A Stochastic Optimization Framework for Private and Fair Learning From Decentralized Data

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models are often trained on sensitive data (e.g., medical records and race/gender) that is distributed across different "silos" (e.g., hospitals). These *federated learning* models may then be used to make consequential decisions, such as allocating healthcare resources. Two key challenges emerge in this setting: (i) maintaining the *privacy* of each person's data, even if other silos or an adversary with access to the central server tries to infer this data; (ii) ensuring that decisions are *fair* to different demographic groups (e.g., race/gender). In this paper, we develop a novel algorithm for private and fair federated learning (FL). Our algorithm satisfies *inter-silo record-level differential privacy* (ISRL-DP), a strong notion of private FL requiring that each silo's communicated messages satisfy record-level differential privacy. In addition to being differentially private, our framework can be used to promote different fairness notions, including demographic parity and equalized odds. We prove that our algorithm converges under mild smoothness assumptions on the loss function (even in nonconvex settings), whereas prior work required strong convexity for convergence. As a byproduct of our analysis, we obtain the first convergent algorithm for ISRL-DP optimization of nonconvex-strongly concave min-max loss functions in federated learning. This convergent DP optimization algorithm is a valuable contribution in its own right. Additionally, our experiments demonstrate the state-of-the-art fairness-accuracy tradeoffs of our algorithm across different privacy levels. Compared to existing state of the art, we obtained an average of around 64% reduction in demographic parity fairness violation and 95% lower for equalized odds.

1 Introduction

Many important decisions are being assisted by machine learning (ML) models (e.g., loan approval or criminal sentencing). Without intervention, ML modes may discriminate against certain demographic groups (e.g., race, gender). For instance, Amazon developed a ML-based recruiting software that showed a strong bias against hiring women for technical jobs (Dastin, 2018). *Algorithmic fairness* research aims to develop algorithms that promote equitable treatment of different demographic groups by correcting biases that may lead to unfair outcomes.

Despite the development of numerous fair learning algorithms, two key challenges impede their real-world application: (1) Training fair models requires access to *sensitive data* (e.g., age, race, gender) in order to ensure fairness of predictions with respect to these attributes. However, data protection and privacy regulations (like E.U.'s General Data Protection Regulation and California's Consumer Privacy Act) restrict the usage of sensitive demographic consumer data (GDP, 2016; BUKATY, 2019). (2) Training data is often *distributed* across different organizations, such as hospitals or banks, who may not share their data with third parties.

To address obstacle (1), prior works (Jagielski et al., 2019; Mozannar et al., 2020; Tran et al., 2021; 2022; Lowy et al., 2023) have used *differential privacy* (Dwork et al., 2006) to preserve the privacy of the sensitive data during fair model training. Informally, differential privacy (DP) ensures that no adversary can infer much more about any individual piece of sensitive data than they could have inferred had that piece of data

never been used. While these works address the first challenge, they fail to address the second challenge, since they require centralized access to the full data.

In this work, we address the two aforementioned challenges via fair private *federated learning* (McMahan et al., 2017) under an appropriate notion of differential privacy. Federated learning (FL) is a distributed learning framework in which silos collaborate to train a global model by exchanging focused updates, often with the orchestration of a central server. By permitting silos to collaborate without sharing their sensitive local data, FL offers an ideal solution to challenge (2).

Although FL offers some privacy benefits to silos via local storage of data, this is not sufficient to prevent sensitive data from being leaked: model parameters or updates can leak data, e.g. via gradient or model inversion attacks (Li et al., 2024b;a). To prevent sensitive data from being leaked during FL, we will require the full transcript of silo i's sent messages (e.g., local gradient updates) to be differentially private. This privacy requirement is known as *inter-silo record-level differential privacy* (ISRL-DP) (Lowy & Razaviyayn, 2023; Liu et al., 2022), defined formally in Section 2. For example, if the silos are hospitals, then ISRL-DP preserves the privacy of each patient's record, even if an adversary with server access colludes with the other hospitals to try to decode the data of hospital i.

Prior work. There are several existing work on centralized private and fair learning (Jagielski et al., 2019; Lowy et al., 2023; Tran et al., 2021), on private federated learning (Lowy et al., 2022b; Lowy & Razaviyayn, 2021; Girgis et al., 2021; Gao et al., 2024), and on fair FL (Ezzeldin et al., 2023). However, *the literature on private and fair federated learning is sparse.* In fact, the only related works we are aware of are due to Rodríguez-Gálvez et al. (2021); Ling et al. (2024); Padala et al. (2021); Gu et al. (2022). The work of Rodríguez-Gálvez et al. (2021) does not prove any ISRL-DP guarantee for their algorithm, nor do they provide a convergence guarantee. The work of Ling et al. (2024) only guarantees convergence for *strongly convex* loss functions, limiting its applicability in a wide range of modern ML models. For example, linear/logistic regression and deep learning loss functions are not strongly convex, limiting the applicability of their developments. Moreover, the algorithm of Ling et al. (2024) promotes a particular form of fairness notion known as *balanced performance fairness,* which is less popular than other notions such as *demographic parity* or *equalized odds*. The work of Padala et al. (2021) goes through a two-stage training. They first train a fair model using FairSGD, in a non-private manner, and then train a DP model which emulates the output of the FairSGD model. But that final trained model is not necessarily fair. On the other hand, the work of Gu et al. (2022) cannot work with mini-batches of data due to their method of enforcing fairness, limiting its use in applications with large training datasets.

Contributions. Motivated by the shortcomings of prior works, our work addresses the following question:

> Can we develop an algorithm for fair and private federated learning that provably converges, even with loss functions that are not necessarily strongly convex?

To answer this question, we develop a novel framework for promoting fairness and ISRL-DP with respect to sensitive attributes in a federated learning setting. Our framework is flexible, covering different fairness notions such as demographic parity and equalized odds. Further, our algorithm provides:

1. **Guaranteed ISRL-DP and convergence:** We prove that our ISRL-DP algorithm converges for any smooth (*potentially non-convex*) loss function, even when mini-batches of data are used (i.e. stochastic optimization). Thus, our algorithm can be used in large-scale FL settings, where full batch training is not feasible.

