (PATE) (Papernot et al., 2018). It combines multiple agnostic 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.The paper analyzes which algorithmic and data properties are responsible for the disproportionate impacts, why these aspects are affecting different groups of individuals disproportionately, and proposes a solution to mitigate these effects.

In summary, the paper makes the following contributions: **(1)** It uses the concept of excess risk to define a notion of fairness that generalizes accuracy parity and measures the direct impact of privacy to the model outputs for different groups of individuals. **(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 that are responsible for the observed disparate impacts. **(4)** It studies when and why these components affect different groups disproportionately during private training. **(5)** Finally, based on these findings, it proposes a method that may aid in mitigating these unfairness effects while retaining high accuracy.

Given the empirical advantages of privacy-preserving ensemble models with respect to other frameworks like DP-SGD (Abadi & et al., 2016; Ghazi et al., 2021; Uniyal et al., 2021), we believe that this work may represents an important and broadly applicable step toward understanding and mitigating the disparate impacts observed in semi-supervised private learning systems.

2 Related Work

The study of the disparate impacts caused by privacy-preserving algorithms has recently seen several important developments. Ekstrand et al. (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, which culminated in a variety of proposals, including (Jagielski et al., 2018; Mozannar et al., 2020; Tran et al., 2021b).

Pujol et al. (2020) were the first to show, empirically, that decision tasks made using DP datasets may disproportionately affect some groups of individuals over others. These studies were complemented theoretically by Tran et al. (2021c). Similar observations were also made in the context of model learning. Bagdasaryan et al. (2019) empirically observed that the accuracy of a DP model trained using DP-Stochastic Gradient Descent (DP-SGD) decreased disproportionately across groups causing larger negative impacts to the underrepresented groups. Farrand et al. (2020) and Uniyal et al. (2021) reached similar conclusions and showed that this disparate impact was 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, an important privacy-enhancing ML system, as well as introduces mitigating guidelines.

3 Preliminaries: Differential Privacy

Differential privacy (DP) is a strong privacy notion stating 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 (Dwork et al. (2006))**.** *A mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-DP, 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 et al., 2014).

4 Problem Settings and Goals

This paper considers a *private* dataset D consisting of n individuals' data $(\boldsymbol{x}_i, y_i)$, with $i \in [n]$, drawn i.i.d. from an unknown distribution Π. Therein, $\boldsymbol{x}_i \in \mathcal{X}$ is a sensitive feature vector containing 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

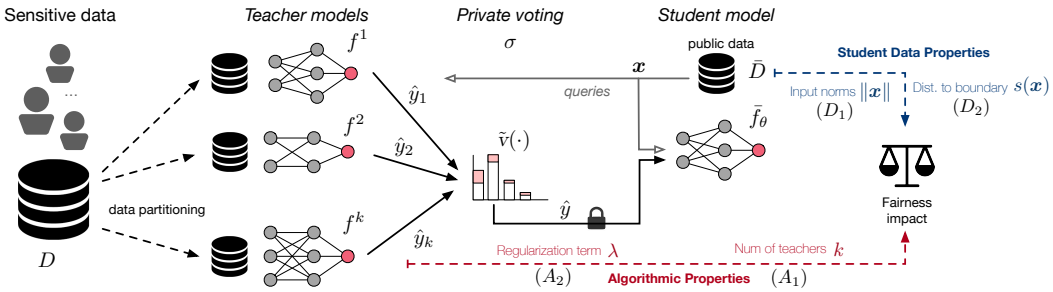

Figure 1: Illustration of PATE and aspects contributing to fairness.

classifier that needs to predict criminal defendant's recidivism. The data features $\boldsymbol{x}_i$ may describe the individual's demographics, education, and crime committed, the protected attribute $\boldsymbol{a}_i$ may describe the individual's gender or ethnicity, and y_i whether the individual has high risk to reoffend.

