Differentially Private and Fair Deep Learning:A Lagrangian Dual Approach

Cuong Tran[1], **Ferdinando Fioretto**[1], **Pascal Van Hentenryck**[2]
[1] Syracuse University, [2] Georgia Institute of Technology
cutran@syr.edu, ffiorett@syr.edu, pvh@isye.gatech.edu

Abstract

A critical concern in data-driven decision making is to build models whose outcomes do not discriminate against some demographic groups, including gender, ethnicity, or age. To ensure non-discrimination in learning tasks, knowledge of the sensitive attributes is essential, while, 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 the individuals' sensitive information while also allowing it to learn non-discriminatory predictors. The method relies on the notion of differential privacy and the use of Lagrangian duality to design neural networks that can accommodate fairness constraints while guaranteeing the privacy of sensitive attributes. The paper analyses the tension between accuracy, privacy, and fairness and the experimental evaluation illustrates the benefits of the proposed model on several prediction tasks.

Introduction

A number of socio-technical decisions, such as criminal assessment, landing, and hiring, are increasingly being aided by machine learning systems. A critical concern is that the learned models are prone to report outcomes that are discriminatory against some demographic group, including gender, ethnicity, or age. These concerns have spurred the recent development of fairness definitions and algorithms for decision-making, focusing attention on the tradeoff between the model accuracy and fairness.

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. Other requirements may be even more stringent, and prevent the collection of protected user attributes, such as for the case of racial attributes in the E.U. General Data Protection Regulation (GDPR), or require protection of the consumer data privacy. 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, and (3) the restriction on the data or protected attributes that can be used. There is thus a need to provide learning models that can both guarantee non discriminatory decisions and protect the privacy of the individuals' sensitive attributes.

To this end, this paper introduces a differential privacy framework to train deep learning models that satisfy several group fairness notions, including *equalized odds*, *accuracy parity*, and *demographic parity* (Zafar et al. 2017a; Hardt et al. 2016; Agarwal et al. 2018), while providing privacy of the protected attributes. The key elements of the framework can be summarized as follows:

1. The fairness requirement is captured by casting the learning task as a constrained optimization problem. A Lagrangian dual approach is then applied to the learning task, dualizing the fairness constraints using augmented Lagrangian terms (Hestenes 1969).
2. The privacy requirement is enforced by using a *clipping approach* on the primal and dual steps and adding noise calibrated by the sensitivities of the constraint terms and their gradients. The primal step only applies clipping on constraint gradients involving sensitive attributes, and thus, does not have a major effect on the model accuracy.
3. The framework addresses the bias-variance trade-off of clipping by providing bounds on the expected errors of constraint gradients and constraint violations. The clipping bounds can then be calibrated by minimizing these upper bounds.

The rest of the paper presents the proposed *Private and Fair Lagrangian Dual* (PF-LD) framework, its theoretical results, and its empirical evaluation on several prediction tasks. The empirical results show that, on selected benchmarks, PF-LD achieves an excellent trade-off among accuracy, privacy, and fairness. It may represent a promising step towards a practical tool for privacy-preserving and fair decision making.

An extended version of this work, that includes proofs of all theorems, can be found in (Tran, Fioretto, and Van Hentenryck 2020).

Related Work

The topics of privacy and fairness have been study mostly in isolation. A few exceptions are represented by the work of Dwork et al. (2012), which is one of the earliest contribution linking fairness and differential privacy, showing that

individual fairness is a generalization of differential privacy. More recently, Cummings et al. (2019) study the tradeoff between differential privacy and equal opportunity, a notion of fairness that restricts a classifier to produce equal true positive rates across different groups. The work shows that there is no classifier that achieves $(\epsilon, 0)$-differential privacy, satisfies equal opportunity, and has accuracy better than a constant classifier. Ekstrand, Joshaghani, and Mehrpouyan (2018) raise questions about the tradeoff between privacy and fairness and, Jagielski et al. (2019) and Mozannar, Ohannessian, and Srebro (2020) propose two simple, yet effective algorithms that satisfy (ϵ, δ)-differential privacy and equalized odds. Xu, Yuan, and Wu (2019) proposes a privacy-preserving and fair logistic regression model making use of the functional mechanism (Chaudhuri, Monteleoni, and Sarwate 2011). Finally, a recent line of work has also observed that private models may have a negative impact towards fairness (Pujol et al. 2020; Bagdasaryan, Poursaeed, and Shmatikov 2019). In contrast to the work discussed above, this paper, presents a Lagrangian dual method to enforce several fairness constraints directly into the training cycle of a deep neural network and proposes a differentially private and fair version of the learning algorithm.

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. Therein, $X_i \in \mathcal{X}$ is a *non-sensitive* feature vector, $A_i \in \mathcal{A}$, with $\mathcal{A} = [m]$ (for some finite m) is a protected attribute[1], and $Y_i \in \mathcal{Y} = \{0, 1\}$ is a binary 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 sensitive attribute A. The model quality is measured in terms of a nonnegative, 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 minimization (ERM) function:

$$\min_\theta J(\mathcal{M}_\theta, D) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathcal{M}_\theta(X_i), Y_i). \qquad \text{(L)}$$

The paper focuses on learning general classifiers, such as neural networks, that satisfy group fairness (as defined next) and protect the disclosure of the sensitive 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.