2. **State-of-the-art empirical performance**: our ISRL-DP alogrithm achieves significantly improved fairness-accuracy tradeoffs on benchmark tasks across different privacy levels. For example, the equalized odds *fairness violation of our algorithm is* 95% *lower than the previous state-of-the-art* (Ling et al., 2024) for the same fixed accuracy level. Additionally, our algorithm even outperforms strong centralized DP fair baselines that do not provide the strong protection of ISRL-DP Tran et al. (2021).

Furthermore, our analysis yields a significant theoretical byproduct: the first convergent FL algorithm for ISRL-DP optimization in nonconvex-strongly concave min-max optimization. This convergent DP optimization algorithm is a valuable contribution in its own right. Moreover, our framework extends to *hybrid centralization* settings, where some data features are centralized and others are decentralized (e.g., centralization of sensitive and decentralization of non-sensitive data). We discuss this in a detailed manner in section 4.

2 Problem Setting and Preliminaries

Consider a federated learning setting with N silos (e.g., hospitals or banks), each of which has data that is partitioned into sensitive and non-sensitive data divisions: $\{Z_j = (X_j, Y_j), S_j\}_{j=1}^N$, where $(X_j, Y_j) = \{x_{j,i}, y_{j,i}\}_{i=1}^{\tilde{n}}$, $S_j = \{s_{j,i}\}_{i=1}^{\tilde{n}}$, and $\tilde{n} = n/N \in \mathbb{N}$ is the number of local samples per silo. $x_{j,i} \in \mathcal{X}$ are the non-sensitive features, $s_{j,i} \in [k] \triangleq \{1, \dots, k\}$ are the discrete sensitive attributes (e.g. race, gender), and $y_{j,i} \in [l] \triangleq \{1, \dots, l\}$ are the ground-truth labels.[1] Let $\widehat{y}_\theta(x)$ denote the model predictions parameterized by θ, and $\ell(\theta, x, y) = \ell(\widehat{y}_\theta(x), y)$ be a loss function (e.g. cross-entropy loss). Our goal is to (approximately) solve the empirical risk minimization (ERM) problem

$$\min_\theta \left\{ \widehat{\mathcal{L}}(\theta) := \frac{1}{N\tilde{n}} \sum_{j=1}^{N} \sum_{i=1}^{\tilde{n}} \ell(\theta, x_{ji}, y_{ji}) \right\} \tag{1}$$

in a fair manner, while maintaining the differential privacy of the sensitive data $\{S_j\}_{j=1}^n$ under ISRL-DP. We consider two different notions of fairness in this work:[2]

Definition 1 (Fairness Notions)**.** Let $\mathcal{A} : \mathcal{Z} \to \mathcal{Y}$ be a classifier.

- $\mathcal{A}$ satisfies *demographic parity* (Dwork et al. 2012) if the predictions $\mathcal{A}(Z)$ are statistically independent of the sensitive attributes.
- $\mathcal{A}$ satisfies *equalized odds* (Hardt et al. 2016a) if the predictions $\mathcal{A}(Z)$ are conditionally independent of the sensitive attributes given $Y = y$ for any $y \in \mathcal{Y}$.

The choice of fairness notion depends on the application at hand (See Chouldechova & Roth (2020) for discussion.)

It has been demonstrated that achieving perfect fairness is *impossible* for a differentially private algorithm that also achieves non-trivial accuracy (Cummings et al. 2019). Therefore, we focus on developing an algorithm that minimizes a certain measure of *fairness violation* on the given dataset Z. Fairness violations can be quantified in various ways; see, Dwork et al. (2012); Hardt et al. (2016a); Lowy et al. (2022a) for an overview. As an example, if demographic parity is the desired fairness criterion, we can quantify the (empirical) demographic parity violation using the following measure:

$$\max_{\widehat{y} \in \mathcal{Y}} \max_{s \in \mathcal{S}} \left| \hat{p}_{\widehat{Y}|S}(\widehat{y}|s) - \hat{p}_{\widehat{Y}}(\widehat{y}) \right|, \tag{2}$$

where $\hat{p}$ represents the empirical probability computed from the data $(Z, \{\widehat{y}_i\}_{i=1}^n)$. Note that the demographic parity violation in equation 2 is zero if and only if demographic parity is satisfied.

Next, we define differential privacy. Following the DP fair learning literature (Jagielski et al. 2019) and motivated by the discussion in the Introduction, we consider a relaxation of DP, in which only the *sensitive attributes* require privacy.[3] In the centralized setting, we say Z and Z' are *adjacent with respect to sensitive data* if $Z = \{(x_i, y_i, s_i)\}_{i=1}^n$, $Z' = \{(x_i, y_i, s_i')\}_{i=1}^n$, and there is a unique $i \in [n]$ such that $s_i \neq s_i'$.

[1] Our algorithm and analysis readily extends to the case in which silo data sets contain different numbers of samples, via standard techniques (see e.g., Lowy & Razaviyayn (2023))

[2] Our method can also handle any other fairness notion that can be defined in terms of statistical (conditional) independence, such as equal opportunity. However, our method cannot handle all fairness notions: for example, false discovery rate and calibration error are not covered by our framework.

[3] However, the convergence guarantee of our algorithm easily extends to the case where privacy of the entire data set is needed.

Definition 2 (Differential Privacy w.r.t. Sensitive Attributes)**.** Let $\varepsilon \geqslant 0,\ \delta \in [0, 1)$. A randomized algorithm $\mathcal{A}$ is (ε, δ)-*differentially private (DP) w.r.t. sensitive attributes* S if for all pairs of data sets Z, Z' that are *adjacent w.r.t. sensitive attributes*, we have

$$\mathbb{P}(\mathcal{A}(Z) \in O) \leqslant e^{\varepsilon}\mathbb{P}(\mathcal{A}(Z') \in O) + \delta, \tag{3}$$

for all measurable sets $O \subseteq \mathcal{Y}$.

In the context of FL with N silos, we say two distributed datasets $Z = (Z_1, \ldots, Z_N)$ and $Z' = (Z'_1, \ldots, Z'_N)$ with $Z_j = \{(x_{j,i}, y_{j,i}, s_{j,i})\}_{i=1}^{\tilde{n}}$ $Z'_j = \{(x_{j,i}, y_{j,i}, s'_{j,i})\}_{i=1}^{\tilde{n}}$ are adjacent if for every $j \in [N]$, there is at most one $i \in [\tilde{n}]$ such that $s_{j,i} \neq s'_{j,i}$. Thus, adjacent distributed datasets Z and Z' may differ in up to N samples, one from each silo.

