SF-PATE: Scalable, Fair, and Private Aggregation of Teacher Ensembles

Cuong Tran[1] , **Keyu Zhu**[2] , **Ferdinando Fioretto**[3] and **Pascal Van Hentenryck**[2]
[1] Syracuse University
[2] Georgia Institute of Technology
[3] University of Virginia
cutran@syr.edu, kzhu67@gatech.edu, fioretto@virginia.edu, pvh@isye@gatech.edu

Abstract

A critical concern in data-driven processes is to build models whose outcomes do not discriminate against some protected groups. In learning tasks, knowledge of the group attributes is essential to ensure non-discrimination, but in practice, these attributes may not be available due to legal and ethical requirements. To address this challenge, this paper studies a model that protects the privacy of individuals' sensitive information while also allowing it to learn non-discriminatory predictors. A key feature of the proposed model is to enable the use of off-the-shelves and non-private fair models to create a privacy-preserving and fair model. The paper analyzes the relation between accuracy, privacy, and fairness, and assesses the benefits of the proposed models on several prediction tasks. In particular, this proposal allows both scalable and accurate training of private and fair models for very large neural networks.

1 Introduction

A number of decision processes with high societal impact, such as criminal assessment, lending, and hiring procedures, are increasingly being supported by machine-learning systems. A critical concern is that the learned models may report outcomes that are discriminatory against some demographic group, including gender, ethnicity, or age.

To ensure non-discrimination in learning tasks, knowledge of the *sensitive* attributes is essential. At the same time, legal and ethical requirements often prevent the use of this sensitive data. For example, U.S. law prevents using racial identifiers in the development of models for consumer lending or credit scoring, and the GDPR in the E.U. prevents the collection of protected user attributes. In this scenario, an important tension arise between **(1)** the demand for models to be non-discriminatory, **(2)** the requirement for such model to use the protected attribute during training, as adopted by common fairness models, and **(3)** the restriction on the data or protected attributes that can be used. *There is thus a need to develop learning models that can both guarantee non-discriminatory decisions and protect the privacy of the individuals' groups attributes.*

To this end, this paper introduces a novel differentially private learning framework that satisfy group fairness while providing privacy of the protected attributes. The proposed framework, called *Scalable, Fair, and Private Aggregation of Teacher Ensemble (SF-PATE)* is inspired by the success of private teachers ensemble learning [Papernot *et al.*, 2018]. These frameworks transfer the classification knowledge learned from a pretrained ensemble of models (called teachers) to a target model (called student) via a privacy-preserving aggregation process. This paper exploits this key idea, but rather than transferring the classification capability of the models, it seeks to answer an important and unanswered question: *can fairness properties of a model ensemble be transferred in a privacy-perserving way to a target model*?

In addition to providing an affirmative answer to the question above, this paper makes four key contributions: (**1**) It proposes two flavors of SF-PATE that enforce fairness properties while protecting the demographic group attributes. (**2**) In addition to guaranteeing differential privacy, the paper provides an analysis on the fairness properties of SF-PATE and shows that unfairness can be bounded in many practical settings. (**3**) Importantly, SF-PATE decouples the implementation of fairness and privacy requirements and it can be systematically built on top of (non-private) fair models which are viewed as black-box. This engineering benefit is truly unique: it simplifies the development of private and fair models, and facilitates the adoption of new fairness metrics in privacy-preserving ML, as the paper shows in the evaluation. (**4**) Evaluations on both tabular and image datasets show that SF-PATE not only achieves better accuracy, privacy, and fairness tradeoffs with respect to the current state of the art, but it is also significantly faster. This aspect is important because the added computational cost of considering privacy and fairness in already computationally heavy models (e.g., large vision or language models) can make deployment more expensive and may discourage the release of ethical and trustworthy models.

Supplemental material. [Tran *et al.*, 2022] contains an extended version of this paper including proofs of all theorems, and additional experiments. All references to the appendix in this paper refer to such an extended version.

2 Related Work

The interconnection between privacy and fairness is receiving increasing attention and the reader is referred to [Fioretto *et* *al.*, 2022] for an overview about the state of the field. Within this literature, a recent line of work observed that private models may have a negative impact towards fairness [Bagdasaryan *et al.*, 2019; Pujol *et al.*, 2020; Tran *et al.*, 2021c; Tran *et al.*, 2021a; Tran and Fioretto, 2023]. Building from these observations, [Ekstrand *et al.*, 2018] raise questions about the tradeoff between privacy and fairness and, [Jagielski *et al.*, 2019] and [Mozannar *et al.*, 2020] proposed two simple, yet effective algorithms that satisfy (ϵ, δ) and ϵ-differential privacy, respectively, for equalized odds. In particular, state of the art model *M*, proposed in [Mozannar *et al.*, 2020], adds calibrated noise to the group attributes using randomized response prior to use them as input to a fair classifier. [Xu *et al.*, 2019] proposed a private and fair logistic regression model making use of the functional mechanism [Chaudhuri *et al.*, 2011]. Finally, [Tran *et al.*, 2021b] proposed *PF-LD*, a method to train a private classifier under differential privacy stochastic gradient descent [Abadi *et al.*, 2016] while imposing fairness constraints. These constraints are imposed using a privacy-preserving extension of the Dual Lagrangian framework of [Fioretto *et al.*, 2020]. Although this method was shown effective to balance privacy and fairness, it is computationally expensive.

In contrast, this work introduces a semi-supervised framework that relies on transferring privacy and fairness from a model ensemble to construct high-quality private and fair classifiers, while also being practical and scalable.