Preliminaries: Fairness

The paper consider a classifier $\mathcal{M}$ satisfying some group fairness notion under a distribution over (X, A, Y) for the protected attribute A and focuses on three fairness notions:

- *Demographic Parity*: $\mathcal{M}$'s predictions are statistically independent of the protected attribute A. That is,

  $$\Pr[\mathcal{M}(X) = \hat{y} \mid A = a] = \Pr[\mathcal{M}(X) = \hat{y}] \quad \forall a \in \mathcal{A}, \hat{y} \in \mathcal{Y},$$

  which, since $\hat{y} \in \{0, 1\}$, can be expressed as

  $$\mathbb{E}[\mathcal{M}(X) \mid A = a] = \mathbb{E}[\mathcal{M}(X)], \quad \forall a \in \mathcal{A}.$$

- *Equalized odds*: $\mathcal{M}$'s predictions are conditionally independent of the protected attribute A given the label Y. That is, for all $a \in \mathcal{A}, \hat{y} \in \mathcal{Y}$, and $y \in \mathcal{Y}$:

  $$\Pr[\mathcal{M}(X) = \hat{y} \mid A = a, Y = y] = \Pr[\mathcal{M}(X) = \hat{y} \mid Y = y].$$

  or, equivalently, for all $a \in \mathcal{A}, y \in \mathcal{Y}$,

  $$\mathbb{E}[\mathcal{M}(X) \mid A = a, Y = y] = \mathbb{E}[\mathcal{M}(X) \mid Y = y].$$

- *Accuracy parity*: $\mathcal{M}$'s miss-classification rate is conditionally independent of the protected attribute:

  $$\Pr[\mathcal{M}(X) \neq Y \mid A = a] = \Pr[\mathcal{M}(X) \neq Y], \quad \forall a \in \mathcal{A},$$

  or equivalently,

  $$\mathbb{E}[\mathcal{L}(\mathcal{M}(X), Y) \mid A = a] = \mathbb{E}[\mathcal{L}(\mathcal{M}(X), Y)], \quad \forall a \in \mathcal{A},$$

  where $\mathcal{L}$ is the loss function to minimize in problem (L).

As noted by Agarwal et al. (2018) and Fioretto et al. (2020), several fairness notions, including those above, can be viewed as equality constraints between the properties of each group with respect to the population. These constraints can be expressed as:

$$\mathbb{E}_{z \sim D_{P_i}}[h(z)] - \mathbb{E}_{z \sim D_{G_i}}[h(z)] = 0 \qquad (1)$$

where, for i in some index set $\mathcal{I}$, D_{P_i} is a subset of the dataset D, indicating the *population term*, D_{G_i} is a subset of D_{P_i}, indicating the *group term*, and is obtained by accessing the protected attributes A, the function h characterizes the model output under some fairness definition.

Since the joint data distribution over (X, A, Y) is unknown, the above uses its empirical mean $\hat{\mathbb{E}}_D$, which can be estimated from the training data D.

Example 1 (Demographic parity)**.** *Demographic parity can be expressed as a set of $|\mathcal{A}|$ constraints, with $h(z) = \mathcal{M}_\theta(z)$ and, for each $i \in \mathcal{A}$, the subsets indicating population terms are defined as:*

$$D_{P_i} = \{(X, Y) \mid (X, A, Y) \in D\},$$

and the subsets indicating the group terms as:

$$D_{G_i} = \{(X, Y) \mid (X, A, Y) \in D \wedge A = i\}.$$

Preliminaries: Differential Privacy

Differential privacy (DP) (Dwork et al. 2006) is a strong privacy notion used to quantify and bound the privacy loss of an individual participation to a computation. While traditional DP protects the participation of an individual to a dataset used in a computation, similarly to (Jagielski et al. 2019; Mozannar, Ohannessian, and Srebro 2020), this work focuses on the instance where the protection is restricted to the sensitive attributes only. A dataset $D \in \mathcal{D} = (\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$

[1] While this notation simplifies exposition, the method proposed is not limited to the case where there is a single protected attribute.

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 $D_S \in \mathcal{A}^n$ describes the sensitive attributes. *The privacy goal is to guarantee that the output of the learning model does not differ much when a single individual sensitive 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 1 (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 $R \subseteq \mathcal{R}$:*

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

When $\delta = 0$ the algorithm is said to satisfy ϵ-differential privacy. 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-differential privacy. 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 $\mathcal{N}(0, \Delta_f\, \sigma^2)$ is the Gaussian distribution with 0 mean and standard deviation $\Delta_f\, \sigma^2$, satisfies (ϵ, δ)-DP for $\delta > \frac{4}{5} \exp(-(\sigma\epsilon)^2/2)$ and $\epsilon < 1$ (Dwork, Roth et al. 2014).

Constrained Learning with Lagrange Duality

When interpreted as constraints of the form (1), fairness properties can be explicitly imposed to problem (L), resulting in a constrained ERM problem. Solving this new problem, however, becomes challenging due to the presence of constraints. To address this challenge, this work uses concepts borrowed from Lagrangian duality.

Consider a set of $|\mathcal{I}|$ constraints of the form (1), and expressed succinctly as:

$$\boldsymbol{\mu}(D_P) - \boldsymbol{\mu}(D_G) = \mathbf{0}^\top \tag{2}$$

where $\boldsymbol{\mu}(D_P)$ and $\boldsymbol{\mu}(D_G)$ are vectors containing elements $\mu(D_{P_i}) = \hat{\mathbb{E}}_{z \sim D_{P_i}}[h(z)]$ and $\mu(D_{G_i}) = \hat{\mathbb{E}}_{z \sim D_{G_i}}[h(z)]$, respectively, for each $i \in \mathcal{I}$.