Definition 3 (Inter-Silo Record-Level DP)**.** A federated learning algorithm $\mathcal{A}$ is (ε, δ)-inter-silo-record-level DP (ISRL-DP) if, for each $j \in [N]$, the full transcript of silo j's sent messages satisfies (3) for all adjacent distributed datasets Z, Z' and any fixed settings of other silos' data.

By post-processing property of DP (Dwork & Roth, 2014), Definition 3 ensures that the model parameters and the messages broadcast by the central server and are also DP.

As discussed in Section 1, Definition 2 is useful if a company wants to train a fair model, but is unable to use the sensitive attributes collected in another silo (and is needed to train a fair model) due to privacy concerns and laws. Following Lowy et al. (2023), we shall impose the reasonably practical assumption that all data sets contain at least ρ-fraction of every sensitive attribute for some $\rho \in (0, 1)$.

3 Private and Fair Federated ERM Framework

A popular method in the literature to enforce fairness is to introduce a regularizer that penalizes the model for making unfair decisions (Zhang et al., 2018; Donini et al., 2018; Baharlouei et al., 2020). Let S, Y, and $\hat{Y}$ be the random variables corresponding to sensitive attributes, actual output, and predictions by the models. The regularization approach to fair ERM jointly optimizes for accuracy and fairness by solving

$$\min_{\theta} \left\{ \widehat{\mathcal{L}}(\theta) + \lambda \mathcal{D}(\hat{Y}, S, Y) \right\},$$

where $\mathcal{D}$ is a measure of (conditional) statistical dependence (based on the fairness notion used) between the sensitive attributes S and the predicted outputs $\hat{Y}$. The dependency of $\mathcal{D}$ on S, $\hat{Y}$, and/or Y varies for different fairness notions. For instance, for demographic parity, $\mathcal{D}$ just depends on S and $\hat{Y}$, while equalized odds $\mathcal{D}$ also depends on Y. The parameter $\lambda \geqslant 0$ controls the trade-off between accuracy and fairness. Inspired by the strong performance of Lowy et al. (2022a; 2023), we use variations of the χ^2 divergence as our $\mathcal{D}$.

Definition 4 (χ^2 Divergence)**.** The χ^2 Divergence between two probability mass functions $P(x)$ and $Q(x)$ over the support of X is defined as

$$\chi^2(P||Q) = \sum_{x \in X} Q(x) \left(\frac{P(x)}{Q(x)} - 1 \right)^2$$

The choice of this divergence was motivated by the theoretical results and strong empirical performance of this divergence on stochastic fair optimization highlighted by the work of Lowy et al. (2022a). They provided extensive analysis on using this divergence using the following arguments:

- (Lowy et al., 2022a) proposed an unbiased estimator for the population χ^2 divergence term, enabling the regularizer's use on mini-batches and ensuring convergence guarantees for stationarity.
- They demonstrated that the χ^2 divergence between model outputs and sensitive labels serves as an upper bound for fairness violations, such as demographic parity and equalized odds. Hence, minimizing this divergence ensures the tightening of fairness violations.

For demographic parity, we would ideally like to use $D_R(\widehat{Y}, S) \triangleq \chi^2(p_{\widehat{Y},S}||p_{\widehat{Y}}p_S)$ as our regularizer, where the *true joint distribution* for the random variables $\widehat{Y}$ and S is given by $p_{\widehat{Y},S}$ and marginals are given by $p_{\widehat{Y}}, p_S$, respectively. However, since the true distribution of $(\widehat{Y}, S)$ is unknown in practice, we resort to an empirical estimate of the regularizer: $\widehat{D}_R(\widehat{Y}, S) \triangleq \chi^2(\hat{p}_{\widehat{Y},S}||\hat{p}_{\widehat{Y}}\hat{p}_S)$, where the empirical joint distribution for the random variables $\widehat{Y}$ and S is given by $\hat{p}_{\widehat{Y},S}$ and marginals by $\hat{p}_{\widehat{Y}}$, $\hat{p}_S$ respectively. Similarly, for equalized odds, $D'_R(\widehat{Y}, S) \triangleq \chi^2(p_{\widehat{Y},S|Y}||p_{\widehat{Y}|Y}p_{S|Y})$, and we use $\widehat{D}'_R(\widehat{Y}, S) \triangleq \chi^2(\hat{p}_{\widehat{Y},S|Y}||\hat{p}_{\widehat{Y}|Y}\hat{p}_{S|Y})$ in practice. We write the full expressions of these regularizers in Appendix A.

For concreteness, we consider demographic parity in what follows, but note that our developments extend easily to equalized oddds. Our approach to enforcing fairness is to augment (1) with the χ^2 regularizer and privately solve:

$$\min_{\theta} \left\{ \text{FERMI}(\theta) := \widehat{\mathcal{L}}(\theta) + \lambda \widehat{D}_R(\widehat{Y}_\theta(X), S) \right\}. \tag{FERMI obj.}$$

The empirical divergence $\widehat{D}_R$ is an asymptotically unbiased estimator of population divergence D_R (Lowy et al., 2022a), suggesting that solving (FERMI obj.) should generalize well to the corresponding population risk minimization problem.

The next question we address is: *how do we solve equation FERMI obj. in a distributed fashion, while satisfying ISRL-DP*? It is not obvious how to obtain statistically unbiased estimators of the gradients of $\widehat{D}_R(\widehat{Y}_\theta(X), S)$ without directly computing $\nabla_\theta \widehat{D}_R(\widehat{Y}_\theta(X), S)$ over the entire data set. But computing the gradient over the entire data set is not possible in the federated learning setting, since each silo stores its data locally in a decentralized manner.