3 Problem Settings and Goals

The paper considers datasets D consisting of n individual data points (X_i, A_i, Y_i), with $i \in [n]$ drawn i.i.d. from an unknown distribution Π. Therein, $X_i \in \mathcal{X}$ is a *non-sensitive* feature vector, $A_i \in \mathcal{A}$, where $\mathcal{A} = [m]$ (for some finite m) is a demographic group attribute and $Y_i \in \mathcal{Y}$ is a class label. The goal is to learn a classifier $\mathcal{M}_\theta : \mathcal{X} \to \mathcal{Y}$, where θ is a vector of real-valued parameters, that ensures a specified non-discriminatory notion with respect to A while guaranteeing the *privacy* of the group attribute A. The model quality is measured in terms of a non-negative, and assumed differentiable, *loss function* $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and the problem is that of minimizing the empirical risk function (ERM):

$$\overset{\star}{\theta} = \operatorname*{argmin}_{\theta} J(\mathcal{M}_\theta, D) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathcal{M}_\theta(X_i), Y_i). \qquad \text{(P)}$$

The paper focuses on learning general classifiers, such as neural networks, that satisfy group fairness (as defined next) and protect the disclosure of the group attributes using the notion of differential privacy. Importantly, the paper assumes that the attribute A is not part of the model input during inference. This is crucial in the application of interest to this work as the protected attributes cannot be disclosed.

Fairness. This work constrains a classifier $\mathcal{M}$ to satisfy a given group fairness notion under a distribution over (X, A, Y) for the protected attribute A as defined below.

Definition 1 (α-fairness)**.** *A model $\mathcal{M}_\theta$ is α-fair w.r.t. a joint distribution (X, A, Y) and fairness function $h(\cdot)$ iff:*

$$\xi(D, h, \theta) = \max_{a \in [m]} |\mathbb{E}_{X,Y|A=a}[h(\mathcal{M}_\theta(X), Y)] - \mathbb{E}_{X,Y}[h(\mathcal{M}_\theta(X), Y)]| \leq \alpha, \qquad (1)$$

wehre $\xi(D, h, \theta)$ is referred to as fairness violation.

The above compares a property for a group of individuals with respect to the whole population and quantifies its difference. Therein, $h(\mathcal{M}_\theta(X), Y) \in \mathbb{R}$, referred here as *fairness function*, defines a target group fairness notion while parameter $\alpha \in \mathbb{R}_+$ represents the *fairness violation*.

The above capture an example class of group fairness notions, including Demographic parity, Equalized odds, and Accuracy parity. A comprehensive review of these fairness definitions is provided in Appendix A.

Differential Privacy. Differential Privacy (DP) [Dwork *et al.*, 2006] is a strong privacy notion used to quantify and bound the privacy loss of an individual participation to a computation. Similarly to [Jagielski *et al.*, 2019; Tran *et al.*, 2021b], this work focuses on the instance where the protection is restricted to the group attributes only. A dataset $D \in \mathcal{D} = (\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$ of size n can be described as a pair (D_P, D_S) where $D_P \in (\mathcal{X} \times \mathcal{Y})^n$ describes the *public* attributes and $D_S \in \mathcal{A}^n$ the group attributes. *The privacy goal is to ensure that the output of the learning model does not significantly change when a single group attribute is changed.*

The action of changing a single attribute from a dataset D_S, resulting in a new dataset D'_S, defines the notion of *dataset adjacency*. Two dataset D_S and $D'_S \in \mathcal{A}^n$ are said adjacent, denoted $D_S \sim D'_S$, if they differ in at most a single entry (e.g., in one individual's group membership).

Definition 2 (Differential Privacy)**.** *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ϵ, δ)-differentially private w.r.t. attribute A, if, for any dataset $D_P \in (\mathcal{X} \times \mathcal{Y})^n$, any two adjacent inputs $D_S, D'_S \in \mathcal{A}^n$, and any subset of output responses $R \subseteq \mathcal{R}$:*

$$\Pr(\mathcal{M}(D_P, D_S) \in R) \leq \exp(\epsilon) \cdot \Pr(\mathcal{M}(D_P, D'_S) \in R) + \delta.$$

When $\delta = 0$ the algorithm is said to satisfy ϵ-DP. 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_f of a real-valued function $f : \mathcal{D} \to \mathbb{R}^k$ is defined as the maximum amount by which f changes in two adjacent inputs D and D': $\Delta_f = \max_{D \sim D'} \|f(D) - f(D')\|$. In particular, the Gaussian mechanism, defined by $\mathcal{M}(D) = f(D) + \mathcal{N}(0, \Delta_f^2\, \sigma^2)$, where $\mathcal{N}(0, \Delta_f^2\, \sigma^2)$ is the Gaussian distribution with 0 mean and standard deviation $\Delta_f\, \sigma$, satisfies (ϵ, δ)-DP for $\delta > \frac{4}{5} \exp(-(\sigma\epsilon)^2/2)$ [Dwork *et al.*, 2006].

