On the Fairness Impacts of Private Ensembles Models

Cuong Tran[1] and **Ferdinando Fioretto**[2]
[1] Syracuse University
[2] University of Virginia
cutran@syr.edu, fioretto@virginia.edu

Abstract

The Private Aggregation of Teacher Ensembles (PATE) is a machine learning framework that enables the creation of private models through the combination of multiple "teacher" models and a "student" model. The student model learns to predict an output based on the voting of the teachers, and the resulting model satisfies differential privacy. PATE has been shown to be effective in creating private models in semi-supervised settings or when protecting data labels is a priority. This paper explores whether the use of PATE can result in unfairness, and demonstrates that it can lead to accuracy disparities among groups of individuals. The paper also analyzes the algorithmic and data properties that contribute to these disproportionate impacts, why these aspects are affecting different groups disproportionately, and offers recommendations for mitigating these effects.

1 Introduction

The widespread adoption of machine learning (ML) systems in decision-making processes have raised concerns about bias and discrimination, as well as the potential for these systems to leak sensitive information about the individuals whose data is used as input. These issues are particularly relevant in contexts where ML systems are used to assist in decisions processes impacting individuals' lives, such as criminal assessment, lending, and hiring.

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; Tran *et al.*, 2021a; Tran *et al.*, 2021d]. The resulting outcomes can have significant impacts on individuals with negative effects on financial, criminal, or job-hiring decisions [Fioretto *et al.*, 2021]. *While these surprising observations have become apparent in several contexts, their causes are largely understudied.*

This paper makes a step toward filling this important gap and investigates the unequal impacts that can occur when training a model using Private Aggregation of Teacher Ensembles (PATE), a state-of-the-art privacy-preserving ML framework [Papernot *et al.*, 2018]. PATE involves combining multiple agnostic models, referred to as *teachers*, to create a *student* model that is able to predict an output based on noisy voting among the teachers. This approach satisfies differential privacy and has been demonstrated to be effective for learning high-quality private models in semi-supervised settings. The paper examines which algorithmic and data properties contribute to disproportionate impacts, why these aspects are affecting different groups of individuals disproportionately, and proposes a solution for mitigating these effects.

In summary, the paper makes several key contributions: **(1)** It introduces a fairness measure that extends beyond accuracy parity and assesses the direct impact of privacy on model outputs for different groups. **(2)** It examines this fairness measure in the context of PATE, a leading privacy-focused ML framework. **(3)** It identifies key components of model parameters and data properties that contribute to disproportionate impacts on different groups during private training. **(4)** It investigates the circumstances under which these components disproportionately affect different groups. **(5)** Finally, based on these findings, the paper proposes a method for reducing these unfair impacts while maintaining high accuracy.

The empirical advantages of privacy-preserving ensemble models over other frameworks, such as DP-SGD [Abadi and et al., 2016; Ghazi *et al.*, 2021; Uniyal *et al.*, 2021], make this work a significant and widely relevant contribution to understanding and addressing the disproportionate impacts observed in semi-supervised private learning systems. As far as we are aware, this is the first study to examine the causes of disparate impacts in privacy-preserving ensemble models.

Supplemental material. A privacy analysis, proofs of all theorems, and additional experiments can be found in [Tran and Fioretto, 2023].

2 Related Work

The relationship between privacy and fairness has been a topic of recent debate, as recently surveyed by [Fioretto *et*