Notice that the constraints in $\boldsymbol{\mu}(D_P)$ access public data only, while the constraints in $\boldsymbol{\mu}(D_G)$ access also the sensitive data. Notice also that the paper does not restrict the setting to the case where only demographic attributes are sensitive and, in particular, it does not assume that $D_S = D_G$. The resulting learning task is defined by the following optimization problem

$$\operatorname*{argmin}_{\theta} J(\mathcal{M}_\theta, D_P) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathcal{M}_\theta(X_i), Y_i) \tag{3a}$$

$$\text{subject to} \qquad \boldsymbol{\mu}(D_P) - \boldsymbol{\mu}(D_G) = \mathbf{0}^\top. \tag{3b}$$

Algorithm 1: *Fair-Lagrangian Dual (F-LD)*

input : $D = (X_i, A_i, Y_i)_{i=1}^n$: Training data;
$\alpha, \boldsymbol{s} = (s_1, s_2, \ldots)$: step sizes.
$\lambda^{\max}$: Max multipliers value.

1 $\lambda_{1,i} \leftarrow 0 \;\; \forall i \in \mathcal{I}$
2 **for** *epoch* $k = 1, 2, \ldots T$ **do**
3 **foreach** *Mini-batch* $B \subseteq D$ **do**
4 $\theta \leftarrow \theta - \alpha \nabla_\theta \big[J(\mathcal{M}_\theta, B_P) + \boldsymbol{\lambda}_k^\top |\boldsymbol{\mu}(B_P) - \boldsymbol{\mu}(B_G)| \big]$
5 $\boldsymbol{\lambda}_{k+1} \leftarrow \boldsymbol{\lambda}_k + s_k \, |\boldsymbol{\mu}(D_P) - \boldsymbol{\mu}(D_G)|$
6 $\lambda_{k+1,i} \leftarrow \min(\lambda^{\max}, \lambda_{k+1,i}) \;\; \forall i \in \mathcal{I}$

In *Lagrangian relaxation*, the problem constraints are relaxed into the objective function using *Lagrangian multipliers* $\lambda_i \geq 0$ associated to each of the $|\mathcal{I}|$ constraints and expressing the penalty induced by violating them. When all the constraints are relaxed, the *Lagrangian function* becomes

$$\mathcal{L}_{\boldsymbol{\lambda}}(\theta) = J(\mathcal{M}_\theta, D_P) + \boldsymbol{\lambda}^\top |\boldsymbol{\mu}(D_P) - \boldsymbol{\mu}(D_G)|, \tag{4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{|\mathcal{I}|})$ and the function $|\cdot|$, used here to denote the element-wise operator (i.e, $|\mu(D_{P_i}) - \mu(D_{G_i})|$ for $i \in \mathcal{I}$), captures a quantification of the constraint violations, often used in constraint programming (Fontaine, Laurent, and Van Hentenryck 2014).

Using a Lagrangian function, the optimization becomes

$$\overset{\star}{\theta}(\boldsymbol{\lambda}) = \operatorname*{argmin}_{\theta} \mathcal{L}_{\boldsymbol{\lambda}}(\theta), \tag{5}$$

that produces an approximation $\mathcal{M}_{\overset{\star}{\theta}(\boldsymbol{\lambda})}$ of $\mathcal{M}_{\overset{\star}{\theta}}$. The Lagrangian dual finds the best Lagrangian multipliers, i.e.,

$$\overset{\star}{\boldsymbol{\lambda}} = \operatorname{argmax}_{\boldsymbol{\lambda} \geq 0} J(\mathcal{M}_{\overset{\star}{\theta}(\boldsymbol{\lambda})}, D_P) \tag{6}$$

to obtain $\mathcal{M}_{\overset{\star}{\theta}(\overset{\star}{\boldsymbol{\lambda}})}$, i.e., the strongest Lagrangian relaxation of $\mathcal{M}$. Learning this relaxation relies on an iterative scheme that interleaves the learning of a number of Lagrangian relaxations (for various multipliers) with a subgradient method to learn the best multipliers. The resulting method, called *Fair-Lagrangian Dual* (F-LD) is sketched in Algorithm 1. Given the input dataset D, the optimizer step size $\alpha > 0$, and the vector of step sizes $\boldsymbol{s}$, the Lagrangian multipliers are initialized in line 1. The training is performed for a fixed number of T epochs. At each epoch k, the *primal update* step (lines 3 and 4) optimizes the model parameters θ using stochastic gradient descent over different mini-batches $B \subseteq D$. The optimization step uses the current Lagrangian multipliers $\boldsymbol{\lambda}_k$. Therein, B_P and B_G indicate the population and group terms over a minibatch. After each epoch, the *dual update* step (line 5), updates the value of the Lagrangian multipliers following to a *dual ascent* rule (Boyd et al. 2011; Fioretto, Mak, and Van Hentenryck 2020). The multipliers values are thus restricted to a predefined upper bound $\lambda^{\max}$ (line 6).

A Private and Fair LD Model

To ensure fairness, the primal (line 4) and dual (line 5) updates of Algorithm 1 involve terms to compute the violations associated to constraints (3b). These terms rely on the

attributes A, and therefore, the resulting model leaks the sensitive information. To contrast this issue, this section introduces an extension to F-LD, called *Private and Fair Lagrangian Dual (PF-LD)* method, that guarantees both fairness and privacy. The idea is to render the computations of the primal and dual update steps differentially private with respect to the sensitive attributes.

Private Primal Update At each epoch k, the primal update (line 4 of Algorithm 1) computes the gradients over the loss function $\mathcal{L}_{\boldsymbol{\lambda}_k}(\theta)$, which is composed of two terms (see Equation (4)). The first term, $J(\mathcal{M}_\theta, D_P)$, uses exclusively public information, while the second term, $\boldsymbol{\lambda}^\top|\boldsymbol{\mu}(D_P) - \boldsymbol{\mu}(D_G)|$ requires both the public and sensitive group information. The computation of these gradients can be made differentially private by the introduction of carefully calibrated Gaussian noise. The general concept, relies on performing a *differentially private Stochastic Gradient Descent (DP-SGD)* step (Abadi et al. 2016). In a nutshell, DP-SDG computes the gradients for each data sample in a random mini-batch, clips their L2-norm, computes the average, and adds noise to ensure privacy.