4 Private & Fair Learning: Challenges

Constructing models enforcing both privacy and fairness brings with it three fundamental challenges:

- **Scalability**: When interpreted as constraints of the form (1), fairness properties can be explicitly imposed to problem (P). The resulting problem is typically non-convex,

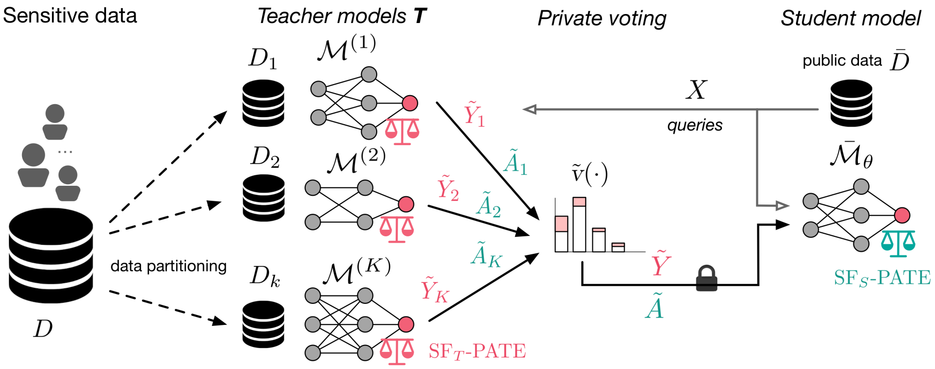


Figure 1: Illustration of the SF-PATE framework. Green (red) colored text and labels depict the SF$_S$ (SF$_T$) version, in which the student (teachers ensemble) is trained under fairness constraints.

but has been approached using a variety of solutions [Agarwal *et al.*, 2018; Du *et al.*, 2021; Kim *et al.*, 2022; Shui *et al.*, 2022]. Researchers have used adaptations of the DP-SGD algorithm [Abadi *et al.*, 2016] to make these methods private. However, the use of sequential clipping operations and fairness constraints in DP-SGD leads to a slow training process, especially in large, overparametrized networks [Subramani *et al.*, 2021]. Even further, the adoption of fairness constraints can also reduce the effectiveness of frameworks such as [Bradbury *et al.*, 2018] and Opacus [Yousefpour *et al.*, 2021], which use vectorization to speed up operations in DP-SGD (see [Tran *et al.*, 2021b], for example).

- **Privacy analysis**: Analyzing the privacy of these methods is also challenging. Ensuring differential privacy through the appropriate amount of noise requires specialized solutions to assess the sensitivity of fairness constraints [Tran *et al.*, 2021b]. These ad-hoc solutions make it difficult to systematically adopt these methods when new fairness notions are introduced or when better fair methods are developed.
- **Poor accuracy/fairness tradeoffs**: Finally, state-of-the-art methods for enforcing privacy in fair learning often sacrifice utility and/or fairness. For example, [Mozannar *et al.*, 2020] perturbs group attributes in a pre-processing step and uses a non-private, fair ML model to post-process the private data. While this avoids the challenge of analyzing privacy, it has been shown to introduce a large amount of unnecessary noise, especially in settings with more than two groups [Tran *et al.*, 2021b].

The approach proposed in this paper avoids these difficulties by providing a teachers ensemble model that **(1)** provides state-of-the-art accuracy, privacy, and fairness tradeoffs, **(2)** adds negligible computational cost to standard (non-private) training, and **(3)** directly exploits the existence of fair (non-private) algorithms. This last property significantly simplifies the engineering of fair and private models, facilitates the adoption of new fairness notions, and avoids the complications introduced by fairness constraints in the DP analysis.

5 The SF-PATE Framework

This section discusses two algorithms for transferring fairness considerations during the private learning process. Both algorithms rely on the presence of an ensemble of *teacher* models $\boldsymbol{T} = \{\mathcal{M}^{(i)}\}_{i=1}^K$, with each $\mathcal{M}^{(i)}$ trained on a non-overlapping portion D_i of the dataset D. This ensemble is used to transfer knowledge to a student model $\bar{\mathcal{M}}_\theta : \mathcal{X} \to \mathcal{Y}$. The student model $\bar{\mathcal{M}}$ is trained using a subset $\bar{D} \subseteq D_P$ whose samples are randomly selected from the original training dataset but whose protected group attributes are unrevealed. As characteristic of ensemble models, the student's queries (data entries $X \in \bar{D}$) are processed by the teacher ensemble to predict the associated label Y through a voting process (which is elaborated below). This process allows the teachers knowledge to be transferred to the student model. The framework, and the two variants introduced next, are depicted schematically in Figure 1.

5.1 Transfer Learning with Fair Student

The first algorithm presented, called SF$_S$-PATE, trains a student model with privacy-preserving group attributes chosen by an ensemble of teachers. Subscript $_S$ in the algorithm's name stands for "student" to emphasize that it is the student that enforces the fairness constraints during training.

The teachers ensemble $\boldsymbol{T} = \{\mathcal{M}^{(i)}\}_{i=1}^K$ is comprised of models $\mathcal{M}^{(i)} : \mathcal{X} \to \mathcal{A}$. Its goal is to predict the group attributes from a sample features, but not the labels as classically done in classification tasks. Note that, importantly, the teacher models are standard classifiers: they are neither private nor fair. Their role is to transfer the information of the group attributes associated with the samples $(X_i, Y_i) \in \bar{D}$ provided by the student. The student model $\bar{\mathcal{M}}_\theta : \mathcal{X} \to \mathcal{Y}$ solves the following regularized constrained minimization:

$$\min_\theta \sum_{(X_i,Y_i)\in\bar{D}} \mathcal{L}(\mathcal{M}_\theta(X_i), Y_i) + \lambda\|\theta - \theta^*\|_2^2 \tag{2a}$$

$$\text{s.t.}\;\; \xi\left(\{X_i, \tilde{v}_A\left(\boldsymbol{T}(X_i)\right), Y_i\}_{(X_i,Y_i)\in\bar{D}}, h, \theta\right) \le \alpha, \tag{2b}$$

where $\tilde{v}_A : \mathcal{A}^K \to \mathcal{A}$ is a private voting scheme returning the group attribute $\tilde{A}$ chosen by the teachers ensemble:

$$\tilde{A} = \tilde{v}_A(\boldsymbol{T}(X)) = \operatorname*{argmax}_{a\in\mathcal{A}}\{\#_a(\boldsymbol{T}(X)) + \mathcal{N}(0,\sigma^2)\}, \quad (3)$$

$\tilde{v}_A$ perturbs the reported counts $\#_a(\boldsymbol{T}(X)) = \left|\{k \in [K] \mid \mathcal{M}_\theta^{(k)}(X)=a\}\right|$ associated to group $a \in \mathcal{A}$ with Gaussian noise of zero mean and standard deviation σ.