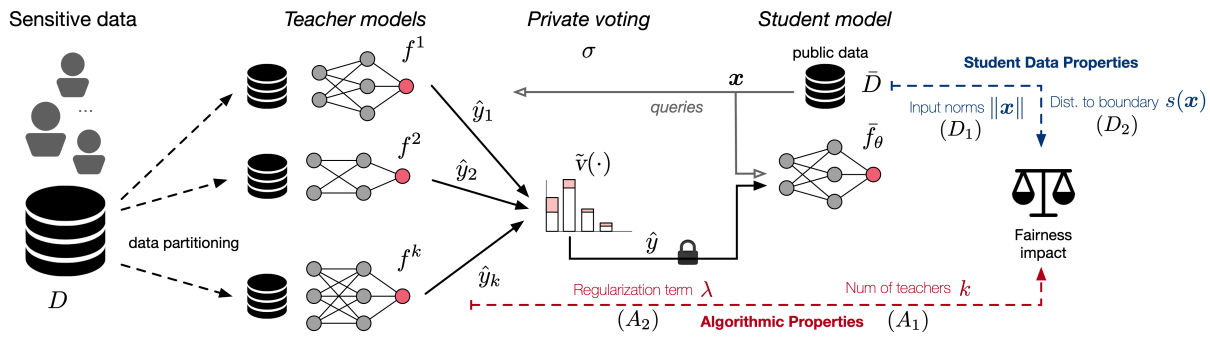

Figure 1: Illustration of PATE and aspects contributing to fairness.

al., 2022], with several researchers raising questions about the tradeoffs involved [Ekstrand *et al.*, 2018]. [Cummings *et al.*, 2019] specifically studied the tradeoffs between differential privacy and equal opportunity, a fairness criterion that requires a classifier to have equal true positive rates for different groups. They demonstrated that it is not possible to simultaneously achieve $(\epsilon, 0)$-differential privacy, satisfy equal opportunity, and have accuracy better than a constant classifier. Additionally, it has been proven that when training data has a long-tailed distribution, it is impossible to develop a private learning algorithm that has high accuracy for minority groups [Sanyal *et al.*, 2022]. These findings led to asking if fair models can be created while preserving sensitive information, and have spurred the development of various approaches such as those presented in [Jagielski *et al.*, 2018; Mozannar *et al.*, 2020; Tran *et al.*, 2021a; Tran *et al.*, 2021c; Tran *et al.*, 2021b; Fioretto *et al.*, 2020].

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.* [2021d]. 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 of the causes of unfairness in the context of private learning ensembles, a significant privacy-enhancing ML system, and introduces guidelines for mitigating these effects.

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)-differential privacy, 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 classifier that needs to predict criminal defendants' 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.

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 $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$:

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \bar{D}, \boldsymbol{T}) + \lambda \|\boldsymbol{\theta}\|^2 \tag{1}$$

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

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.

We focus 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:

$$\tilde{v}(\boldsymbol{T}(\boldsymbol{x})) = \operatorname{argmax}_c \{\#_c(\boldsymbol{T}(\boldsymbol{x})) + \mathcal{N}(0, \sigma^2)\} \tag{3}$$

which perturbs the reported counts $\#_c(\boldsymbol{T}(\boldsymbol{x})) = |\{j : j \in [k], f^j(\boldsymbol{x}) = c\}|$ for class $c \in \mathcal{C}$ 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 C of [Tran and Fioretto, 2023]. 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. One widely used metric for measuring utility in private learning is the *excess risk* [Zhang *et al.*, 2017], which is defined as the difference between the private and non-private risk functions:

$$R(S, \boldsymbol{T}) \stackrel{\text{def}}{=} \mathbb{E}_{\tilde{\boldsymbol{\theta}}}\left[\mathcal{L}(\tilde{\boldsymbol{\theta}}; S, \boldsymbol{T})\right] - \mathcal{L}(\boldsymbol{\theta}^*; S, \boldsymbol{T}), \tag{4}$$

where the expectation is taken over the randomness of the private mechanism, S is a subset of $\bar{D}$, $\tilde{\boldsymbol{\theta}}$ is the private student model's parameters, and $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \bar{D}, \boldsymbol{T}) + \lambda \|\boldsymbol{\theta}\|^2$.

In this paper, the unfairness introduced by privacy in the learning task is measured using the difference in excess risks of each protected subgroup. This notion is significant because it captures the unintended impact of privacy on task accuracy for a given group, and it relates to the concept of accuracy parity, a standard metric in fair and private learning. More specifically, the paper focuses on measuring the excess risk $R(\bar{D}_{\leftarrow a}, \boldsymbol{T})$ for groups $a \in \mathcal{A}$, where $\bar{D}_{\leftarrow a}$ is the subset of $\bar{D}$ containing only samples from a group a. We use the shorthand $R(\bar{D}_{\leftarrow a})$ to refer to $R(\bar{D}_{\leftarrow a}, \boldsymbol{T})$ and assume that the private mechanisms are non-trivial, i.e., they minimize the population-level excess risk $R(\bar{D})$.