This paper studies the fairness implications arising when training private semi-supervised transfer learning models. The setting is depicted in Figure 1. We are given an ensemble of *teacher* models $\boldsymbol{T} = \{f^j\}_{j=1}^k$, with each $f^j : \mathcal{X} \to \mathcal{Y}$ trained on a non-overlapping portion D_i of D. This ensemble is used to transfer knowledge to a *student* model $\bar{f}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, where $\boldsymbol{\theta}$ is 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. We focus on learning classifier $\bar{f}_{\boldsymbol{\theta}}$ 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 with loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \bar{D}, \boldsymbol{T}) + \lambda \|\boldsymbol{\theta}\|^2 = \sum_{\boldsymbol{x} \in \bar{D}} \ell\left(\bar{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), v(\boldsymbol{T}(\boldsymbol{x}))\right) + \lambda \|\boldsymbol{\theta}\|^2, \tag{1}$$

where $v : \mathcal{Y}^k \to \mathcal{Y}$ 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 $\{f^j(\boldsymbol{x})\}_{j=1}^k$, and $\lambda > 0$ is a regularization term.

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

Privacy. *Privacy* is achieved by using a DP version $\tilde{v}$ of the voting function v:

$$\tilde{v}(\boldsymbol{T}(\boldsymbol{x})) = \arg\max_c \{\#_c(\boldsymbol{T}(\boldsymbol{x})) + \mathcal{N}(0, \sigma^2)\} \tag{2}$$

which perturbs the reported counts $\#_c(\boldsymbol{T}(\boldsymbol{x})) = |\{j : j \in [k], f^j(\boldsymbol{x}) = c\}|$ for class $c \in C$ with zero-mean Gaussian and standard deviation σ. The overall approach, called *PATE* (Papernot et al., 2018), guarantees (ϵ, δ)-DP, with privacy loss scaling with the magnitude of the standard deviation σ and the size of the public dataset $\bar{D}$. A detailed review of the privacy analysis of PATE is reported in Appendix B. Throughout the paper, the privacy-preserving parameters of the model $\bar{f}$ trained with noisy voting $\tilde{v}(\boldsymbol{T}(\boldsymbol{x}))$ are denoted with $\tilde{\boldsymbol{\theta}}$.

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

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

where the expectation is 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 $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \bar{D}, \boldsymbol{T}) + \lambda \|\boldsymbol{\theta}\|^2$. In particular, the paper focuses on measuring the excess risk $R(\bar{D}_{\leftarrow a}, \boldsymbol{T})$ for groups $a \in \mathcal{A}$, where $\bar{D}_{\leftarrow a}$ denotes the subset of $\bar{D}$ containing exclusively samples from group a. *This notion captures the (unintended) impact of privacy on the task accuracy for a given group.*
This paper uses shorthand $R(\bar{D}_{\leftarrow a})$ to denote $R(\bar{D}_{\leftarrow a}, \boldsymbol{T})$ and assumes that the private mechanisms are non-trivial, i.e., they minimize the population-level excess risk $R(\bar{D})$.

Fairness is measured as the highest excess risk difference among all groups:

$$\xi(\bar{D}) = \max_{a,a' \in \mathcal{A}} R(\bar{D}_{\leftarrow a}) - R(\bar{D}_{\leftarrow a'}). \tag{4}$$

Notice how the above relates with the notion of accuracy parity (Bagdasaryan et al., 2019), which measures the disparity of the task accuracy across groups. This is the case when the adopted loss ℓ is a hard (0/1) loss. All the experiments reported in the paper will, in fact, use such a 0/1-loss, while, the theoretical analysis uses general differentiable loss functions.

5 PATE Fairness Analysis: Roadmap

The goal of this paper is to identify what induces unfairness in PATE and why. The next sections isolate these key factors which will be categorized as *algorithm's parameters* and the *public student data characteristics*. The theoretical analysis assumes that, for a group $a \in \mathcal{A}$, the group loss function $\mathcal{L}(\boldsymbol{\theta}; D_{\leftarrow a}, \boldsymbol{T})$ is convex and β_a-smooth, for some $\beta_a \geq 0$, w.r.t. the model parameters $\boldsymbol{\theta}$. The evaluation, however, does not restrict the form of the loss function. A detailed description of the experimental settings is reported in Appendix C and proofs of all theorems in Appendix A.