The result below bounds the global sensitivity Δ_p of the *sensitive* term in the primal update, which is needed to calibrate the noise necessary to guarantee privacy.

Theorem 1. *Let $\|\nabla_\theta h(z)\| \leq C_p$, for all $z \in B_{G_i}$, $i \in \mathcal{I}$, and some $C_p > 0$. The global sensitivity Δ_p of the gradients of the constraints violation $\nabla_\theta \boldsymbol{\lambda}^\top|\boldsymbol{\mu}(B_P) - \boldsymbol{\mu}(B_G)|$ is*

$$\Delta_p \leq \frac{2C_p\lambda^{\max}}{\min_{i\in\mathcal{I}}|B_{G_i}| - 1}. \tag{7}$$

The above uses a clipping term, C_p, to control the maximal change of the gradients. Crucially, this is non-limiting, as it can be enforced by clipping the gradient contribution $\|\nabla_\theta h(z)\|$ to C_p, similarly to what is done in DP-SGD (Abadi et al. 2016). Using Theorem 1, the privacy-preserving primal update step for a mini-batch $B \subseteq D$ can be executed by clipping exclusively the gradients of the functions $h(z)$ associated with the group terms in B_G. *It is not necessary to perform gradient clipping for the functions $h(z)$ associated with the population terms in B_P*. While this may induce propagating population and group terms gradients of different magnitudes, the authors observed often improved performance in the adopted setting. Thus, PF-LD substitutes line 4 of Algorithm 1 with the following

$$\theta \leftarrow \theta - \alpha\Big(\nabla_\theta\left[J(\mathcal{M}_\theta, B_P)\right] + \boldsymbol{\lambda}^\top\left|\nabla_\theta\boldsymbol{\mu}(B_P) - \bar{\nabla}_\theta^{C_p}\boldsymbol{\mu}(B_G)\right| + \mathcal{N}(0, \sigma_p^2\,\Delta_p^2\boldsymbol{I})\Big), \tag{8}$$

with $\boldsymbol{I} \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{I}|}$, $\sigma_p > 0$, and $\bar{\nabla}_\theta^{C_p}$ is applied to each element $\mu(B_{G_i})$ of vector $\boldsymbol{\mu}(B_G)$, where $\bar{\nabla}_\theta^{C_p}(x) = \nabla x / \max(1, \frac{\|\nabla x\|}{C_p})$ denotes the gradients of a given scalar loss x clipped in a C_p-ball, for $C_p > 0$.

Private Dual Update Similar to the primal step, the dual update requires access to the sensitive group information (see line 5 of Algorithm 1). It updates the multipliers based on amount of constraint violation $|\boldsymbol{\mu}(D_P) - \boldsymbol{\mu}(D_G)|$ computed over the entire dataset D. Privacy can be attained by injecting Gaussian noise to the computation of the multipliers, but computing the global sensitivity Δ_d of the constraint violations is non-trivial since the range of the violations is unbounded. Once again, the paper recurs to the adoption of a clipping term, C_d, that controls the maximal contribution of the constraint violation to the associated multiplier value.

Theorem 2. *Let $|h(z)| \leq C_d$, for all samples $z \in D_{G_i}$, $i \in \mathcal{I}$, and some $C_d > 0$. The global sensitivity Δ_d of the constraint violation $|\boldsymbol{\mu}(D_P) - \boldsymbol{\mu}(D_G)|$ is*

$$\Delta_d \leq \frac{\sqrt{2}C_d}{\min_{i\in\mathcal{I}}|D_{G_i}| - 1}. \tag{9}$$

The privacy-preserving dual update step, used in lieu of line 5 of Algorithm 1, is given by the following

$$\boldsymbol{\lambda}_{k+1} \leftarrow \boldsymbol{\lambda}_k + s_k\left(|\boldsymbol{\mu}(D_P) - \bar{\boldsymbol{\mu}}^{C_d}(D_G)| + \mathcal{N}(0, \sigma_d^2\Delta_d^2\boldsymbol{I})\right) \tag{10}$$

with $\boldsymbol{I} \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{I}|}$, $\sigma_d > 0$, and where, for every $i \in \mathcal{I}$,

$$\bar{\mu}^{C_d}(D_{G_i}) = \hat{\mathbb{E}}_{z\sim D_{G_i}}\left[\frac{h(z)}{\max(1, \frac{|h(z)|}{C_d})}\right].$$

Note that while Theorem 1 bounds the individual gradient norms of each functions $h(z)$ for samples $z \in B_{G_i}$ and $i \in \mathcal{I}$, Theorem 2 bounds their maximum absolute values. The choice of terms C_p and C_d plays a special role in limiting the impact of an individual change in the protected attributes. It controls, indirectly, the privacy loss, as it impacts the global sensitivities Δ_p and Δ_d. However, these terms also affect the model accuracy and fairness. In particular, larger C_p values will propagate more precise gradients ensuring better accuracy and model fairness. On the other hand, larger clipping values will also introduce more noise, and thus degrade the information propagated. The converse is true for smaller clipping values. A theoretical and experimental analysis of the impact of these terms to the model accuracy and fairness is provided in the next sections.