Definition 2. *Fairness is measured as the highest excess risk difference among all groups:*

$$\xi(\bar{D}) = \max_{a, a' \in \mathcal{A}} R(\bar{D}_{\leftarrow a}) - R(\bar{D}_{\leftarrow a'}). \tag{5}$$

Notice how this definition of fairness relates to the concept of accuracy parity [Bagdasaryan *et al.*, 2019], which measures the disparity of task accuracy across groups, when the adopted loss ℓ is a 0/1-loss. All the experiments in the paper use, in fact, this 0/1-loss, while the theoretical analysis considers general differentiable loss functions. Additional details regarding this fairness definition and its relations with other fairness notions can be found in Appendix A of [Tran and Fioretto, 2023].

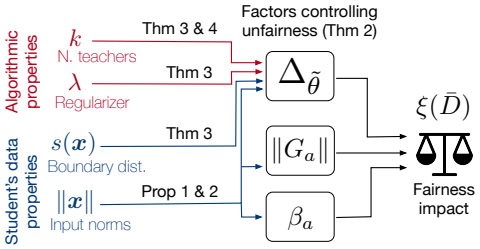

Figure 2: Factors impacting PATE fairness.

5 PATE Fairness Analysis: Roadmap

The objective of this paper is to identify the factors that cause unfairness in PATE and understand why they have this effect. The following sections isolate these key factors, which will be divided into two categories: *algorithm parameters* and *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 with respect to the model parameters $\boldsymbol{\theta}$ for some $\beta_a \geq 0$. However, the evaluation does not impose any restrictions on the form of the loss function. A detailed description of the experimental settings can be found in Appendix D, and the proofs of all theorems are included in Appendix A of [Tran and Fioretto, 2023].

A fairness bound. We start by introducing a bound on the model disparity, which will be crucial for identifying the algorithm and data characteristics that contribute to 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 teachers' voting on the student learned model. Here, $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}$ represent the parameters of student model $\bar{f}$ learned using a clean or noisy voting scheme, respectively.

Theorem 1. *The model fairness is upper bounded as:*

$$\xi(\bar{D}) \leq 2 \max_a \|G_a\| \, \mathbb{E}\left[\Delta_{\tilde{\boldsymbol{\theta}}}\right] + \tfrac{1}{2} \max_a \beta_a \, \mathbb{E}\left[\Delta^2_{\tilde{\boldsymbol{\theta}}}\right], \tag{6}$$

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 maximum gradient norm $\max_a \|G_a\|$ among all groups, and **(3)** the largest smoothness parameter $\max_a \beta_a$ among all groups.

The paper delves into which **A**lgorithms' parameters and **D**ata characteristics affect the factors that contribute to model unfairness. Within the **A**lgorithm's parameters, in addition to the privacy variable ϵ (captured by the noise parameter σ),

the paper identifies two 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 $s(\boldsymbol{x})$) are key factors that can exacerbate the excess risks induced by the student model. The relationships between these factors and how they impact model fairness are illustrated in Figure 2. Several aspects of the analysis in this paper rely on the following definition.

Definition 3. *Given a data sample $(\boldsymbol{x}, y) \in D$, for an ensemble $\boldsymbol{T}$ and voting scheme v, the* flipping probability *is:*

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

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.

The theoretical results presented in the following sections are supported and corroborated by empirical evidence from tabular datasets (UCI Adults, Credit card, Bank, and Parkinsons) and an image dataset (UTKFace). These results were obtained using feed-forward networks with two hidden layers and nonlinear ReLU activations for both the ensemble and student models for tabular data, and CNNs for image data. All reported metrics are the average of 100 repetitions used to compute empirical expectations and report 0/1 losses, *which capture the concept of accuracy parity*. While the paper provides a brief overview of the empirical results to support the theoretical claims, extended experiments and more detailed descriptions of the datasets can be found in Appendix D of [Tran and Fioretto, 2023].