A fairness bound. We start by introducing a bound on the model disparity, which will be key to pinpoint the algorithm's and data characteristics responsible to exacerbate unfairness in PATE. Throughout the paper, we refer to the quantity $\Delta_{\tilde{\boldsymbol{\theta}}} \stackrel{\text{def}}{=} \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$ as to *model deviation due to privacy*, or simply *model deviation*, as it captures the effect of the private teacher voting on the student learned model. Therein, $\boldsymbol{\theta}^*$ 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. *The model fairness is upper bounded as:*

$$\xi(\bar{D}) \leq 2 \max_a \|G_a\| \, \mathbb{E}\left[\Delta_{\tilde{\boldsymbol{\theta}}}\right] + 1/2 \max_a \beta_a \, \mathbb{E}\left[\Delta^2_{\tilde{\boldsymbol{\theta}}}\right], \tag{5}$$

where $G_a = \mathbb{E}_{\boldsymbol{x} \sim \bar{D}_{\leftarrow a}}\left[\nabla_{\boldsymbol{\theta}^*} \ell(\bar{f}_{\boldsymbol{\theta}^*}(\boldsymbol{x}), y)\right]$ *is the gradient of the group loss evaluated at* $\boldsymbol{\theta}^*$*, and* $\Delta_{\tilde{\boldsymbol{\theta}}}$ *and* $\Delta^2_{\tilde{\boldsymbol{\theta}}}$ *capture the first and second order statistics of the model deviation.*

The above illustrates that the model unfairness is proportionally regulated by three direct factors: **(1)** the model deviation $\Delta_{\tilde{\boldsymbol{\theta}}}$, **(2)** the largest group's gradient norm $\max_a \|G_a\|$, and **(3)** the largest group's smoothness parameter $\max_a \beta_a$.

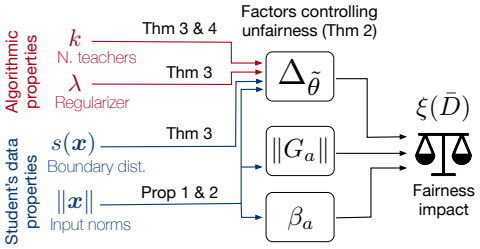

Figure 2: Factors impacting PATE fairness.

The paper delves into which **A**lgorithms' parameters and **D**ata characteristics affect these factors, and thus, infer with the model unfairness. Within the **A**lgorithm's parameters, in addition to the privacy variable ϵ (captured by the noise parameter σ), the paper reveals two interesting factors having a direct impact on fairness: ($\boldsymbol{A_1}$) the regularization term λ associated with the student risk function and ($\boldsymbol{A_2}$) the size k of the teachers' ensemble. Regarding the public student **D**ata's characteristics, the paper shows that ($\boldsymbol{D_1}$) the magnitude of the sample input norms $\|\boldsymbol{x}\|$ and ($\boldsymbol{D_2}$) the distance of a sample to the decision boundary (denoted $s(\boldsymbol{x})$) play decisive roles to exacerbate the excess risks induced by the student model. A schematic illustration of the relations between these factors and how they impact the model fairness is provided in Figure 2.

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

Definition 2. *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:*

$$p^{\leftrightarrow}_{\boldsymbol{x}} \stackrel{def}{=} \Pr\left[\tilde{\mathrm{v}}(\boldsymbol{T}(\boldsymbol{x})) \neq \mathrm{v}(\boldsymbol{T}(\boldsymbol{x}))\right].$$

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

Finally, the theoretical claims reported in the next sections are supported and complemented by empirical evidence on both tabular datasets (UCI Adults, Credit card, Bank, and Parkinsons) and

image dataset (UTKFace). These results use feed-forward networks with two hidden layers and nonlinear ReLU activations for both the ensemble and student models for tabluar data and CNNs for image data. All reported metrics are average of 100 repetitions, used to compute the empirical expectations, and report 0/1 losses which *capture the notion of accuracy parity*. The paper reports a glimpse of the empirical results, with the purpose of supporting the theoretical claims, and extended experiments and additional description of the dataset are reported in Appendix C.

6 Algorithm's Parameters

This section analyzes the algorithm's parameters that affect the disparate impact of the student model outputs. The fairness analysis reported in this section assumes that the student model loss $\ell(\cdot)$ is convex and *decomposable*:

Definition 3. *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 $y \in \mathcal{Y}$:*

$$\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = z(h_{\boldsymbol{\theta}}(\boldsymbol{x})) + c\, y\, h_{\boldsymbol{\theta}}(\boldsymbol{x}). \tag{6}$$

A number of loss functions commonly adopted in ML, including the logistic loss (used in our experiments) or the least square loss function, are decomposable (Patrini et al., 2014). Additionally, while restrictions are commonly imposed on the loss functions to render the analysis tractable, our findings are empirically validated on non-linear models.