Privacy Analysis

The privacy analysis of PF-LD relies on the moment accountant for Sampled Gaussian (SG) mechanism (Mironov, Talwar, and Zhang 2019), whose privacy is analyzed using the definition of *Rényi Differential Privacy* (RDP) (**?**). In synthesis, a function f satisfies (α, ϵ)-RDP, if the Rényi divergence $\mathcal{D}_\alpha$ of order α between any adjacent inputs of f is upper bounded by ϵ. Additionally, the SG mechanism with sampling ratio q (that is, uses a fraction q of the n data points) and standard deviation σ satisfies (α, ϵ)-RDP with:

$$\epsilon \leq \mathcal{D}_\alpha\left(\mathcal{N}(0, \sigma^2) \;\|\; (1-q)\mathcal{N}(0, \sigma^2) + q\mathcal{N}(1, \sigma^2)\right) \tag{11a}$$

$$\epsilon \leq \mathcal{D}_\alpha\left((1-q)\mathcal{N}(0, \sigma^2) + q\mathcal{N}(1, \sigma^2) \;\|\; \mathcal{N}(0, \sigma^2)\right). \tag{11b}$$

The following results consider a PF-LD algorithm that trains a model over T epochs using a dataset D containing n training samples, uses mini-batches B at each iteration, and standard deviation parameters σ_p and σ_d, associated to the primal and dual steps, respectively. Note that, Equations (8) and (10), correspond to instances of the SG mechanism.

Lemma 1. *The PF-LD primal step satisfies (α, ϵ_p)-RDP, where ϵ_p satisfies Equations* (11) *with $q = |B|/n$ and $\sigma = \sigma_p \Delta_p$ are, respectively, the SG sampling ratio q and standard deviation parameters.*

Lemma 2. *The PF-LD dual step satisfies (α, ϵ_d)-RDP, where ϵ_d satisfies Equations* (11) *with $q = 1$ and $\sigma = \sigma_d \Delta_d$ are, respectively, the SG sampling ratio and standard deviation parameters.*

PF-LD uses a predefined amount of noise (specified by parameters σ_p and σ_d) at each iteration, so that each iteration has roughly the same privacy loss, and uses the moment accountant (Abadi et al. 2016) to track detailed information of the cumulative privacy loss.

Theorem 3. *PF-LD satisfies $(\alpha, \frac{Tn\epsilon_p}{|B|} + T\epsilon_d)$-RDP.*

The results above follow directly from Equations (11) and RDP composability results (Mironov 2017), since the primal and dual steps of PF-LD uses Gaussian noise with parameter σ_p and σ_d, respectively. The final privacy loss in the (ϵ, δ)-differential privacy model is obtained by observing that a mechanism satisfying (α, ϵ)-RDP also satisfies $(\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-differential privacy, for any $0 < \delta < 1$ (Mironov 2017).

Bias-Variance Analysis

A key aspect of PF-LD is the choice of values C_d and C_p used to bound the functions $h(z)$, and their gradients, respectively, for every sample $z \in D_{G_i}$ and $i \in \mathcal{I}$. The choice of these clipping terms affects the global sensitivity of the functions of interest, which, in turn, impacts the amount of noise used by the differentially private mechanisms. As a result, the clipping terms are associated to a *bias-variance* trade-off: Small values can discard significant amounts of *constraints information*, thus introduce bias; large values retain more information but force the differential privacy mechanism to introduce larger noise, inducing more variance. It is important to recall that, at every iteration, PF-LD induces some privacy loss, thus, for a fixed privacy budget, the use of small values cannot be compensated by longer runs. This section formulates a bias-variance analysis that is helpful to select clipping values for gradient norms under the SG mechanism.

Let $\boldsymbol{G} = \nabla_\theta \boldsymbol{\lambda}^\top |\boldsymbol{\mu}(B_P) - \boldsymbol{\mu}(B_G)|$ be the gradient computed over a minibatch B during the primal update of F-LD (Algorithm 1 line 4) and $\tilde{\boldsymbol{G}} = \boldsymbol{\lambda}^\top |\nabla_\theta \boldsymbol{\mu}(B_P) - \bar{\nabla}_\theta^{C_p} \boldsymbol{\mu}(B_G)| + \mathcal{N}(0, \sigma_p^2 \Delta_p^2 \boldsymbol{I})$ be its privacy-preserving counterpart, as computed by PF-LD (Equation (8)).

Theorem 4. *The expected error between the real and noisy gradients, $\boldsymbol{G}$ and $\tilde{\boldsymbol{G}}$, incurred during the primal step can be upper bounded as:*

$$\mathbb{E}\left[\left\|\boldsymbol{G} - \tilde{\boldsymbol{G}}\right\|\right] \leq \frac{2\sqrt{S_{\tilde{\boldsymbol{G}}}}\sigma_p \lambda^{\max} C_p}{\min_{i \in \mathcal{I}} |B_{G_i}| - 1} + \sum_{i \in \mathcal{I}} \lambda_i \hat{\mathbb{E}}_{z \sim B_{G_i}} \left[\max\left(0, \|\nabla_\theta h(z)\| - C_p\right)\right], \tag{12}$$

where $S_{\tilde{\boldsymbol{G}}}$ is the shape (i.e., the number of entries) of $\tilde{\boldsymbol{G}}$.

The proof relies on isolating and bounding the variance and bias terms of the expected error).

Note that the bound above is a convex function of C_p. Its unique minimizer satisfies:

$$\frac{2\sqrt{S_{\tilde{\boldsymbol{G}}}}\sigma_p \lambda^{\max}}{\min_{i \in \mathcal{I}} |B_{G_i}| - 1} = \sum_{i \in \mathcal{I}} \lambda_i \hat{\mathbb{E}}_{z \sim B_{G_i}} \Big[\mathbb{1}\big[\|\nabla_\theta h(z)\| \geq C_p\big]\Big].$$

While beyond the scope of this work, the above illustrates that a procedure to find the optimal C_p privately can be constructed effectively.