Problem (2) minimizes the standard empirical risk function (first component of (2a)), while encouraging the student model parameters θ to be close to the optimal, non-fair, parameters θ^* (second component of (2a)). Note that θ^* can be considered as the most accurate model since it solves problem (P) using all training data D. This model will not leak any private information w.r.t. the protected attributes A because it is trained using only features X and labels Y. The regularization parameter $\lambda > 0$ controls the trade-off between accuracy (when λ is large, the model is pushed towards θ^*) and fairness (when λ is small, the model prioritizes fairness violations). The fairness constraints expressed in Equation (2b) can be enforced through the adoption of off-the-shelf techniques and this paper relies on the use of the Lagrangian dual deep learning framework of [Fioretto *et al.*, 2020]. This choice also allows us to make a fair comparison with state-of-the-art PF-LD [Tran *et al.*, 2021b] in our experiments.

SF$_S$-PATE achieves (ϵ, δ)-DP by introducing calibrated noise on the counts of the protected attributes predicted by the teachers ensemble, as illustrated in Equation (3). The privacy analysis follows from [Papernot *et al.*, 2018] with a few additional considerations and is provided in Appendix C.

Fairness analysis. The rest of this section discusses the fairness guarantees provided by SF$_S$-PATE. The proofs of all theorems are reported in the Appendix B. Note that SF$_S$-PATE relies on a (non-private) fair classifier to train the student model. The performance of many such fair classifiers has been established in the literature. The key question however is to assess the impact of the privacy-preserving voting scheme on the fairness of the student model. The following theorem bounds the fairness violation of the student $\bar{\mathcal{M}}_\theta$ w.r.t. (X, A, Y)—i.e., the original, non-noisy data—when the group attribute A and its privacy-preserving counterpart $\tilde{A}$ are close enough to each other statistically, i.e.,

$$\left|\Pr(\tilde{A} = a \mid x, y) - \Pr(A = a \mid x, y)\right| \leq \eta_A,$$

for events $(X = x, Y = x)$ and value a. Note that, above, η_A characterizes the confidence of the noisy voting process. Small η_A values correspond to ensembles with high prediction agreement (e.g., many teachers predict the same output $\tilde{A}$). Conversely, large η_A values correspond to ensembles with low agreements, rendering the noisy votes less tolerant to noise. Additionally, this property specifies that the group attributes can be inferred from the sample features and labels, which is also what allows a notion of disparity to arise.

Theorem 1. *Let $\bar{\mathcal{M}}_\theta$ be α-fair w.r.t. $(X, \tilde{A}, Y)$ and $h(\cdot)$. Then, $\bar{\mathcal{M}}_\theta$ is α'-fair w.r.t. (X, A, Y) and $h(\cdot)$ with:*

$$\alpha' = \frac{2\eta_A \cdot B}{\min_{a\in[m]} \Pr(A = a) \cdot \Pr(\tilde{A} = a)} + \alpha\,, \quad (4)$$

where B is the supremum of the fairness function $h(\cdot)$.

Thus, this result help us quantifying the fairness guarantees attained on the original data (X, A, Y) when only private noisy data $(X, \tilde{A}, Y)$ (as used by SF$_S$-PATE) is used. Quantity $\Pr(\tilde{A} = a)$, however, may be difficult to compute, due to the noisy voting process. The next result allows us to compute α' as a function of quantity $\Pr(A = a)$ only, which is independent from the noisy process and can be computed empirically from the data, in non-restrictive settings.

Corollary 1. *When the probability of A belonging to any class $a \in [m]$ is at least η_A (i.e., $\Pr(A = a) > \eta_A$, $\forall\, a \in [m]$), the fairness bound α' can then be given by*

$$\alpha' = \frac{2\eta \cdot B}{\min_{a\in[m]} \Pr(A = a) \cdot (\Pr(A = a) - \eta)} + \alpha\,.$$

The appendix in [Tran *et al.*, 2022] also provides a result (described in Theorem 3) to compare the fairness guarantees achieved when the protected attributes $\tilde{A}$ are computed using randomized response, which is the core process used by the baseline and state of the art model *M* [Mozannar *et al.*, 2020]. This result shows that, when compared with the baseline model *M*, SF$_S$-PATE makes the privacy-preserving group attributes $\tilde{A}$ much closer to their original counterparts A under the same privacy budgets. This aspect is also further discussed in the experiments (as emphasized in Figure 3).

5.2 Transfer Learning with Fair Teachers

SF$_S$-PATE transfers privacy-preserving group attributes from an ensemble of teachers to a fair student model. This section introduces an SF-PATE variant that transfer (non-private) fairness through the semi-supervised learning scheme. The proposed algorithm, called SF$_T$-PATE, trains a student model from an ensemble of fair teacher models. Subscript $_T$ in the algorithm's name stands for "teachers" and emphasizes that the fairness constraints are enforced by the teachers and transferred to the student during training.