Recall that the model deviation is a central factor proportionally controlling the unfairness of PATE (Theorem 1). We now provide a useful bound on such quantity and highlight its relations with key algorithms parameters.

Theorem 2. *Consider a student model $\bar{f}_{\boldsymbol{\theta}}$ trained with a convex and decomposable loss function $\ell(\cdot)$. Then, the first order statistics of the model deviation is upper bounded as:*

$$\mathbb{E}\left[\Delta_{\bar{\boldsymbol{\theta}}}\right] \leq \frac{|c|}{m\lambda}\left[\sum_{\boldsymbol{x}\in\bar{D}} p^{\leftrightarrow}_{\boldsymbol{x}} \|G^{\max}_{\boldsymbol{x}}\|\right], \tag{7}$$

where c is a real constant and $G^{\max}_{\boldsymbol{x}} = \max_{\boldsymbol{\theta}} \|\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}(\boldsymbol{x})\|$ represents the maximum gradient norm distortion introduced by a sample $\boldsymbol{x}$. Both c and h are defined as in Equation 6.

The proof relies on λ-strong convexity of the loss function $\mathcal{L}(\cdot) + \lambda\|\boldsymbol{\theta}\|$ (see Appendix A) and its tightness is reported empirically in Appendix C.2. Theorem 2 uncovers how the student model changes due to privacy and relates it with two mechanism-dependent components: **(1)** the regularization term λ of the empirical risk function $\mathcal{L}(\boldsymbol{\theta}, \bar{D}, \boldsymbol{T})$ (see Equation 1), and **(2)** the flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$, which, as it will be shown later, is strongly controlled by the size k of the teacher ensemble. These mechanisms-dependent components and the subject of study of this section. The discussion about data dependent components, including those related with the maximum gradient norm distortion $G^{\max}_{\boldsymbol{x}}$ is delegated to Section 7.

A_1: The impact of the regularization term λ. The first immediate observation of Theorem 2 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 deviation $\mathbb{E}[\Delta_{\bar{\boldsymbol{\theta}}}]$ relates directly to the fairness goal (see first term of RHS of Equation 5 in Theorem 1) *the regularization term affects the disparate impact of the privacy-preserving student model*. These effects are further illustrated in Figure 3 (top). The figure shows how increasing λ reduces the expected difference between the privacy-preserving and original model parameters $\mathbb{E}[\Delta_{\bar{\boldsymbol{\theta}}}]$ (left), as well as the excess 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 model's accuracy, as shown in the right plot. The latter is an intuitive and recognized effect of large regularizers (Mahjoubfar et al., 2017).

A_2: The impact of the teachers ensemble size k. Next we consider the relation between the ensemble size k and the resulting private model's fairness. The following result relates the size of the ensemble with its voting confidence.

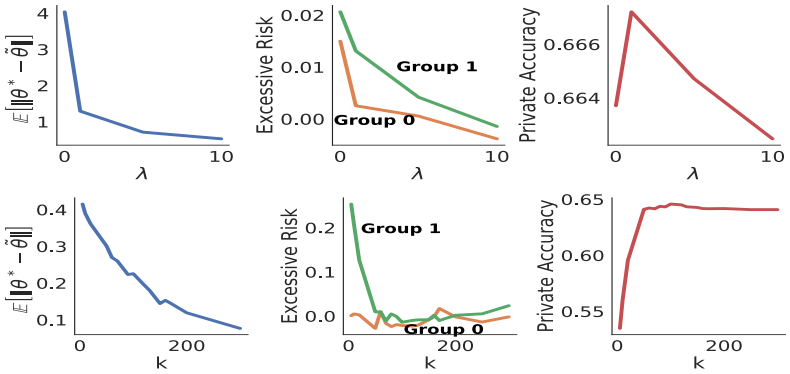

Figure 3: Credit card dataset with $\sigma = 50, k = 150$ (top) and $\lambda = 100$ (bottom). Expected model deviation (left), excess risk (middle), and model accuracy (right) as a function of the regularization term (top) and ensemble size (bottom).