Next, the paper shows how to bound the expected error incurred in using the noisy constraint violations during the dual step. Let $V_i = |\mu(D_{P_i}) - \mu(D_{G_i})|$ be the value corresponding to the i-th constraint violation ($i \in \mathcal{I}$), and $\tilde{V}_i = |\mu(D_{P_i}) - \bar{\mu}^{C_d}(D_{G_i})| + \mathcal{N}(0, \sigma_d^2 \Delta_d^2)$ be its privacy-preserving version (see Equation (10)).

Theorem 5. *The expected absolute error between the real and noisy constraint violations V_i and $\tilde{V}_i$, for $i \in \mathcal{I}$, is bounded by the following*

$$\mathbb{E}\left[|V_i - \tilde{V}_i|\right] \leq \frac{\sqrt{2} C_d \, \sigma_d}{\min_{i \in \mathcal{I}} |D_{G_i}| - 1} + \hat{\mathbb{E}}_{z \sim D_{G_i}} \left[\max(0, |h(z)| - C_d)\right]. \tag{13}$$

The proof uses similar arguments as those in the proof of Theorem 4 (see (Tran, Fioretto, and Van Hentenryck 2020) for details).

Experimental Analysis

Datasets, Models, and Metrics This section studies the behavior of the proposed algorithm on several datasets, including *Income*, *Bank*, and *Compas* (**?**) datasets. Since the trends are similar, the following discussion focuses on the Bank datasets, whose task is to detect client subscriptions to the term deposit, and the protected attributes define two age groups.

The experiments consider a baseline classifier (*CLF*), implemented as a neural network with two hidden layers, that maximize accuracy only, without considerations for fairness or privacy, and compare the proposed *PF-LD* model against the following state-of-the-art algorithms: `Z`, it implements a fair logistic regression models that achieves group fairness. These models were presented in (Zafar et al. 2017b) for demographic parity and in (**?**) for accuracy parity and equalized odds. `A`, it implements the fair logistic regression model based on reduction approaches, introduced in (**?**). Note that the models above preserves group fairness but *do not guarantee privacy*. They are used to highlight the effectiveness of the proposed approach based on the Lagrangian dual to ensure fairness. Finally, `M`, proposed in (Mozannar, Ohannessian, and Srebro 2020), the model most related to the proposed work, ensures both fairness and ϵ-differential privacy with respect to the sensitive attributes. The algorithm uses the fair model `A` on perturbed noisy data generated according to a randomized response mechanism (Kairouz, Oh, and Viswanath 2014). While these models where studied in the context of equalized odds, this work extends them to satisfy

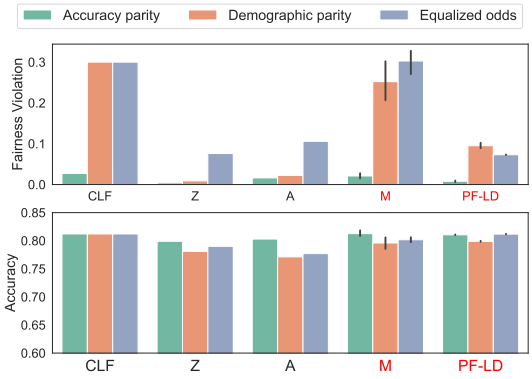

Figure 1: Accuracy and fairness comparison.

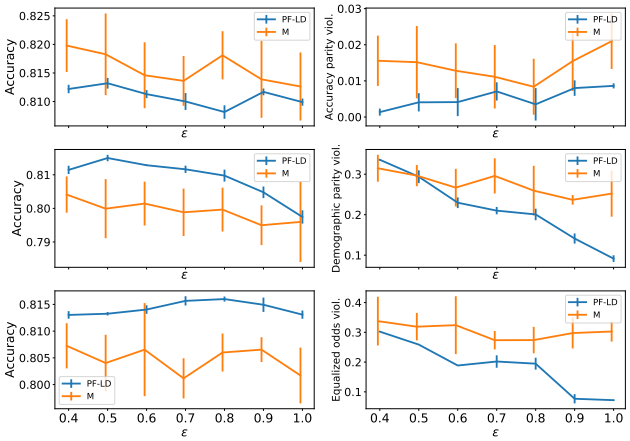

Figure 2: Privacy, fairness, and accuracy tradeoff.

all fairness definitions considered in this work. Compared to the proposed model M has the disadvantage of introducing large amounts of noise to the sensitive attribute, especially when the domain of these attributes is large and/or when A is high-dimensional. [2]

The experiments analyze the accuracy, fairness violations, and privacy losses (when applicable) of the models above. The fairness violations are measured as the maximal difference in fairness constraint violations between any two protected groups. The privacy losses are set to $\epsilon = 1.0$ and $\delta = 10^{-5}$, unless otherwise specified. PF-LD uses clipping bound values $C_p = 10.0$ and $C_d = 5.0$ and each experiment and configuration is repeated 10 times and presents average and standard deviation results.

Accuracy and Fairness This section analyzes the impact on accuracy and fairness of the privacy-preserving models introduced above. The results are summarized in Figure 1. Note that the plots have different scales. First, observe that the prediction accuracy of the proposed model is in line with that of the baseline *non-private, non-fair* classifier. This is true for all the group fairness constraints adopted. Next, observe that the fairness violations reported by PF-LD are competitive with those reported by the fair, *non-private* models A and Z. Finally, notice that PF-LD reports considerably lower fairness violations when compared to M, the only other private and fair model analyzed, and the results are consistent across all fairness metrics and benchmarks. This is remarkable as the privacy budget adopted is very modest when compared to what typically adopted in the privacy-preserving machine learning literature (Xie et al. 2018; Jagielski et al. 2019).