The teachers transfer fairness properties via model prediction to the student. The teachers ensemble $\boldsymbol{T} = \{\mathcal{M}^{(i)}\}_{i=1}^K$ is composed of pre-trained classifiers $\mathcal{M}^{(i)} : \mathcal{X} \to \mathcal{Y}$ that are non-private but fair over their training data D_i. Each SF$_T$-PATE teacher solves an empirical risk minimization problem subject to fairness constraint:

$$\theta_i = \operatorname*{argmin}_{\theta} \sum_{(X,Y)\in D_i} \mathcal{L}(\mathcal{M}_\theta(X_i), Y_i) \quad \text{s.t. } \xi(D_i, h, \theta) \leq \alpha.$$

Similar to SF$_S$-PATE, the implementation uses the Lagrangian dual method [Fioretto *et al.*, 2020] to achieve this goal. This is again an important aspect of the SF-PATE framework since, by relying on black-box fair (and not private) algorithms, it decouples the dependency of developing joint private and fair analysis. The student model $\bar{\mathcal{M}}_\theta : \mathcal{X} \to \mathcal{Y}$ solves a standard empirical risk minimization:

$$\min_{\theta} \sum_{X\in\bar{D}} \mathcal{L}\left(\mathcal{M}_\theta(X), \tilde{v}_Y(\boldsymbol{T}(X))\right) + \lambda\|\theta - \theta^*\|_2^2, \quad (5)$$

where θ^* represents the student model parameters obtained solving the standard classification task (without fairness considerations) (P) on $\bar{D}$ and the private voting scheme $\tilde{v}_Y : \mathcal{Y}^K \to \mathcal{Y}$ reports the label $\tilde{Y}$ selected by the teachers:

$$\tilde{Y} = \tilde{v}(\boldsymbol{T}(X)) = \operatorname*{argmax}_{y \in \mathcal{Y}} \{\#_y(\boldsymbol{T}(X) + \mathcal{N}(0, \sigma^2)\}. \quad (6)$$

SF$_T$-PATE protects the privacy of the sensitive group information A and of the labels Y when $\lambda = 0$. The reason comes from the fact that when $\lambda = 0$, the student only utilize the private class label $\tilde{v}_Y(\boldsymbol{T}(X)$ but not the true label Y, implicitly encoded in θ^*, to train the model. Thus, SF$_T$-PATE can be adopted in contexts where both the protected group and the labels are sensitive information. Finally, note that the voting scheme above, emulates the GNMAX framework proposed in [Papernot *et al.*, 2018], but the two have fundamentally different goals: GNMAX aims at protecting the participation of individuals into the training data, while SF$_T$-PATE aims at privately transferring the fairness properties of the teachers ensemble to a student model.

SF$_T$-PATE achieves (ϵ, δ)-DP by introducing calibrated noise on the counts of the predicted labels reported by the teachers ensemble (see Appendix C).

Fairness analysis. The next results considers how well the fairness properties of the ensemble are transferred to the student model. They provides theoretical insights on the potential fairness violations induced by the voting mechanism $\tilde{v}_Y$. For notational convenience, let $B := \sup_{y \in \mathcal{Y}, y' \in \mathcal{Y}} h(y, y')$ denote the supremum of the fairness function $h(\cdot)$. Let Z be the random vector $(\mathcal{M}_\theta^{(1)}(X), \ldots, \mathcal{M}_\theta^{(K)}(X), Y)$ while Z_a the conditional random vector given the group label $a \in [m]$, i.e., $Z_a := (\mathcal{M}_\theta^{(1)}(X), \ldots, \mathcal{M}_\theta^{(K)}(X), Y) \mid A = a$.

The following theorem bounds the fairness violation of the voting mechanism $\tilde{v}_Y(\boldsymbol{T})$ when the joint distributions of the teachers ensemble and labels Y are roughly similar across the different group classes. That is, for each group attribute $a \in [m]$, the total variation distance between Z and Z_a is bounded from the above by $\eta_Y > 0$, i.e.,

$$\mathrm{d}_{\mathrm{TV}}(Z, Z_a) := \sup_{S \subseteq \Omega} |\Pr(Z \in S) - \Pr(Z_a \in S)| \leq \eta_Y,$$

where Ω is used to denote the shared probability space.

Theorem 2. *The voting mechanism $\tilde{v}_Y(\boldsymbol{T})$ is α' fair w.r.t. (X, A, Y) and $h(\cdot)$ with $\alpha' = \eta_Y \cdot B$.*

The result above shed light on the (non-restrictive) conditions required for the ensemble votes to transfer fairness knowledge accurately. The next corollary is a direct consequence of Theorem 2 and provides a sufficient condition for perfect fairness of the voting mechanism $\tilde{v}_Y(\boldsymbol{T})$.

Corollary 2. *Suppose that the random vector Z is independent of A, i.e., $\{Z_a\}_{a \in [m]}$ and Z are identically distributed. Then, the voting mechanism $\tilde{v}_Y(\boldsymbol{T})$ is perfectly fair (i.e., fairness violation $\alpha = 0$) w.r.t. (X, A, Y) and $h(\cdot)$.*

These results are reassuring: informally speaking, they show that SF$_T$-PATE provides fairness guarantees that are "inversely propertional" to the predictive impact of the protected attributes. The experimental results presented subsequently confirm that SF$_T$-PATE achieves state-of-the-art tradeoffs among accuracy, privacy, and fairness.