Fortunately, Lowy et al. (2022a) gives us a statistically unbiased estimator through a min-max problem formulation. For feature input x, let the predicted class labels be given by $\widehat{y}(x, \theta) = j \in [l]$ with probability $\mathcal{F}_j(x, \theta)$, where $\mathcal{F}(x, \theta) \in [0, 1]^l$ is differentiable in θ, and $\sum_{j=1}^{l} \mathcal{F}_j(x, \theta) = 1$. For instance, $\mathcal{F}(x, \theta) = (\mathcal{F}_1(x, \theta), \ldots, \mathcal{F}_l(x, \theta))$ could represent the output of a neural net after softmax layer or the probability label assigned by a logistic regression model. Then we have the following min-max re-formulation of (FERMI obj.):

Theorem 5 (Lowy et al. (2022a)). *There are differentiable functions $\widehat{\psi}_{ji}$ such that (FERMI obj.) is equivalent to*

$$\min_{\theta} \max_{W \in \mathbb{R}^{k \times l}} \left\{ \widehat{F}(\theta, W) := \widehat{\mathcal{L}}(\theta) + \lambda \frac{1}{N\tilde{n}} \sum_{j=1}^{N} \sum_{i=1}^{\tilde{n}} \widehat{\psi}_{ji}(\theta, W) \right\}. \tag{4}$$

Further, $\widehat{\psi}_{ji}(\theta, W)$ is strongly concave in W for any θ.

The functions $\widehat{\psi}_{ji}$ are given explicitly in Appendix D. With Theorem 5, we can now claim that: for any batch on a particular silo $\mathcal{B}_j$ with size $m \in [\tilde{n}]$, the gradients (with respect to θ and W) of $\frac{1}{Nm} \sum_{j=1}^{N} \sum_{i \in \mathcal{B}_j} \ell(x_{ji}, y_{ji}; \theta) + \lambda \widehat{\psi}_{ji}(\theta, W)$ are statistically unbiased estimators of the gradients of $\widehat{F}(\theta, W)$, if $\mathcal{B}$ is drawn uniformly from Z. However, when differential privacy of the sensitive attributes is also desired, the formulation (4) presents some challenges, due to the non-convexity of $\widehat{F}(\cdot, W)$. Lowy et al. (2023) solve this problem in the centralized setting, but the proposed method may leak central data to the server and does not satisfy ISRL-DP.

Next, we develop our distributed ISRL-DP fair learning algorithm.

3.1 ISRL-DP Fair Federated Learning via SteFFLe

Our algorithm for privately solving the min-max FL problem equation 4 is given in Algorithm 1. Algorithm 1 is essentially a noisy distributed variation of *stochastic gradient descent ascent* (SGDA). Gaussian noise is added to each silo's sensitive stochastic gradients $\nabla_\theta \widehat{\psi}, \nabla_w \widehat{\psi}$ to ensure ISRL-DP with respect to the sensitive attributes. Then, the server aggregates these noisy sensitive gradients and the noiseless non-sensitive gradients $\nabla_\theta \ell(x, y, \theta)$ and updates the model parameters θ_{t+1} and W_{t+1} by taking descent and ascent steps.

Algorithm 1 SteFFLe: Stochastic Private Fair Federated Learning

Input: $\{Z_j = \{x_{j,i}, y_{j,i}\}_{i=1}^{\tilde{n}}, \{s_{j,i}\}_{i=1}^{\tilde{n}}\}_{j=1}^{N}$, $\theta_0 \in \mathbb{R}^{d_\theta}$, $W_0 = 0 \in \mathbb{R}^{k\times l}$, step-sizes (η_θ, η_w), fairness parameter $\lambda \geqslant 0$, iteration number T, minibatch size $|B_t| = m \in [\tilde{n}]$, set $\mathcal{W} \subset \mathbb{R}^{k\times l}$, noise parameters $\{\sigma_{j,w}^2, \sigma_{j,\theta}^2\}_{j=1}^{N}$.
Compute $\widehat{P}_S^{-1/2} = \mathrm{diag}(\widehat{p}_S(1)^{-1/2}, \ldots, \widehat{p}_S(k)^{-1/2})$, where $\widehat{p}_S(r) := \frac{1}{N\tilde{n}} \sum_{j=1}^{N} \sum_{i=1}^{\tilde{n}} \mathbb{1}_{\{s_{ji}=r\}} \geqslant \rho > 0$.
for $t = 0, 1, \ldots, T-1$ **do**
  Central server sends θ_t, W_t to all silos.
  for $j \in [N]$ **in parallel do**
    Silo j draws a mini-batch B_t of data points $\{(x_{j,i}, y_{j,i}), s_{j,i}\}_{i\in B_t}$.
    Silo j's non-sensitive division computes stochastic gradient $g_{t,j} := \frac{1}{|B_t|} \sum_{i\in B_t} \nabla_\theta \ell(x_{j,i}, y_{j,i}, \theta_t)$ and sends $\{\mathcal{F}(x_{j,i}, \theta_t), \nabla\mathcal{F}(x_{j,i}, \theta_t), j, i\}_{i\in B_t}$ to sensitive data division.
    Silo j's sensitive division computes noisy sensitive stochastic gradients $h_{t,j,\theta} := \frac{1}{|B_t|} \sum_{i\in B_t} \nabla_\theta \widehat{\psi}_{j,i}(\theta_t, W_t) + u_{t,j}$ and $h_{t,j,w} := \frac{1}{|B_t|} \sum_{i\in B_t} \nabla_w \widehat{\psi}_{j,i}(\theta_t, W_t) + V_{t,j}$, where $u_{t,j} \sim \mathcal{N}(0, \sigma_{j,\theta}^2 \mathbf{I}_{d_\theta})$ and $V_{t,j}$ is a $k \times l$ matrix with independent random Gaussian entries $(V_t)_{q,r} \sim \mathcal{N}(0, \sigma_{j,w}^2)$.
    Silo j broadcasts $g_{t,j}$, $h_{t,j,\theta}$, and $h_{t,j,w}$ to the central server.
  end for
  Central server updates $\theta_{t+1} \leftarrow \theta_t - \frac{\eta_\theta}{N} \sum_{j=1}^{N} [g_{t,j} + \lambda h_{t,j,\theta}]$ and $W_{t+1} \leftarrow \Pi_{\mathcal{W}}\Big(W_t + \frac{\lambda\eta_w}{N} \sum_{j=1}^{N} h_{t,j,w}\Big)$.
end for
Pick $\hat{t}$ uniformly at random from $\{1, \ldots, T\}$.
Return: $\hat{\theta}_T := \theta_{\hat{t}}$.