Theorem 3. *For a sample $\boldsymbol{x} \in \bar{D}$ let the teacher models outputs $f^i(\boldsymbol{x})$ be in agreement, $\forall i \in [k]$. The flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ is given by $p^{\leftrightarrow}_{\boldsymbol{x}} = 1 - \Phi(\frac{k}{\sqrt{2}\sigma})$, 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. This analysis indicates that the ensemble size k (as well as the privacy parameter σ) has a direct impact on the outcome of the teacher voting and, thus, it affects the model deviation and its disparate impact. The theorem shows that larger k values correspond to smaller flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$. *Combined with Theorem 1, it suggests that the model deviation due to privacy and the various groups' excess risks are inversely proportional to the ensemble size k.*

Figure 4 (top) 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. Note that, in these experiments, *different teachers may have different agreements on each sample*, thus they generalize the result of Theorem 3. Next, Figure 3 (bottom) shows that increasing k reduces the expected model deviation (left), reduces the group excess risk difference (middle), and increases the model $\bar{f}$ accuracy (right). Similarly as for the regularization term λ, large values k can reduce the accuracy of the (private and non-private) models. This behavior is related with the bias-variance tradeoff imposed on the growing ensemble with less training data available to each teacher.

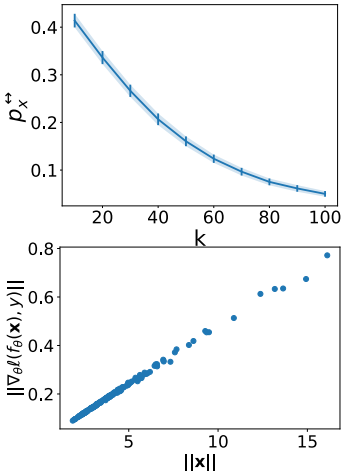

Figure 4: Credit-card: Average flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ for samples $\boldsymbol{x} \in \bar{D}$ as a function of the ensemble size k (top) and relation between gradient and input norms (bottom).

This section concludes with a useful corollary of Theorem 2.

Corollary 1 (Theorem 2)**.** *For a* logistic regression *classifier $\bar{f}_{\boldsymbol{\theta}}$, the model deviation is upper bounded as:*

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

The result above highlights several interesting points. First, it indicates the presence of a relation between gradient norms and input norms, which is further highlighted in Figure 4 (bottom). The plot illustrates the strong correlation between inputs and their associated gradient norms. Second, it shows that samples with large 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.

In summary, the regularization paramter λ and the ensemble size k are two key algorithmic parameters that, by bounding the model deviation $\Delta_{\tilde{\theta}}$, can control the disparate impacts of the student model. These relations are further illustrated in the causal graph on Figure 1.

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 results below show that the norms of the student's data samples and their distance to the decision boundary are two key factors tied to the exacerbation of excess risk in PATE. As we will discuss next, this is particularly interesting as it demystifies a recurrent belief about *unfairness being solely a consequence of imbalanced training data*. The following is a second corollary of Theorem 2 and bounds the second order statistics of the model deviation to privacy.

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

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

Note that, similarly to what shown by Corollary 1, when $\bar{f}_{\boldsymbol{\theta}}$ is a logistic regression model, the gradient norm $\|G_{\boldsymbol{x}}^{\max}\|$ above can be substituted with the input norm $\|\boldsymbol{x}\|$.

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.

(D_1): The impact of the data input norms. First notice that the norm $\|\boldsymbol{x}\|$ of a sample $\boldsymbol{x}$ strongly influences the model deviation controlling quantity $\Delta_{\tilde{\theta}}$ as already observed by Corollaries 1 and 2. This aspect is further highlighted in Figure 5 (top), which illustrates the strong correlation between the input norms and the model deviation. *Thus, samples with high input norms may have a nontrivial impact to the model deviation and, in turn, to its unfairness (see Theorem 1).*

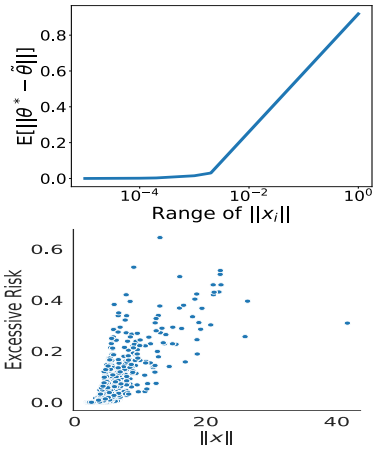

Figure 5: *Credit*: Relation between input norms and model deviation (top) and Spearman correlation between input and excess risk (bottom).