Privacy, Fairness, and Accuracy Tradeoff This section illustrates the tradeoff between privacy, fairness, and accuracy attained by PF-LD and compares them with algorithm M. The results are summarized in Figure 2, that depicts the average and standard deviation of 100 model runs. The statistical difference between the performance PF-LD and M was verified using a paired t-test, which reported a p-value < 0.001 in each setting.

Firstly, observe that the fairness violation score decreases as the privacy budget ϵ increases (note that the scale differs across the plots). Large privacy losses allow PF-LD to either run more iterations, given fixed noise values used at each iteration, or reduce the level of noise applied to a given number of iterations. These cases imply propagating more (former case) or more accurate (latter case) noisy constraint violations that results in better capturing the fairness constraints violations during the primal and dual update steps. This aspect is not obvious for M.

Next, notice that the model accuracy slightly decreases as ϵ increases. While this may seem surprising, our analysis shows that the fairness constraints, having their violations being propagated more exactly when ϵ increases, have a negative impact on the model accuracy.

Finally, notice that, in most cases, PF-LD is more accurate and produce models that have smaller fairness violations, and, importantly, it produces models that are more robust than those produced by M. This is noticeable by comparing the standard deviations on accuracy and fairness violations of the two models. These observations demonstrates the practical benefits of the proposed model.

PF-LD: Analysis of the Clipping Values This sections analyses the factors affecting the privacy, accuracy, and fairness tradeoff outlined above. The analysis focuses on the primal clipping bound C_p as trends for the dual bound C_d are similar. Figure 3 illustrates the effects of C_p on the model accuracy and fairness, at varying of the privacy parameter ϵ. Observe that, for different fairness definitions, the best accuracy/fairness tradeoff is obtained when $C_p \in [10, 20]$ (green and yellow curves). The use of small clipping values (blue curve) slows the drop in fairness violations at the increasing of the privacy budget ϵ. This is because small C_p values limit the impact of the constraints violations to the model. On the other extreme, for high C_p values (e.g., brown curve), not

[2] The authors note there is an additional work which addresses learning a fair and private classifier (Jagielski et al. 2019). While an important contribution, it has been shown to induce significant privacy losses (see Figure 1 of (Jagielski et al. 2019)). Model M was shown to outperform these algorithms presented in (Jagielski et al. 2019) in terms of classification error bounds. Therefore, this paper adopts M, as the state-of-the-art.

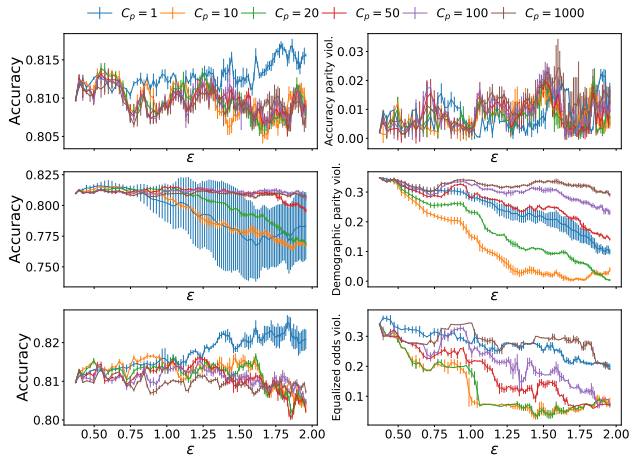

Figure 3: Effects of C_p on fairness and accuracy.

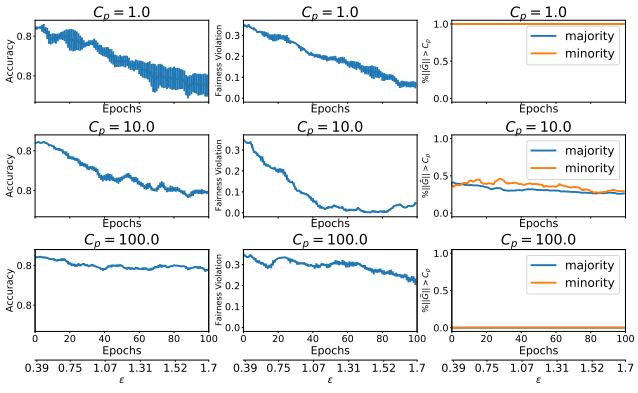

Figure 4: Individual gradient norms associated to a protected group and their relation to the clipping values C_p.

Hidden layers	$\lvert B \rvert$	*N*	*F-LD*	*P*	*PF-LD*
2	16	0.53 (0.01)	0.60 (0.01)	0.9 (0.04)	1.0 (0.02)
2	64	0.19 (0.01)	0.25 (0.01)	3.2 (0.09)	3.3 (0.08)
2	256	0.11 (0.01)	0.16 (0.01)	5.7 (0.05)	5.7 (0.10)
10	16	1.20 (0.02)	1.30 (0.01)	2.0 (0.08)	2.1 (0.07)
10	64	0.41 (0.02)	0.47 (0.01)	6.9 (0.09)	7.0 (0.12)
10	256	0.17 (0.01)	0.20 (0.01)	11.8 (0.15)	11.8 (0.23)

Table 1: Training time (sec) comparison.

only it is observed a degradation in fairness violations, but also in the model accuracy. This is because large C_p values imply larger amount of noise to be added to the gradient of the constraint violations, resulting in less accurate information to be back-propagated at each step. Additionally, the noisy constraints gradients can negatively impact the classifier loss function. Thus, the resulting models tend to have worse accuracy/fairness tradeoff than those trained with intermediate C_p values. These observations support the theoretical analysis which showed that the expected error between the true and private gradients of the fairness constraints is upper bounded by a convex function of the primal and the dual clipping bounds.