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 4. *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 $y \in \mathcal{Y}$:*

$$\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = z(h_{\boldsymbol{\theta}}(\boldsymbol{x})) + c\, y\, h_{\boldsymbol{\theta}}(\boldsymbol{x}). \tag{7}$$

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 proportionally controls the unfairness of PATE (Theorem 1). In the following, we provide a useful bound on the model deviation and highlight its relationship with key algorithm 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:*

$$\mathbb{E}\Big[\Delta_{\tilde{\boldsymbol{\theta}}}\Big] \leq \frac{|c|}{m\lambda}\left[\sum_{\boldsymbol{x}\in\bar{D}} p_{\boldsymbol{x}}^{\leftrightarrow} \|G_{\boldsymbol{x}}^{\max}\|\right], \tag{8}$$

where c is a real constant and $G_{\boldsymbol{x}}^{\max} = \max_{\boldsymbol{\theta}} \|\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}(\boldsymbol{x})\|$ represents the maximum gradient norm distortion introduced by a sample $\boldsymbol{x}$. Both c and h are defined as in Equation 7.

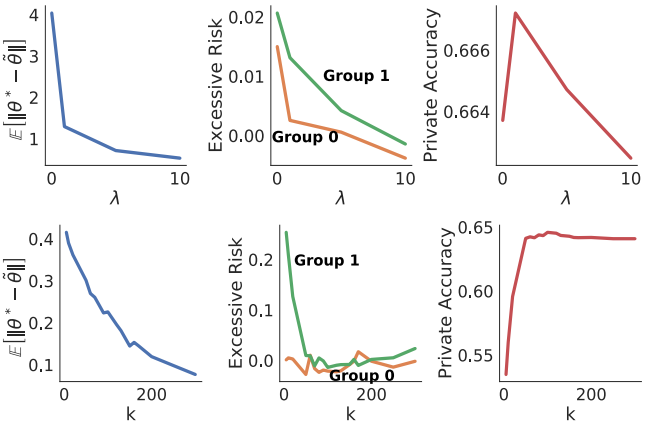

Figure 3: Credit card dataset with $\sigma = 50$, $k = 150$ (top) and $\lambda = 100$ (bottom). Expected model deviation (left), excess risk (middle), and model accuracy (right) as a function of the regularization term (top) and ensemble size (bottom).

The proof relies on λ-strong convexity of the loss function $\mathcal{L}(\cdot) + \lambda\|\boldsymbol{\theta}\|$ (see Appendix B of of [Tran and Fioretto, 2023]) and its tightness is demonstrated empirically in Appendix D.2 of [Tran and Fioretto, 2023]. Theorem 2 reveals 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 $p_{\boldsymbol{x}}^{\leftrightarrow}$, which, as it will be shown later, is heavily controlled by the size k of the teacher ensemble. These mechanisms-dependent components and the focus of this section, while data-dependent components, including those related to the maximum gradient norm distortion $G_{\boldsymbol{x}}^{\max}$ are discussed to Section 7.

$\boldsymbol{A_1}$: The impact of the regularization term λ. The first immediate observation of Theorem 2 is that variations of the regularization term λ can increase or decrease the difference between the private and non-private student model parameters. Since the model deviation $\mathbb{E}[\Delta_{\tilde{\boldsymbol{\theta}}}]$ has adirect relationship with the fairness goal (see the first term of RHS of Equation 6 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 $\mathbb{E}[\Delta_{\tilde{\boldsymbol{\theta}}}]$ (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].

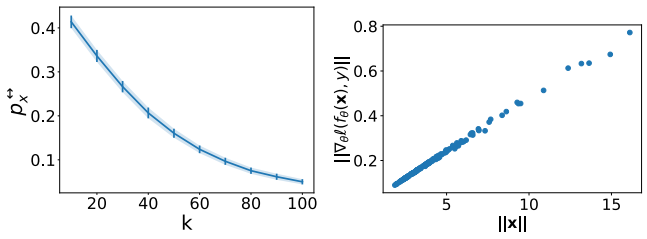

Figure 4: Credit-card: Average flipping probability $p_{\boldsymbol{x}}^{\leftrightarrow}$ for samples $\boldsymbol{x} \in \bar{D}$ as a function of the ensemble size k (left) and the relation between gradient and input norms (right).