Next, recall that the group gradient norms G_a have a proportional effect on the upper bound of the model unfairness, as shown in Theorem 1 (as well as on the excess risk $R(\bar{D}_{\leftarrow a})$, as shown in Lemma 1, Appendix A). The following results first highlights the relation between gradient norm for a sample $\boldsymbol{x} \in \bar{D}$ and its associated input norm and then it connects such remark to the observed student model's unfairness.

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

$$\nabla_{\boldsymbol{\theta}^*}\ell(\bar{f}_{\boldsymbol{\theta}^*}(\boldsymbol{x}), y) = (\bar{f}_{\boldsymbol{\theta}^*}(\boldsymbol{x}) - y) \otimes \boldsymbol{x},$$

where $\otimes$ expresses the Kronecker product.

Thus, the relation above suggests that the *input norm* of data samples play a key role in controlling their associated excess risk, and, thus, that of the group in which they belong to. This aspect can be appreciated in Figure 5 (bottom), which shows a strong correlation between the input norms and excess risk. This observation is significant. It demystifies a common belief that unfairness is solely caused by the imbalances in group sizes. *Rather, are the data properties themselves that directly contributes to unfairness.*

Finally, the discussion notes that the smoothness parameter β_a captures the local flatness of the loss function relative to samples of a group a. A derivation of β_a for logistic regression classifiers is extended from a results of Shi et al. (2021) and further illustrates the relationship between input norms $\|\boldsymbol{x}\|$ of a group $a \in \mathcal{A}$ and the smoothness parameters β_a.

Proposition 2. *Consider again a binary logistic regression as in Proposition 1. The smoothness parameter β_a for a group $a \in \mathcal{A}$ is given by: $\beta_a = 0.25 \max_{\boldsymbol{x}\in D_a} \|\boldsymbol{x}\|^2$.*

Thus, Propositions 1 and 2 illustrate that groups with large (small) inputs' norms tend to have large (small) gradient's norms and smoothness parameters. As these factors control the model deviation, they also affect their associated excess risk, resulting in larger disparate effects. An extended analysis of the above claim is provided in Appendix C.7.

(D_2): The impact of the distance to decision boundary. As mentioned in Theorem 2, the flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ of a sample $\boldsymbol{x} \in \bar{D}$ directly controls the model deviation $\Delta_{\bar{\theta}}$. Intuitively, samples close to the decision boundary are associated to small ensemble voting confidence and vice-versa. Thus, groups with samples close to the decision boundary will be more sensitive to the noise induced by the private voting process. To illustrate this intuition the paper reports the concept of *closeness to boundary*.

Definition 4 (Tran et al. (2021a)). *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: $s(\boldsymbol{x}) \stackrel{def}{=} 1 - \sum_{c=1}^{C} f_{\boldsymbol{\theta}^*,c}(\boldsymbol{x})^2$, where $f_{\boldsymbol{\theta},c}$ denotes the softmax probability for class c.*

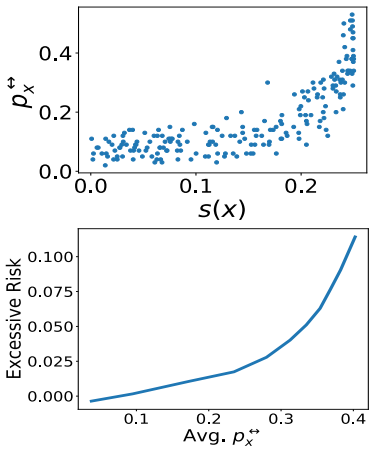

Figure 6: *Credit*: Spearman correlation between closeness to boundary $s(\boldsymbol{x})$ and flipping probability $p^{\leftrightarrow}_{\boldsymbol{x}}$ (top) and relation between input norms and excess risk (bottom).