6 Experiments

This section evaluates the performance of the SF-PATE algorithms against the prior approaches of [Tran *et al.*, 2021b] and [Mozannar *et al.*, 2020], denoted by *PF-LD* and *M*, respectively. They represents the state-of-the-art for learning private and fair classifiers in the context studied in this paper. In addition to assess the competitiveness of the proposed solutions, the evaluation focuses on two key aspects that set the proposed framework apart from existing methods. **(1)** It shows that SF-PATE can naturally handle new fairness notions, even when no viable privacy analysis exists about these notions. **(2)** It shows that SF-PATE has a low computational overhead compared to classical (non-private, non-fair) classifiers, rendering it a practical choice for the training of very large models. These two properties are unique to SF-PATE and make it to applicable to a broad class of challenging decision tasks.

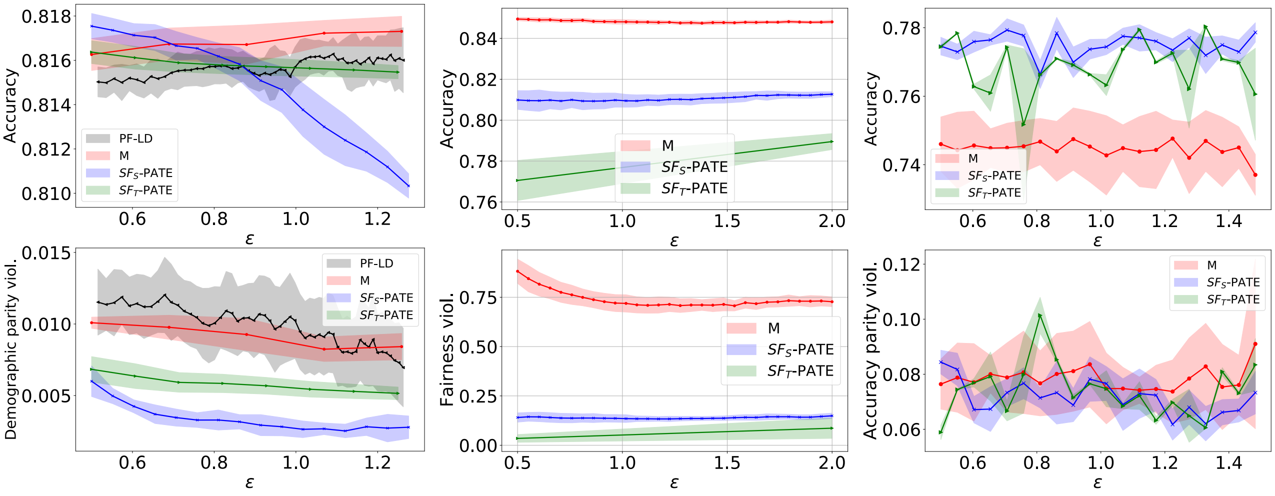
Datasets and settings. The evaluation is conducted using four UCI tabular datasets: Bank, Parkinson, Income and Credit Card [Blake, 1998], and UTKFace [Hwang *et al.*, 2020], a vision dataset. The latter is used to demonstrate the scalability of SF-PATE when trained on very large models. All experiments are repeated using 100 random seeds.

Models and hyperparameters. To ensure a fair comparison, the experimental analysis uses the same architectures, model initialization θ, and parameters for all models (including the baselines models *PF-LD* and *M*). For tabular datasets, the underlying classifier is a feedforward neural network with two hidden layers and nonlinear ReLU activations. The fair, non-private, classifiers adopted by the two SF-PATE variants implement a Lagrangian dual scheme [Fioretto *et al.*, 2020], which is also the underlying scheme adopted by baseline models. For vision tasks on the UTK-Face dataset, the evaluation uses a Resnet 50 classifier. A more detailed description of these approaches, the settings adopted, hyperparameters optimization, and the datasets is deferred to Appendix D.

6.1 Accuracy, Privacy, and Fairness Trade-off

We now compare the accuracy, fairness, and privacy tradeoffs of the proposed model variants SF$_S$- and SF$_T$-PATE against the baseline models *PF-LD* and *M* on the tabular datasets. Figure 2(a) illustrates the accuracy (top subplots) and fairness violations $\xi(\theta, h, \bar{D})$ (bottom subplots) when varying the privacy loss ϵ (x-axis). The fairness notion adopted is demographic parity and additional results on other fairness metrics and datasets are deferred to the Appendix (in [Tran *et al.*, 2022]), showing similar trends. The figures clearly illustrate that both SF-PATE variants achieve better accuracy/fairness tradeoffs for various privacy parameters. The property of retaining high accuracy with low fairness violation is especially relevant in the tight privacy regime adopted ($\epsilon < 2$). Observe that the figures y-axes have different scales.

First, notice that, consistently with previous work showing that teachers ensemble models can outperform DP-SGD



(a) Demographic parity fairness on Credit Card dataset

(b) Generalized demographic parity fairness on Income dataset

(c) Accuracy parity fairness on UTK Face dataset

Figure 2: Accuracy, privacy and fairness trade-offs for different fairness metrics and on different datasets.

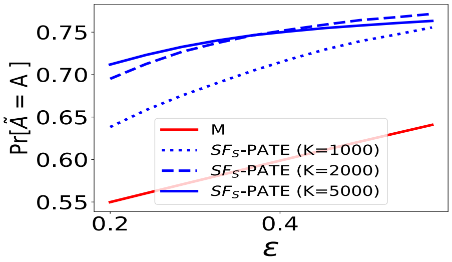


Figure 3: Income data: Private group attributes accuracy.

based models in terms of accuracy [Uniyal *et al.*, 2021], all SF-PATE models typically outperform the *FP-LD* model based on DP-SGD. Remarkably, both SF-PATE variants also report lower fairness violations, for the fairness notions analyzed, indicating the strength of this approach.