A_2: The impact of the teachers ensemble size k. Next, we consider the relationship 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.

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 $p_{\boldsymbol{x}}^{\leftrightarrow}$ is given by $p_{\boldsymbol{x}}^{\leftrightarrow} = 1 - \Phi(\frac{k}{\sqrt{2}\sigma})$, 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 shows that the ensemble size k (as well as the privacy parameter σ) directly affects the outcome of the teacher voting and, therefore, the model deviation and its disparate impact. The theorem shows that larger k values correspond to smaller flipping probability $p_{\boldsymbol{x}}^{\leftrightarrow}$. *In conjunction with Theorem 1, this suggests that the model deviation due to privacy and the excess risks for various groups are inversely proportional to the ensemble size k.*

Figure 4 (top) illustrates the relationship between the number k of teachers and the flipping probability $p_{\boldsymbol{x}}^{\leftrightarrow}$ of the ensemble, indicating that larger ensembles result in smaller flipping probabilities. It is worth noting that in these experiments, *different teachers may have different agreements on each sample*, thus this result generalizes the one presented in Theorem 3. Additionally, Figure 3 (bottom) shows that increasing k reduces the expected model deviation (left), reduces the group excess risk difference (middle), and increases the accuracy of the model $\bar{f}$ (right). Similar to theregularization term λ, large values k can decrease the accuracy of the (private and non-private) models. This behavior is related to the bias-variance tradeoff imposed on the growing ensemble with less training data available to each teacher.
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:*

$$\mathbb{E}\left[\Delta_{\tilde{\boldsymbol{\theta}}}\right] \leq \frac{1}{m\lambda}\left[\sum_{\boldsymbol{x}\in\bar{D}} p_{\boldsymbol{x}}^{\leftrightarrow}\|\boldsymbol{x}\|\right]. \tag{9}$$

This result highlights the presence of a relationship between gradient norms and input norms, which is further illustrated in Figure 4 (bottom). The plot shows a strong correlation between inputs and their associated gradient norms. The result also shows that samples with large norms can significantly impact fairness, emphasizing the importance of considering the characteristics of the student data, which are the subject of study in the next section.

In summary, the regularization parameter 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 in Figure 1.

7 Student's Data Properties

Having examined the algorithmic properties of PATE affecting fairness, this section turns on analyzing the role of certain characteristics of the student data in regulating the disproportionate impacts of of the algorithm. The results below will show that the norms of the student's data samples and their distance to the decision boundary can significantly impact the excess risk in PATE. This is particularly interesting as it dispels the notion that unfairness in these models is solely due to 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:*

$$\mathbb{E}\left[\Delta_{\tilde{\boldsymbol{\theta}}}^2\right] \leq \frac{|c|^2}{m\lambda^2}\left[\sum_{\boldsymbol{x}\in\bar{D}} p_{\boldsymbol{x}}^{\leftrightarrow 2}\|G_{\boldsymbol{x}}^{\max}\|^2\right]. \tag{10}$$

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 $\Delta_{\tilde{\boldsymbol{\theta}}}$ as already observed by Corollaries 1 and 2. This aspect is further highlighted in Figure 5 (top), which illustrates that samples with high input norms have a significant impact on the model deviation. *As a result, these samples may contribute to the unfairness of the model, as per Theorem 1.*

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. These norms also have an effect on the excess risk $R(\bar{D}_{\leftarrow a})$, as shown in Lemma 1, Appendix B of [Tran and Fioretto, 2023] The following results reveal a connection between the gradient norm for a sample $\boldsymbol{x} \in \bar{D}$ and its associated input norm, and how these factors relate to the unfairness observed in the student model.

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:*

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

where $\otimes$ expresses the Kronecker product.