The above, relates large (small) $s(\boldsymbol{x})$ values 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 (top). The plots are once again generated using a neural network with nonlinear objective and the relation holds for all datasets analyzed. The plot indicates that the samples that are close to the decision boundary have a higher probability of "flipping" their label, thus resulting in a worse excess risk, and thus unfairness. Finally, the strong proportional effect of the flipping probability on the excess risks is illustrated in Figure 6 (bottom).

In summary, the norms $\|\boldsymbol{x}\|$ of a group' samples and their associated distance to boundary $s(\boldsymbol{x})$ are two key characteristics of the student data that, by controlling the model deviation $\Delta_{\tilde{\theta}}$, as well as the smoothness parameters β_a and the group gradients G_a, can control the disparate impacts of the student model (see Figure 2 for a schematic representation).

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 is their effects on the model deviation $\Delta_{\tilde{\theta}}$ which, in turn, is related with the excess risk of different groups, whose difference we would like to minimize (see again Theorem 1).

Motivated by these observations, this section proposes a strategy that aims at reducing the deviation of the private model parameters. To do so, we exploit the idea of *soft labels* (as defined below). The traditional voting scheme (denoted *hard labels* in this section) may be significantly affected by small perturbations due to noise, especially when teachers have low voting confidence. Consider, for example, the case of a binary classifier where for a sample $\boldsymbol{x}$, $k/2+1$ teachers vote label 0 and $k/2-1$, label 1, for some even ensemble size k. When perturbations are induced to these counts to guarantee privacy, the process may report the incorrect label ($\hat{y} = 1$) with high probability. As a result, the student model private parameters may be sensitive to the noisy voting and deviate significantly from the non-private ones. 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: $\alpha(\boldsymbol{x}) = \left(\frac{\#_c(\boldsymbol{T}(\boldsymbol{x}))}{k}\right)_{c=1}^{C}$ and their privacy-preserving counterparts $\tilde{\alpha}(\boldsymbol{x})$ adds Gaussian noise $\mathcal{N}(0,\sigma^2)$ in the numerator of $\alpha(\boldsymbol{x})$.*

To exploit soft labels, the training step of the student model uses loss $\ell'(\bar{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \tilde{\alpha}) = \sum_{c=1}^{C} \tilde{\alpha}_c \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), c)$, which can be considered as a weighted version of the original loss function $\ell(\bar{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), c)$ on class label c, whose weight is its confidence $\tilde{\alpha}_c$. Note that $\ell'(\bar{f}_{\boldsymbol{\theta}}(\boldsymbol{x}), \tilde{\alpha}) = \ell(\bar{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))$ when all teachers in the ensemble chose the same label. The privacy loss for this model is equivalent to that of classical PATE. The analysis is reported in Appendix B.

The effectiveness of this scheme is demonstrated in Figure 7. The experiment settings are reported in details in the Appendix and reflect those described in Section 5. The left subplot shows the relation

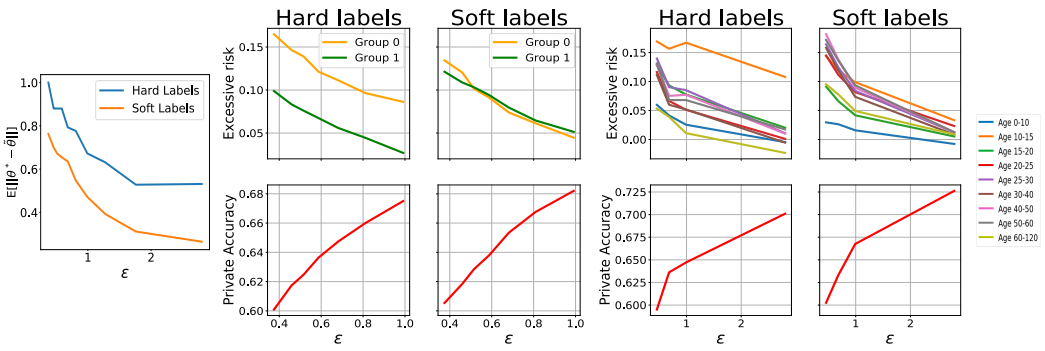

Figure 7: Training privately PATE with hard and soft labels: Model deviation at varying of the privacy loss (left) on Credit dataset and excess risk at varying of the privacy loss for Credit (middle) and UTKFace (right) datasets.