Theorem 6. *Let $\varepsilon \leqslant 2\ln(1/\delta)$, $\delta \in (0,1)$, and $T \geqslant \left(\tilde{n}\frac{\sqrt{\varepsilon}}{2|B_t|}\right)^2$. Assume $\mathcal{F}(x,\cdot)$ is L_θ-Lipschitz for all x, and $|(W_t)_{r,q}| \leqslant D$ for all $t \in [T], r \in [k], q \in [l]$. Then, for $\sigma_{j,w}^2 \geqslant \frac{16T\ln(1/\delta)}{\varepsilon^2\tilde{n}^2\rho}$ and $\sigma_{j,\theta}^2 \geqslant \frac{16L_\theta^2 D^2 \ln(1/\delta)T}{\varepsilon^2\tilde{n}^2\rho}$, Algorithm 1 is (ε,δ)-ISRL-DP with respect to the sensitive attributes for all data sets containing at least ρ-fraction of minority attributes.*

See Appendix B for the proof. Next, we provide a convergence guarantee for Algorithm 1.

Theorem 7. *Assume that the loss function $\ell(\cdot, x, y)$ is Lipschitz and $\ell(\cdot, x, y)$ and $\mathcal{F}(x,\cdot)$ have Lipschitz gradients. Then, there exist algorithmic parameters such that Algorithm 1 returns a $\hat{\theta}_T$ which is (ϵ,δ)-ISRL-DP with*

$$\mathbb{E}\|\nabla FERMI(\hat{\theta}_T)\|^2 = \mathcal{O}\left(\frac{\sqrt{\max(d_\theta, kl)\ln(1/\delta)}}{\varepsilon\tilde{n}\sqrt{N}}\right).$$

Note that we choose T to achieve the best accuracy that our algorithm can achieve. We will clarify the choice of T that implies the accuracy result in the above result. Compared to the central DP stationarity gap bound obtained in Lowy et al. (2023) (with $n = \tilde{n}N$), the bound in 7 is larger by a factor of $\sqrt{N}$. This is because ISRL-DP is a stronger privacy notion than central DP (Lowy & Razaviyayn, 2023) and our analysis accounts for *data heterogeneity* across silos.

The proof of Theorem 10 follows from careful tracking of noise variance and sampling of data obtained from the different silos. A key observation is that even though the sampling is distributed across silos, the expected value of gradient after this modified form of sampling is an unbiased estimator of the global loss function due to linearity of expectation. Moreover, by averaging silos' noisy gradients, we reduce the total privacy noise variance.

While our approach leverages the DP min-max optimization techniques of Lowy et al. (2023), extending this framework to FL has its own challenges. In particular, the sampling is distributed across silos with different data, which introduces additional challenges in analysis of the central updates. However, the expected value of gradient after this modified form of sampling is an unbiased estimator of the global loss function due to linearity of expectation helps us to overcome this challenge and derive bounds on sampling as to that of Lowy

et al. (2023). See Appendix C for the detailed proof. In fact, in Appendix C, we prove Theorem 10, which is a general result that applies to *all smooth non-convex strongly-concave min-max optimization problems*, being of independent interest to the private optimization and federated learning community.

In Algorithm 1, we implicitly assume that the *frequency* of each sensitive attribute is known in order to compute $\hat{P}_S$ and broadcast it to the silos. This assumption is not very restrictive: In practice, releasing the frequency of the sensitive attributes of data is very common. Moreover, it is straightforward to privately estimate $\hat{P}_S$ using DP histograms. Thus, for simplicity, we assumed $\hat{P}_S$ to be known.

In the next section, we show our framework extends to *hybrid* centralization settings.

4 Different Modes of Data Centralization

Recall that we have assumed each silo is divided into two distinct parts: one that holds the *sensitive* data and another that holds *non-sensitive* data. The two divisions within each silo can communicate with each other and with all the sensitive and non-sensitive divisions of other silos. Leveraging this subtlety, we show how to model a wide range of hybrid centralized/distributed learning tasks that involve privacy of sensitive attributes. We will illustrate how Algorithm 1 readily extends to these hybrid tasks.

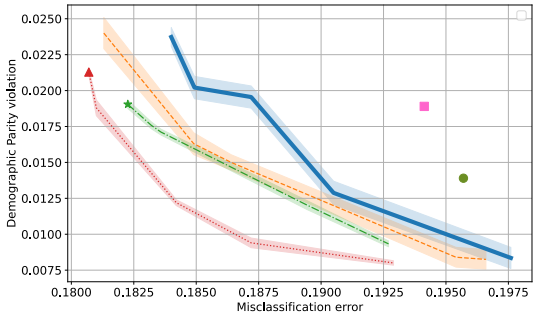
One Silo, Centralized Sensitive and Non-Sensitive Data in Separate Subdivisions. An example of this can be seen in healthcare organizations where the sensitive part of data can only be accessed by authorized personnel. In this case, we have $N = 1$ silo in SteFFLe. The updates from the sensitive subdivision are private due to the ISRL-DP guarantee in Theorem 6. Theorem 7 recovers the stationarity bound in Lowy et al. (2023).

Centralized Sensitive Data and Decentralized Non-Sensitive Data. In this case, we have 1 silo containing sensitive features and N silos containing non-sensitive features. Our algorithm can be used to train models in this setting: in round t, instead of querying silo i's sensitive division, the central server queries the central sensitive silo and receives noisy ISRL-DP sensitive gradients. These noisy sensitive gradients are combined with the noiseless non-sensitive gradients from each of the non-sensitive silos, and then the model is updated. An example of a silo containing centralized sensitive data is the *United States Census Bureau*. It provides essential demographic, social, and economic data that various institutions utilize for a wide range of purposes. Some examples of these institutions include government agencies, academics, and non-profit organizations. The data with these institutions correspond to silos with public data and any machine learning model they train for decision making would require a combination of their own data and the centralized sensitive data provided by the Census.