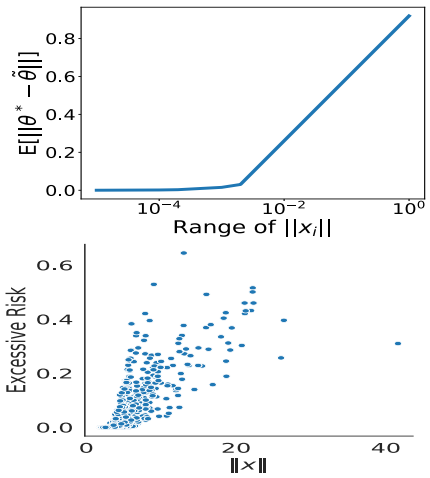

Figure 5: *Credit*: Relation between input norms and model deviation (top) and excess risk (bottom).

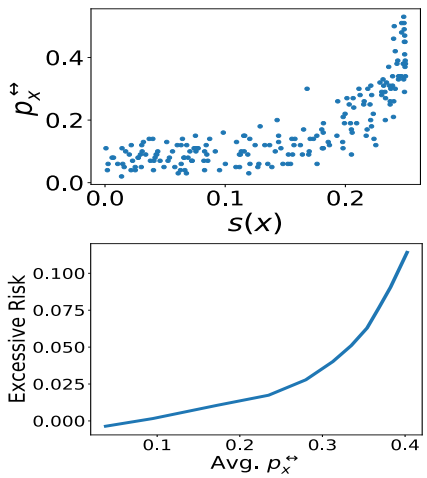

Figure 6: *Credit*: Spearman correlation between closeness to boundary $s(\boldsymbol{x})$ and flipping probability $p_{\boldsymbol{x}}^{\leftrightarrow}$ (top) and relation between input norms and excess risk (bottom).

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 because it challenges the common belief that unfairness is solely caused by imbalances in group sizes. Instead, it suggests that the properties of the data itself directly contribute to unfairness.

Finally, note that the smoothness parameter β_a reflects the local flatness of the loss function in relation to samples from a group a. An extension of the results from [Shi *et al.*, 2023] is provided to derive β_a for logistic regression classifiers, further illustrating the connection between the 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: $\beta_a = 0.25 \max_{\boldsymbol{x} \in D_a} \|\boldsymbol{x}\|^2$.*

Therefore, Propositions 1 and 2 show that groups with large (small) inputs' norms tend to have large (small) gradient norms and smoothness parameters. Since these factors influence the model deviation, they also affect the associated excess risk, leading to larger disparate impacts. An extended analysis of the above claim is provided in Appendix D.7 of [Tran and Fioretto, 2023].

(D_2): The impact of the distance to decision boundary. As mentioned in Theorem 2, the flipping probability $p_{\boldsymbol{x}}^{\leftrightarrow}$ of a sample $\boldsymbol{x} \in \bar{D}$ directly controls the model deviation $\Delta_{\tilde{\boldsymbol{\theta}}}$. 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 5 ([Tran *et al.*, 2021b])**.** *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 $f_{\boldsymbol{\theta},c}$ is the softmax of class c.*

The above discussion relates large (small) values of $s(\boldsymbol{x})$ to projections of point $\boldsymbol{x}$ that are close (distant) to the model decision boundary. *The concept of closeness to decision boundary provides a way to indirectly quantify the flipping probability of a sample.* Empirically, the correlation between the distance of sample $\boldsymbol{x}$ to the decision boundary and its flipping probability $p_{\boldsymbol{x}}^{\leftrightarrow}$ 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, leading to a worse excess risk and unfairness. Finally, Figure 6 (bottom) further illustrates the strong proportional effect of the flipping probability on the excess risk.

To summarize, the norms $\|\boldsymbol{x}\|$ of a group's samples and their associated distance to boundary $s(\boldsymbol{x})$ are two key characteristics of the student data that influence fairness through their control of the model deviation $\Delta_{\tilde{\theta}}$, the smoothness parameters β_a, and the group gradients G_a, (see Figure 2 for a schematic representation).