To further shed lights on the impacts of C_p to the model fairness and accuracy, Figure 4 illustrates the model accuracy (left column) the fairness violations (middle column) and the percentage of times the norm of the gradients associated to the constraint violations of a protected group exceeds the clipping value C_p: $\|\tilde{G}\| > C_p\%$ (right column). The last column indicates the frequency of propagating the correct or the clipped information. The figure uses demographic parity, but the results are consistent across the other fairness metrics studied. Observe that, the percentage of individual constraint gradients exceeding C_p is very high when C_p is small. Thus, a significant amount of information is lost due to clipping. Conversely, at large C_p regimes most individual gradients (for both protected groups) are smaller than C_p. This choice reduces bias, but it introduces large variances due to noise necessary to preserve privacy. Therefore, both cases result in models that have large fairness violations. Conversely, at intermediate C_p regimes, the produced models have lower constraint violations while retaining high accuracy.

Performance Finally, this section analyses the performance of the proposed Lagrangian dual framework. Table 1 reports the average training time and standard deviations (in parenthesis) required to execute one epoch at the varying of the number of hidden layers and mini-batch size $|B|$ for a model enforcing no privacy nor fairness (*N*), one enforcing privacy only (*P*), one enforcing fairness only (*F-LD*), and the proposed one enforcing both privacy and fairness (*PF-LD*). The tests use a common laptop (MacBook Air 2013, 1.7GHz, 8GB RAM) on the Bank dataset and are consistent for all the fairness notions adopted.

Note that imposing the fairness constraints comes at almost no-overhead on top of the non-fair counterpart models. It has been observed that clipping (used to preserve privacy) is computationally expensive. However, this drawback has been recently mitigated by the work of Subramani, Vadivelu, and Kamath (2020), which uses JAX to speed up these operations and can achieve up to an order magnitude speedups.

Conclusions

This paper was motivated by the discrepancy between concerns in building models whose outcomes do not discriminate against some demographic groups and the requirements that the sensitive attributes, which are essential to build these models, may not be available due to legal and ethical requirements. It proposed a framework to train deep learning models that satisfy several notions of group fairness, including equalized odds, accuracy parity, and demographic parity, while ensuring that the model satisfies differential privacy for the protected attributes. The framework relies on the use of Lagrangian duality to accommodate the fairness constraints and the paper showed how to inject carefully calibrated noise to the primal and dual steps of the Lagrangian dual process to guarantee privacy of the sensitive attributes. The paper further analyses the tension between accuracy, privacy, and fairness and an extensive experimental evaluation illustrates the benefits of the proposed framework showing that it may be come a practical tool for privacy-preserving and fair decision making.

Ethic Statement

This paper studies a framework to train deep neural networks resulting in models that protect the privacy of the individuals' sensitive information while also allowing it learn non-discriminatory predictors. The outcome of this research has the potential to operate ethically, privately, and effectively in key decision processes arising in many classification problems, as illustrated in the experimental section. The adoption of the proposed methodology would benefit both the individual data owners, whose privacy protection is guaranteed, and the data curators that can use privacy-preserving and fair solutions which are also operationally effective. The main challenge or limitation of the proposed framework lies in the inherent tradeoff between the privacy of its users, the level of fairness retained, and the model accuracy.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J.; et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3(1): 1–122.

Chaudhuri, K.; Monteleoni, C.; and Sarwate, A. D. 2011. Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research* .

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

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

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

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

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

Fioretto, F.; Mak, T.; and Van Hentenryck, P. 2020. Predicting AC Optimal Power Flows: Combining Deep Learning and Lagrangian Dual Methods. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(01): 630–637.

Fioretto, F.; Van Hentenryck, P.; Mak, T.; Tran, C.; Baldo, F.; and Lombardi, M. 2020. Lagrangian Duality for Constrained Deep Learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*.

Fontaine, D.; Laurent, M.; and Van Hentenryck, P. 2014. Constraint-Based Lagrangian Relaxation. In *Principles and Practice of Constraint Programming*, 324–339.

Hardt, M.; Price, E.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Hestenes, M. R. 1969. Multiplier and gradient methods. *Journal of optimization theory and applications* 4(5): 303–320.

Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Malvajerdi, S. S.; and Ullman, J. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Kairouz, P.; Oh, S.; and Viswanath, P. 2014. Extremal Mechanisms for Local Differential Privacy. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Mironov, I. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275. IEEE.

Mironov, I.; Talwar, K.; and Zhang, L. 2019. R\'enyi Differential Privacy of the Sampled Gaussian Mechanism. *ArXiv* abs/1908.10530.

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

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

Subramani, P.; Vadivelu, N.; and Kamath, G. 2020. Enabling Fast Differentially Private SGD via Just-in-Time Compilation and Vectorization. *ArXiv* abs/2010.09063.

Tran, C.; Fioretto, F.; and Van Hentenryck, P. 2020. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *ArXiv* abs/2009.12562.

Xie, L.; Lin, K.; Wang, S.; Wang, F.; and Zhou, J. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* .

Xu, D.; Yuan, S.; and Wu, X. 2019. Achieving Differential Privacy and Fairness in Logistic Regression. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017b. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of Artificial Intelligence and Statistics (AISTAT)*.