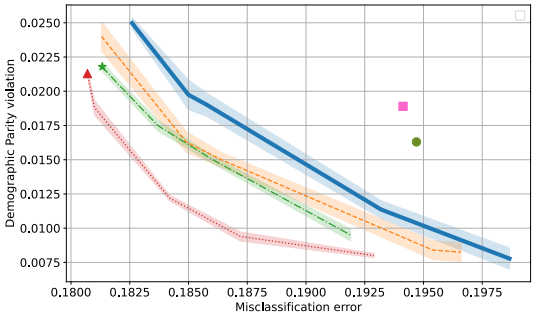
General Case: Arbitrary Numbers of Sensitive and Non-Sensitive Silos. Recall that every data-point we have is represented by a tuple $\{((x_u, y_u), s_u)\}_{u=1}^n$. We refer to (x_u, y_u) as the non-sensitive part of the datapoint and s_u to be the sensitive part. We assume that every silo indexes the data by universal index assigned to each data-point (say indexed by $1, 2, .., n$) instead of their local index. The data is distributed between the silos as follows:

- Let there be p silos (represented by $1, ..., p$) with non-sensitive parts of the data (non-sensitive silos) and s silos (represented by $1, ..., s$) with the sensitive parts of data (sensitive silos).
- Let $i \in [p]$ be the non-sensitive silo containing the non-sensitive attributes of datapoints which have indices $P_i \subset [n]$ such that $P_i \cap P_j = \phi$ for all $i, j \in [p]$ for $i \neq j$ and $\bigcup_{i=1}^p P_i = [n]$.
- Similarly, let any sensitive silo $i \in [s]$, contain the non-sensitive attributes of datapoints which have indices $S_i \subset [n]$ such that $S_i \cap S_j = \phi$ for all $i, j \in [s]$ for $i \neq j$ and $\bigcup_{i=1}^s S_i = [n]$.

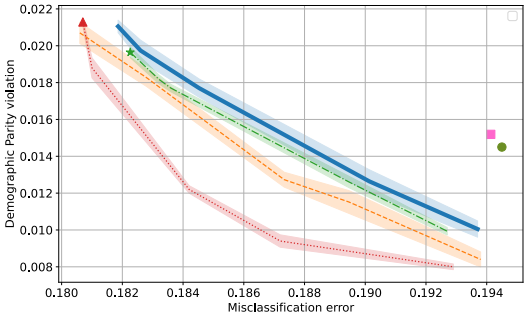
For the training to happen, the non-sensitive silos locally sample a batch of data $\mathcal{J}_j \subset P_j$ for all $j \in [p]$. They compute the gradients of the loss with using their part of data and broadcast it to the server. For the gradient of the regularizer, each non-sensitive silo j broadcasts $\mathcal{J}_j$ along with their respective model outputs. Then, the sensitive silos c (such that $S_c \bigcap \bigcup_{j=1}^p \mathcal{J}_j \neq \varnothing$) which have the data corresponding to the indices



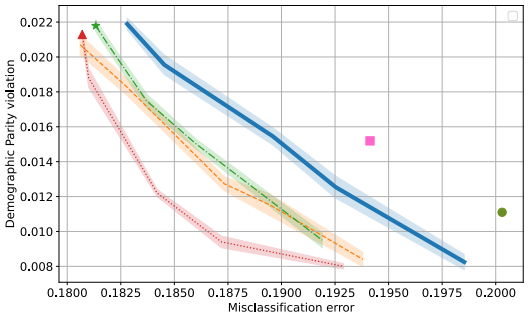
(a) $\varepsilon = 1$, *Homogeneous* ($h = 0$)



(b) $\varepsilon = 1$, *Heterogeneous* ($h = 0.75$)



(c) $\varepsilon = 3$, *Homogeneous* ($h = 0$)



(d) $\varepsilon = 3$, *Heterogeneous* ($h = 0.75$)

Category	Line Colour	Method	Federated	Differentially Private	Fair
Proposed		SteFFLe	✓	✓	✓
Baseline		Ling et al. (2024)	✓	✓	✓
		Tran et al. (2021b)	✗	✓	✓
Cost Comparison		Central DP Non-Federated SteFFLe	✗	✓	✓
		Non-Private SteFFLe	✓	✗	✓
		Non-Private Centralized (Lowy et al. 2022)	✗	✗	✓
		Non-Private Non-Fair FL (McMahan et al. 2017)	✓	✗	✗
		Non-Private Non-Fair Central ERM	✗	✗	✗

(e) Plot Legend

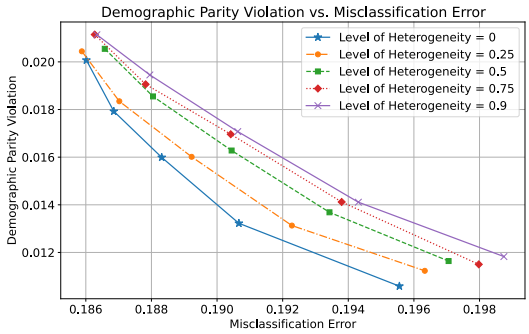
Figure 1: Demographic parity vs Misclassification error on *Credit Card* dataset (*Number of Silos* = 3)

sampled by all the non-sensitive silos ($S_c \bigcap \bigcup_{j=1}^{p} \mathcal{J}_j$) compute the gradient using the broadcasted outputs by the model and their local sensitive data. Then, these silos locally add noise according to batch size being $|S_c \bigcap \bigcup_{j=1}^{p} \mathcal{J}_j|$, and broadcast these noisy gradients to the server. The server aggregates both gradients from the non-sensitive and the sensitive silos and updates the model parameters.

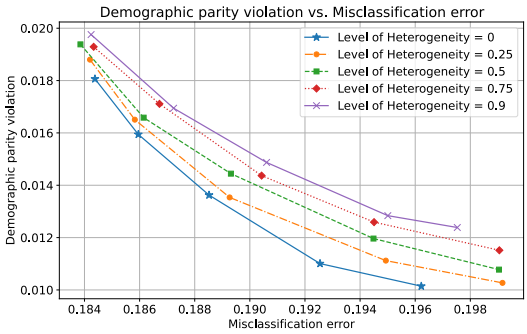
It is important that for every sensitive silo c scales its noise according to batch size being $|S_c \bigcap \bigcup_{j=1}^{p} \mathcal{J}_j|$ to preserve ISRL DP. However, since we have assumed that only the sensitive silos corresponding to the data will participate implying that $|S_c \bigcap \bigcup_{j=1}^{p} \mathcal{J}_j| \geqslant 1$. Hence, the upper bound on the stationarity gap would still exist with the value of batch size being one, thus still preserving a theoretical guarantee.