between the model deviation $\mathbb{E}[\Delta_{\tilde{\theta}}]$ at varying of the privacy loss ϵ (dictated by the noise level σ). Notice how the student models trained using soft labels reduce their model deviation ($E[\Delta_{\tilde{\theta}}]$) 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 excess risks $R(\bar{D}_{\leftarrow a})$ associated with each group $a \in \mathcal{A}$ for Credit (left) and UTKFace datasets (right), while the bottom subplots illustrate the accuracy of the models (a simple ReLU network for the tabular dataset and a much more complex CNN for the image dataset). Recall that the fairness goal $\xi(\bar{D})$ is captured by the gap between excess risk curves in the figures. Notice how soft labels can reduce the disparate impacts in private training (top). Notice also 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 Appendix C and the trends are all consistent with what described above. Finally, an important benefit of this solution is that it *does not* uses the protected group information ($a \in \mathcal{A}$) during training. Thus, it is applicable in challenging situations when it is not feasible to collect or use protected features (e.g., under the General Data Protection Regulation (GDPR) Lahoti et al. (2020)). *These results are significant. They suggest that this mitigating solution can be effective for improving the disparate impact of private model ensembles without sacrificing accuracy.*

9 Discussion, Limitations, and Conclusions

This section discusses two key messages arising from this study. First, 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 deviation is affected by the voting outcomes of the teacher ensemble (Theorems 1 and 2). Similarly to robust ML models being insensitive to input perturbations, strongly agreeing ensemble will be less sensitive to noise and vice-versa. This observation raises a question about the connection of robustness and fairness in private models, which, we believe is an important open question. Next, we also notice that 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), which may trigger some nontrivial unfairness issues on its own.

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

Abadi and et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, pp. 15479–15488, 2019.

C.L. Blake and C.J. Merz. Uci repository of machine learning databases, 1988. URL https://archive.ics.uci.edu/ml/datasets.php.

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. Combining unsupervised and supervised learning in credit card fraud detection, 05 2019.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 309–315, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

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

Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pp. 35–47, 2018.

Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 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*, pp. 15–19, 2020.

Ferdinando Fioretto, Cuong Tran, and Pascal Van Hentenryck. Decision making with differential privacy under a fairness lens, 2021.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34*, 2021.

Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation, 2020.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. *arXiv preprint arXiv:1812.02696*, 2018.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning, 2020.

Max Little, Patrick Mcsharry, Stephen Roberts, Declan Costello, and Irene Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomedical engineering online*, 6:23, 02 2007. doi: 10.1186/1475-925X-6-23.

Ata Mahjoubfar, Claire Lifan Chen, and Bahram Jalali. Deep learning and classification. In *Artificial Intelligence in Label-free Microscopy*, pp. 73–85. Springer, 2017.

Ilya Mironov. Rényi differential privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, Aug 2017. doi: 10.1109/csf.2017.11. URL http://dx.doi.org/10.1109/CSF.2017.11.

Hussein Mozannar, Mesrob I. Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. 02 2018.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. *Advances in Neural Information Processing Systems*, 27:190–198, 2014.

David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 189–199, 2020.

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

Shai Shalev-Shwartz. Online learning: Theory, algorithms, and applications. 08 2007.

Zheng Shi, Nicolas Loizou, Peter Richtárik, and Martin Takáč. Ai-sarah: Adaptive and implicit stochastic recursive gradient methods, 2021.

Cuong Tran, My H. Dinh, and Ferdinando Fioretto. Differentially private deep learning under the fairness lens, 2021a.

Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 9932–9939. AAAI Press, 2021b.

Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 560–566, 2021c.

Archit Uniyal, Rakshit Naidu, Sasikanth Kotti, Sahib Singh, Patrik Kenfack, Fatemehsadat Mireshghallah, and Andrew Trask. Dp-sgd vs pate: Which has less disparate impact on model accuracy?, 06 2021.

Depeng Xu, Wei Du, and Xintao Wu. 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, pp. 1924–1932, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467268. URL https://doi.org/10.1145/3447548.3467268.

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 3922–3928, 2017. doi: 10.24963/ijcai.2017/548. URL https://doi.org/10.24963/ijcai.2017/548.