Additionally, recall that both SF$_S$-PATE and *M* generate privacy-preserving group features and input them to a fair model. However, in contrast to *M*, the model ensemble used in SF$_S$-PATE exploits the relationship between the sample features X and its associated group information A to derive more accurate private group information $\tilde{A}$. This results in student models which are often more accurate, and fairer, than the baseline *M*. This result is further highlighted in Figure 3, which reports the expected accuracy of the private group attributes $\Pr(\tilde{A}_i = A_i)$ produced by SF$_S$-PATE for different sizes K of the ensemble and compares it with that obtained by model *M*. Note the distinctive ability of the teachers ensemble to generate high-quality privacy-preserving group attributes $\tilde{A} = \tilde{v}_A(\boldsymbol{T}(X))$, especially under tight privacy regimes. Additionally, increasing the ensemble size also enables the teachers to transfer higher quality private group attributes, which is a fundamental aspect to ensure fairness.

The second SF-PATE variant, SF$_T$-PATE, which operates by privately transferring the fairness knowledge from a teachers ensemble to a student, is found to always outperform both *M* and *PF-LD* on tabular datasets (see Figure 2(a) and Appendix D). Finally, the analysis also shows that the average accuracy of the SF-PATE models is within 2% of their non-private counterpart and brings up to an order of magnitude fairness violation gains over existing methods. *This is significant due to the challenging nature of the tasks.*

6.2 Handling New Group Fairness Notions

The next results consider the ability of SF-PATE to handle arbitrary fairness notions, even if a privacy analysis is missing, as long as a fair model can be derived. This feature stems from the use of black-box (non-private but fair) models in SF-PATE. This is in sharp contrast with state-of-the-art model *PF-LD*, that requires the development of a privacy analysis for each fairness property of interest, in order to calibrate the amount of noise to apply in both primal and dual step. To demonstrate this benefit, this section introduces a new fairness notion, which generalizes demographic parity.

Definition 3 (Generalized demographic parity)**.** *requires the distribution over the predictions of $\mathcal{M}$ to be statistically independent of the protected group attribute A. That is,* $\forall a \in \mathcal{A}, \eta \in [0,1]$, $\Pr(\mathcal{M}(X) \geq \eta | A = a) = \Pr(\mathcal{M}(X) \geq \eta)$.

This generalizes demographic parity, which states that $\Pr\left(\mathcal{M}_\theta(X) \geq 0.5 \mid A = a\right) = \Pr\left(\mathcal{M}_\theta(X) \geq 0.5\right)$. Generalized demographic parity is useful in settings where the decision threshold (e.g., 0.5 above) might not be available at the training time. Matching the distribution of score functions (e.g., credit or income scores) among different groups attributes guarantees demographic parity fairness regardless of the decision threshold adopted. Such fairness constraint can be implemented by equalizing different order statistics of the score functions between group and population level:

$$\mathbb{E}[\mathcal{M}_\theta(X)^h | A = a] = \mathbb{E}[\mathcal{M}_\theta(X)^h] \quad \forall\, a \in [m], h \in [H].$$

Dataset	SF$_S$-PATE	SF$_T$-PATE	M	PF-LD
Bank	**14**	**13**	31	116
Parkinson	**8**	**8**	17	31
Income	**55**	**56**	129	1234
Credit Card	**30**	**31**	76	575
UTK dataset	**1669**	**1662**	3248	N/A

Table 1: Runtime (in sec.) to achieve $\epsilon = 1.0$ across different datasets. Blue and red colors highlight the fastest and second-fastest runtime, respectively. Ensemble size for SF-PATE models is 300.

The experiments H = 2 and the Lagrangian Dual method of [Fioretto *et al.*, 2020] to enforce these constraints during training. *Notice that it is highly non-trivial to derive a privacy analysis for such fairness notion*–the PF-LD model only does so for $H = 1$, and, thus, not viable in this setting.

Figure 2(b) reports the accuracy and fairness violations obtained by the SF-PATE models and the baseline model *M* which adopts, as a post-processing step, a non-private but fair classifier. Fairness is evaluated in terms of the Wasserstein distance between the score functions of different groups. The smaller the distance the lower the fairness violation. The plots clearly illustrates the advantages of SF-PATE in terms of both accuracy and fairness when compared to model *M*. Remarkably, the fairness violations reported by SF-PATE are often significantly lower than those reported by model *M* for various privacy loss parameters ϵ.

These results are significant from an engineering standpoint: the development of private and fair analysis is complex and SF-PATE immediately lowers the accessibility barrier for non-privacy experts to develop private and fair ML models.

6.3 Computational Time and Scalability

The last results demonstrate another key benefits of SF-PATE: its ability to scale to large data and perform well on very deep networks. These experiments use a ResNet 50 ($>$ 23M parameters) and PF-LD was unable to train even a single epoch (in 1h) due to its use of computational expensive per-sample gradient clipping performed during training. This renders such model unusable for many realistic settings. The comparison thus focuses on SF-PATE and *M*.

Figure 2(c) shows the accuracy and fairness trade-off for different privacy values ϵ. The models enforce accuracy parity and the figure shows that both versions of SF-PATE significantly improve accuracy compared to model M, while also maintaining similar or even reducing fairness violations. For additional results see Appendix D.

Table 1 shows the training time required for the algorithms to create a private model (with $\epsilon = 1.0$) on the benchmarks set. Notice the significant training time differences across the models with SF$_S$-PATE being up to three times faster than *M* and up to two orders of magnitude faster than *PF-LD*.