5 Numerical Experiments

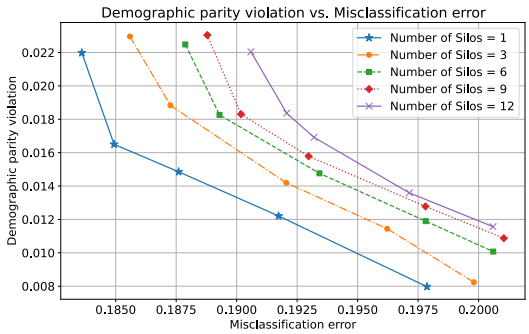
In this section, we evaluate the performance of our algorithm in terms of fairness violation vs. test error for different levels of privacy, levels of silo heterogeneity, and numbers of silos. We present our results in two parts: In Section 5.1, we assess the performance of our method in training logistic regression models



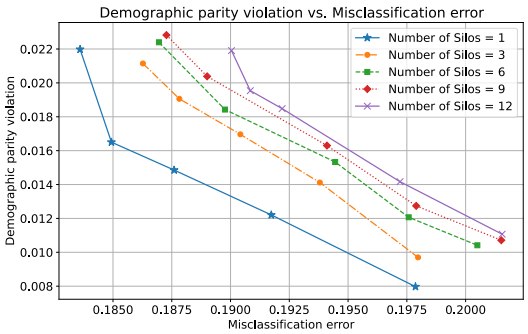
(a) $\varepsilon = 1$, *Number of Silos* $= 3$, Varying h



(b) $\varepsilon = 3$, *Number of Silos* $= 3$, Varying h



(c) $\varepsilon = 1$, *Homogeneous* ($h = 0$), Varying N



(d) $\varepsilon = 1$, *Heterogeneous* ($h = 0.75$), Varying N

Figure 2: Varying Levels of Heterogeneity (h) and Number of Silos (N) on the *Credit Card* dataset

on several benchmark tabular datasets. In Section 5.2, we discuss how the fairness-accuracy tradeoffs are affected by silo heterogeneity and by the number of silos for a fixed privacy level.

Average results. To evaluate the overall performance of our algorithm and the existing baselines, we calculated the performance gain with respect to fairness violation (for fixed accuracy level) that our model yields *over all the datasets*. We obtained reductions in demographic parity violations of around 75.47% and 52.93% compared with Tran et al. (2021) and Ling et al. (2024). Note that the algorithm of Tran et al. (2021) is *not ISRL-DP*, instead satisfying only the weaker notion of central DP. We also obtained an average reduction in equalized odds violation of 95.42% compared to Ling et al. (2024). We specify our experimental setup, datasets, methods and additional results that we compare against in Appendix E.

5.1 Federated, Private, and Fair Logistic Regression

In the first set of experiments we train a logistic regression model using SteFFLe (Algorithm 1) to promote demographic parity. We compare SteFFLe against all applicable publicly available baselines in each experiment. We carefully tuned the hyperparameters of all baselines for fair comparison. We find that *SteFFLe consistently outperforms all state-of-the-art baselines across all data sets in all privacy and heterogeneity levels.*

Baselines. The baselines include: (1) the approach by Tran et al. (2021), which is *central* differentially private and fair but *not federated and not ISRL-DP*; (2) the method of Ling et al. (2024), which incorporates federated learning, ISRL-DP, and fairness. These were the only DP fair baselines with code made publicly available for each experiment.

Additionally, we examine the *cost of incorporating federated learning and ISRL-DP* by measuring the fairness-accuracy trade-offs for different *variations of SteFFLE*. These variations include: *Central DP SteFFLe* Lowy et al. (2023), which is not ISRL-DP or federated, but still satisfies the weaker central DP notion and still promotes fairness; *Non-Private SteFFLe*, which is fair and federated, but not private; *Non-Private Centralized* Lowy et al. (2022a), which is fair, but not private or federated; *Non-Private Non-Fair FL* McMahan et al. (2018), which uses federated averaging; and *Non-Private Non-Fair Central ERM*, which simply uses SGD. See Figure 1 and the legend therein for our results on *Credit Card* dataset.

Datasets. We use three benchmark tabular datasets: Credit-Card, Adult Income, and Retired Adult dataset from the UCI machine learning repository (Dua & Graff (2017)). The predicted variables and sensitive attributes are both binary in these datasets. We analyze fairness-accuracy trade-offs with three different privacy budgets $\varepsilon \in \{1, 3, 9\}$ and two different values of heterogeneity levels $h = 0$ (homogeneous setting) and $h = 0.75$ (heterogeneous setting), keeping the number of silos $N = 3$ for each dataset. Note that these values of ε and heterogeneity levels are standard in the literature for empirically comparing different algorithms in private and federated learning methods Lowy et al. (2023); Ghoukasian & Asoodeh (2024). We compare against state-of-the-art algorithms proposed in Ling et al. (2024) and (the demographic parity objective of) Tran et al. (2021). The results displayed are averages over 15 trials (random seeds) for each value of ε, h and N.

Results for different datasets. Selected results for private and fair federated logistic regression on the Credit Card dataset are shown in Fig. 1. The remaining results of the Credit Card dataset and experiments of Adult and Retired Adult dataset are shown in Appendix E.5.1 and Appendix E.5.2. For logistic regression with equalized odds as the fairness violation, we provide further results (for a modified version of SteFFLe) on the Credit Card dataset in Appendix E.1. Compared to the baselines Tran et al. (2021) and Ling et al. (2024), SteFFLe offers superior fairness-accuracy tradeoffs at all privacy (ε) and heterogeneity levels (h) across all three datasets. Moreover, the method of Tran et al. (2021) is not ISRL-DP.

Training on Large scale Datasets In our second set of experiments, we train a large classifier ($d \approx 11.68$ millon) on the UTKFace dataset consisting of 20,000 images. The classifier categorizes facial images into nine age groups, following a setup similar to Tran et al. (2022), while treating race (with five classes) as the sensitive attribute. Our results clearly demonstrate that *Algorithm 1 converges in a non-binary classification setting with small batch sizes and non-binary sensitive attributes.* Further details about the experiments, numerical results and observations can be found in Appendix E.3.