8 Mitigation Solution

The previous sections have identified a number of algorithmic and data-related factors that can influence the disparate impact of the student model. These factors often affect the model deviation $\Delta_{\tilde{\boldsymbol{\theta}}}$, which is related to the excess risk of different groups (as shown in Theorem 1), whose difference we would like to minimize. With this in mind, this section proposes a strategy to reduce the deviation of the private model parameters. To do so, we exploit the idea of *soft labels* instead of traditional *hard labels* in the voting process. Hard labels may be significantly affected by small perturbations due to noise, especially when the teachers have low confidence in their votes. For example, consider 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. If perturbations are

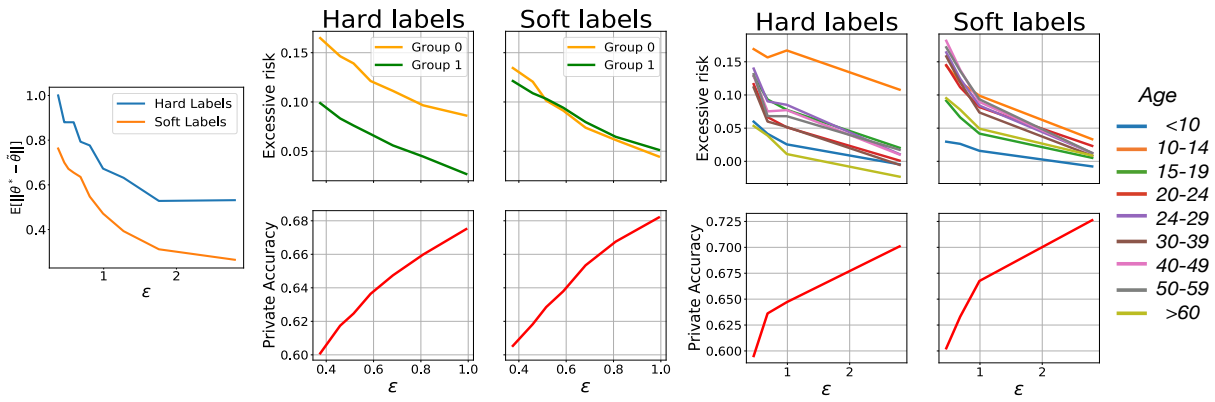

Figure 7: Training privately PATE with hard and soft labels: Model deviation at varying of the privacy loss (left) on Credit dataset and excess risk at varying of the privacy loss for Credit (middle) and UTKFace (right) datasets.

introduced to these counts to esnure privacy, the process may incorrectly report label ($\hat{y} = 1$) with high probability, causing causing the student model's private parameters to deviate significantly from the non-private ones. This issue can be partially addressed by the introduction of soft labels:

Definition 6. *The* soft label *of a sample* $\boldsymbol{x}$ *is* $\alpha(\boldsymbol{x}) = (\#_c(\boldsymbol{T}(\boldsymbol{x}))/_k)_{c=1}^{C}$ *and its private counterpart* $\tilde{\alpha}(\boldsymbol{x})$ *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 C of [Tran and Fioretto, 2023].

The effectiveness of this scheme is demonstrated in Figure 7. The experiment settings are reported in detail in [Tran and Fioretto, 2023] (Appendix) and reflect those described in Section 5. The left subplot shows the relation between the model deviation $\mathbb{E}\left[\Delta_{\tilde{\boldsymbol{\theta}}}\right]$ 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 ($E[\Delta_{\tilde{\boldsymbol{\theta}}}]$) when compared to the counterparts that use hard labels.

The middle and right plots of Figure 7 show the impact of the proposed solution on the private student model in terms of the utility/fairness tradeoff. 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 (right) datasets, respectively. The bottom subplots shows the accuracy of the models, which include a simple ReLU network for the tabular data and a 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 than the counterparts produced with hard labels.

Additional experiments, including illustrating the behavior of the mitigating solution at varying of the number of teachers are reported in [Tran and Fioretto, 2023] (Appendix D). Note also that the proposed solution preserves the original privacy budget. In contrast, mitigating solutions that would consider explicitly the number of teachers K or the smoothness parameter λ will inevitably introduce further privacy/fairness tradeoffs as would require costly privacy-preserving hyperparameter optimization [Papernot and Steinke, 2021].