These results show that SF-PATE can become a practical tool for private and fair ML especially for large, overparametrized models and under stringent privacy regimes.

6.4 Discussion and Limitations

While the previous section highlighted the advantages of the proposed framework over existing models, this section sheds light on key elements and usage guidelines for SF$_S$- and SF$_T$-PATE. While both SF-PATE variants rely on the same framework, they diverge in the fairness enforcing mechanism. SF$_S$-PATE delegates this to its student while SF$_T$-PATE to its teachers. Notice that, by training an ensemble of fair teachers, SF$_T$-PATE is able to transfer fairness knowledge *directly* which, as observed in the experiments, often results in smaller fairness violations than those attained by SF$_S$-PATE. This is notable, especially in the case of “complex” fairness constraints; i.e., when multiple concurrent constraints are to be enforced, as in the case of generalized demographic parity which imposes multiple order moments matching. Being treated as soft penalty functions, these constraints are added to the original empirical loss function (see Problem (2)). SF$_S$-PATE adds noise to each constraint terms (e.g., those represented by the constrains in (2b)) since the private voting scheme affects the accuracy of the constraints. In contrast, SF$_T$-PATE adds noise only to the original loss term (e.g., the first term of (2a)) because the voting scheme acts on the labels Y and not the protected groups A and fairness is enforced by the teachers.

A shortcoming of the SF$_T$-PATE algorithm, however, is that it has to enforce fairness in each teacher model. Thus, one has to ensure that enough data (with large enough representation from all the protected groups) is assigned to each teacher. On the other hand, by training a *single* fair model (the student) SF$_S$-PATE avoids such potential issue.

It is also worth noting that a limitation of all ensemble models, including those proposed in this work, is the need to store a model for each of the K teachers. This, however, also represents an opportunity for future research to develop effective model storage and pruning techniques that minimize the loss in accuracy and fairness while retaining privacy.

7 Conclusions

This paper proposed a framework to train deep learning models that satisfy several notions of group fairness while ensuring that the model satisfies differential privacy for the protected attributes. The proposed framework, called *Scalable, Fair, and Private Aggregation of Teacher Enseble* (SF-PATE) transfer fairness knowledge learned from a pretrained ensemble of models to a target model via a privacy-preserving voting process. The paper analyzes the fairness properties of SF-PATE and shows that unfairness can be bounded in many practical settings. An important property of SF-PATE is to allow the adoption of black-box (non-private) fair models during the knowledge transfer process, which may simplify the development and boosts the adoption of new fairness metrics in privacy-preserving ML.

Finally, evaluation on both tabular and image datasets shows not only that SF-PATE achieves better accuracy, privacy, and fairness tradeoffs with respect to the current state-of-the-art, but it is also significantly faster. These properties render SF-PATE amenable to train large, overparameterized, models, that ensure privacy, accuracy, and fairness simultaneously, showing that it may become a practical tool for privacy-preserving and fair decision making.

Ethical Statement

This research aims to address the ethical concern of building models that do not discriminate against certain demographic groups by studying a model that protects the privacy of individuals' sensitive information while also allowing it to learn non-discriminatory predictors. The proposed model allows the use of existing fair models to create a privacy-preserving and fair model, and it is designed to be scalable and accurate for very large neural networks. The paper analyze the relationship between accuracy, privacy, and fairness and evaluate the benefits of the proposed model on various prediction tasks. Overall, this research aims to promote ethical and fair practices in data-driven processes by finding ways to balance privacy and fairness concerns.

Acknowledgements

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

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[Bagdasaryan *et al.*, 2019] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2019.

[Blake, 1998] Catherine Blake. Uci repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998. Accessed: 2023-01-30.

[Bradbury *et al.*, 2018] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. http://github.com/google/jax, 2018. Accessed: 2023-01-30.

[Chaudhuri *et al.*, 2011] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 2011.

[Du *et al.*, 2021] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34:12091–12103, 2021.

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

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

[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.*, 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.

[Hwang *et al.*, 2020] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. In *31st British Machine Vision Conference 2020, BMVC 2020*, 2020.

[Jagielski *et al.*, 2019] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *ICML*, pages 3000–3008, 2019.

[Kim *et al.*, 2022] Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. *arXiv preprint arXiv:2202.02943*, 2022.

[Mozannar *et al.*, 2020] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *ICML*, pages 7066–7075, 2020.

[Papernot *et al.*, 2018] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[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 *ACM FAccT*, pages 189–199, 2020.

[Shui *et al.*, 2022] Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[Subramani *et al.*, 2021] Pranav Subramani, Nicholas Vadivelu, and Gautam Kamath. Enabling fast differentially private sgd via just-in-time compilation and vectorization. *Advances in Neural Information Processing Systems*, 34:26409–26421, 2021.

[Tran and Fioretto, 2023] Cuong Tran and Ferdinando Fioretto. A fairness analysis on private aggregation of teacher ensembles. In *International Joint Conference on Artificial Intelligence (IJCAI)*, page TBD, 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, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9932–9939, 2021.

[Tran *et al.*, 2021c] Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under the fairness lens. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 560–566, 2021.

[Tran *et al.*, 2022] Cuong Tran, Keyu Zhu, Ferdinando Fioretto, and Pascal Van Hentenryck. SF-PATE: Scalable, Fair, and Private Aggregation of Teacher Ensembles. *arXiv preprint arXiv:2204.05157*, 2022.

[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.

[Xu *et al.*, 2019] Depeng Xu, Shuhan Yuan, and Xintao Wu. Achieving differential privacy and fairness in logistic regression. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 2019.

[Yousefpour *et al.*, 2021] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.