5.2 Impact of Silo Heterogeneity and the Number of Silos

In this section, we analyze the impact on SteFFLe's performance due to *varying heterogeneity levels* and *the number of silos* on the fairness-error trade-off, with a fixed privacy budget of $\varepsilon = 1$. We analyze how these factors affect demographic parity violation and misclassification error on the Credit Card dataset, as depicted in Fig. 2.

Heterogeneous silo data is challenging in private fair FL. We conducted experiments with silo heterogeneity levels ranging from 0 to 0.9, with 0 being homogeneous and 1 being heterogeneous. In Fig.2(a) and 2(b), the results demonstrate a *clear increase in both misclassification error and demographic parity violation as heterogeneity increases*, for a fixed number of $N = 3$ silos. This indicates that higher silo heterogeneity exacerbates the model's difficulty in achieving an optimal balance between fairness and accuracy.

Fig. 2(c) and 2(d) illustrates the effect of the number of silos on performance in both the homogeneous and heterogeneous settings. We vary the number of silos between $N \in [1, 12]$. In the homogeneous settings, as the number of silos increases from 1 to 12, both demographic parity violation rise and misclassification error grows. A similar trend is apparent in the heterogeneous setting, where an increase in the number of silos results in a proportional rise in both demographic parity violation and misclassification error in Fig. 2 (d). These findings suggest that *increasing the number of silos amplifies the challenges of maintaining fairness and accuracy, particularly under federated learning frameworks which incorporate privacy constraints.*

6 Conclusion and Discussion

Motivated by pressing ethical and legal concerns, we considered the problem of training fair and private ML models with decentralized data. We developed an algorithm that satisfies the strong ISRL-DP guarantee. We proved that our ISRL-DP algorithm converges for any minibatch size, without requiring (strong) convexity of the loss function. Finally, numerical experiments on several benchmark fairness data sets demonstrated that our method offers substantial fairness-accuracy benefits over the prior art, across different levels of privacy and silo heterogeneity. Our experiments also highlighted the challenges of silo heterogeneity for fair and accurate ISRL-DP FL. We hope this work inspires further research in private and fair federated learning. Future directions include exploring fundamental trade-offs between ISRL-DP, fairness, and accuracy, as well as improving performance in heterogeneous settings.

References

Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016.

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *ICLR*, 2020.

PRESTON BUKATY. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing, 2019. ISBN 9781787781320. URL http://www.jstor.org/stable/j.ctvjghvnn.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, April 2020. ISSN 0001-0782. doi: 10.1145/3376898. URL https://doi.org/10.1145/3376898.

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

Jeffrey Dastin. Insight - amazon scraps secret ai recruiting tool that showed bias against women, 2018. URL https://shorturl.at/d5Hhv.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

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

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: enabling group fairness in federated learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25911. URL https://doi.org/10.1609/aaai.v37i6.25911.

Changyu Gao, Andrew Lowy, Xingyu Zhou, and Stephen J. Wright. Private heterogeneous federated learning without a trusted server revisited: Error-optimal and communication-efficient algorithms for convex losses, 2024. URL https://arxiv.org/abs/2407.09690.

Hrad Ghoukasian and Shahab Asoodeh. Differentially private fair binary classifications. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 611–616, 2024. doi: 10.1109/ISIT57864.2024.10619147.

Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2521–2529. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/girgis21a.html.

Xiuting Gu, Zhu Tianqing, Jie Li, Tao Zhang, Wei Ren, and Kim-Kwang Raymond Choo. Privacy, accuracy, and model fairness trade-offs in federated learning. *Computers Security*, 122:102907, 2022. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2022.102907. URL https://www.sciencedirect.com/science/article/pii/S0167404822003005.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016a.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, New York, New York, USA, 20–22 Jun 2016b. PMLR. URL http://proceedings.mlr.press/v48/hardt16.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pp. 3000–3008. PMLR, 2019.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2345–2355, 2017.

Zhuohang Li, Andrew Lowy, Jing Liu, Toshiaki Koike-Akino, Bradley Malin, Kieran Parsons, and Ye Wang. Exploring user-level gradient inversion with a diffusion prior. *arXiv preprint arXiv:2409.07291*, 2024a.

Zhuohang Li, Andrew Lowy, Jing Liu, Toshiaki Koike-Akino, Kieran Parsons, Bradley Malin, and Ye Wang. Analyzing inference privacy risks through gradients in machine learning. *arXiv preprint arXiv:2408.16913*, 2024b.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Xinpeng Ling, Jie Fu, Zhili Chen, Kuncan Wang, Huifa Li, Tong Cheng, Guanying Xu, and Qin Li. Fedfdp: Federated learning with fairness and differential privacy. *arXiv preprint arXiv:2402.16028*, 2024.

Ziyu Liu, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. On privacy and personalization in cross-silo federated learning. *arXiv preprint arXiv:2206.07902*, 2022.

Andrew Lowy and Meisam Razaviyayn. Output perturbation for differentially private convex optimization with improved population loss bounds, runtimes and applications to private adversarial training. *arXiv preprint:2102.04704*, 2021.

Andrew Lowy and Meisam Razaviyayn. Private federated learning without a trusted server: Optimal algorithms for convex losses. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TVY6GoURrw.

Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization. *Transactions on Machine Learning Research*, 2022a.

Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a trusted server. *arXiv preprint arXiv:2203.06735*, 2022b.

Andrew Lowy, Devansh Gupta, and Meisam Razaviyayn. Stochastic differentially private and fair learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3nM5uhPlfv6.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/pdf?id=BJ0hF1Z0b.

Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pp. 7066–7075. PMLR, 2020.

Manisha Padala, Sankarshan Damle, and Sujit Gujar. Federated learning meets fairness and differential privacy. In *International Conference on Neural Information Processing*, pp. 692–699. Springer, 2021.

Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS Workshop*, 2021. URL https://arxiv.org/pdf/2109.08604.pdf.

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, pp. 9932–9939, 2021.

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.

Zhang Zhifei Song Yang and Qi Hairong. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.