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 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 study highlights two key messages. First, the proposed mitigating solution relates to concepts in robust machine learning. In particular, Papernot *et al.* [2016] showed that training a classifier with soft labels can increase its robustness against adversarial samples. This connection is not coincidental, as the deviation of the model is influenced by the voting outcomes of the teacher ensemble (as demonstrated in Theorems 1 and 2). In the same way that robust ML models are insensitive to input perturbations, an ensemble that strongly agrees will be less sensitive to noise and vice versa. This raises the question of the relationship between robustness and fairness in private models, which is an important open question. Second, we also note that more advanced voting schemes, such as interactive GNMAX [Papernot *et al.*, 2018], may produce different fairness results. While this is an interesting area for further analysis, these sophisticated voting schemes may introduce sampling bias (e.g., interactive GNMAX may exclude samples with low ensemble voting agreement), which could create its own fairness issues.

Given the growing use of privacy-preserving learning tasks in consequential decisions, this work represents a significant and widely applicable step towards understanding the causes of disparate impacts in differentially private learning systems.

Acknowledgements

This research is partially supported by NSF grants 2232054 and 2133169. Fioretto is also supported by an Amazon Research Award and a Google Research Scholar Award. Its views and conclusions are those of the authors only.

References

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

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

[Cummings *et al.*, 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*, pages 309–315, 2019.

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

[Dwork *et al.*, 2014] 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.

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

[Farrand *et al.*, 2020] 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*, pages 15–19, 2020.

[Fioretto *et al.*, 2020] Ferdinando Fioretto, Pascal Van Hentenryck, Terrence W. K. Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. In *Machine Learning and Knowledge Discovery in Databases. European Conference, ECML PKDD 2020*, volume 12461 of *Lecture Notes in Computer Science*, pages 118–135, 2020.

[Fioretto *et al.*, 2021] Ferdinando Fioretto, Cuong Tran, and Pascal Van Hentenryck. Decision making with differential privacy under a fairness lens. *arXiv preprint arXiv: Arxiv-2105.07513*, 2021.

[Fioretto *et al.*, 2022] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. In *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5470–5477, 2022.

[Ghazi *et al.*, 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, editors, *Advances in Neural Information Processing Systems 34*, 2021.

[Jagielski *et al.*, 2018] 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.

[Lahoti *et al.*, 2020] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

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

[Mozannar *et al.*, 2020] 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.

[Papernot and Steinke, 2021] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

[Papernot *et al.*, 2016] 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)*, pages 582–597. IEEE, 2016.

[Papernot *et al.*, 2018] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations*, 2018.

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

[Pujol *et al.*, 2020] 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*, pages 189–199, 2020.

[Sanyal *et al.*, 2022] Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning? In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 1738–1748, 2022.

[Shi *et al.*, 2023] Zheng Shi, Abdurakhmon Sadiev, Nicolas Loizou, Peter Richtárik, and Martin Takáč. AI-SARAH: Adaptive and implicit stochastic recursive gradient methods. *Transactions on Machine Learning Research*, 2023.

[Tran and Fioretto, 2023] Cuong Tran and Ferdinando Fioretto. On the fairness impacts of private ensembles models. *arXiv preprint arXiv:2305.11807*, 2023.

[Tran *et al.*, 2021a] Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 27555–27565. Curran Associates, Inc., 2021.

[Tran *et al.*, 2021b] Cuong Tran, My H. Dinh, and Ferdinando Fioretto. Differentially private deep learning under the fairness lens. *arXiv preprint arXiv: Arxiv-2106.02674*, 2021.

[Tran *et al.*, 2021c] 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*, pages 9932–9939. AAAI Press, 2021.

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

[Uniyal *et al.*, 2021] Archit Uniyal, Rakshit Naidu, Sasikanth Kotti, Sahib Singh, Patrik Joslin Kenfack, Fatemehsadat Mireshghallah, and Andrew Trask. Dp-sgd vs pate: Which has less disparate impact on model accuracy? *arXiv*, 2106.12576, 2021.

[Zhang *et al.*, 2017] 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*, pages 3922–3928, 2017.