Proceedings of Machine Learning Research 1:1–33, 2022 Algorithmic Fairness through the Lens of Causality and Privacy

Stochastic Differentially Private and Fair Learning

Andrew Lowy LOWYA@USC.EDU
University of Southern California

Devansh Gupta* DEVANSH19160@IIITD.AC.IN
Indraprastha Institute of Information Technology, Delhi

Meisam Razaviyayn RAZAVIYA@USC.EDU
University of Southern California

Abstract

Machine learning models are increasingly used in high-stakes decision-making systems. In such applications, a major concern is that these models sometimes discriminate against certain demographic groups such as individuals with certain race, gender, or age. Another major concern in these applications is the violation of the privacy of users. While *fair learning* algorithms have been developed to mitigate discrimination issues, these algorithms can still *leak* sensitive information, such as individuals' health or financial records. Utilizing the notion of *differential privacy (DP)*, prior works aimed at developing learning algorithms that are both private and fair. However, existing algorithms for DP fair learning are either not guaranteed to converge or require full batch of data in each iteration of the algorithm to converge. In this paper, we provide the first *stochastic* differentially private algorithm for fair learning that is guaranteed to converge. Here, the term "stochastic" refers to the fact that our proposed algorithm converges even when *minibatches* of data are used at each iteration (i.e. *stochastic optimization*). Our framework is flexible enough to permit different fairness notions, including demographic parity and equalized odds. In addition, our algorithm can be applied to non-binary classification tasks with multiple (non-binary) sensitive attributes. As a byproduct of our convergence analysis, we provide the first utility guarantee for a DP algorithm for solving nonconvex-strongly concave min-max problems. Our numerical experiments show that the proposed algorithm *consistently offers significant performance gains over the state-of-the-art baselines*, and can be applied to larger scale problems with non-binary target/sensitive attributes.

1. Introduction

In recent years, machine learning algorithms have been increasingly used to inform decisions with far-reaching consequences (e.g. whether to release someone from prison or grant them a loan), raising concerns about their compliance with laws, regulations, societal norms, and ethical values. Specifically, machine learning algorithms have been found to discriminate against certain "sensitive" demographic groups (e.g. racial minorities), prompting a profusion of *algorithmic fairness* research (Dwork et al., 2012; Sweeney, 2013; Datta et al., 2015; Feldman et al., 2015; Bolukbasi et al., 2016; Angwin et al., 2016; Calmon et al., 2017; Hardt et al., 2016a; Fish et al., 2016; Woodworth et al., 2017; Zafar et al., 2017; Bechavod and Ligett, 2017; Kearns et al., 2018; Prost et al., 2019; Baharlouei et al., 2020; Lowy et al., 2022). Algorithmic fairness literature aims to develop fair machine learning algorithms that output non-discriminatory predictions.

* Work done as a Visiting Scholar at University of Southern California.

Fair learning algorithms typically need access to the sensitive data in order to ensure that the trained model is non-discriminatory. However, consumer privacy laws (such as the E.U. General Data Protection Regulation) restrict the use of sensitive demographic data in algorithmic decision-making. These two requirements–*fair algorithms* trained with *private data*–presents a quandary: how can we train a model to be fair to a certain demographic if we don't even know which of our training examples belong to that group?

The works of Veale and Binns (2017); Kilbertus et al. (2018) proposed a solution to this quandary using secure *multi-party computation (MPC)*, which allows the learner to train a fair model without directly accessing the sensitive attributes. Unfortunately, as Jagielski et al. (2019) observed, *MPC does not prevent the trained model from leaking sensitive data.* For example, with MPC, the output of the trained model could be used to infer the race of an individual in the training data set (Fredrikson et al., 2015; He et al., 2019; Song et al., 2020; Carlini et al., 2021). To prevent such leaks, Jagielski et al. (2019) argued for the use of *differential privacy* (Dwork et al., 2006) in fair learning. Differential privacy (DP) provides a strong guarantee that no company (or adversary) can learn much more about any individual than they could have learned had that individual's data never been used.

Since Jagielski et al. (2019), several follow-up works have proposed alternate approaches to DP fair learning (Xu et al., 2019; Ding et al., 2020; Mozannar et al., 2020; Tran et al., 2021b,a, 2022). As shown in Fig. 1, each of these approaches suffers from at least two critical shortcomings. In particular, *none of these methods have convergence guarantees when mini-batches of data are used in training.* In training large-scale models, memory and efficiency constraints require the use of small minibatches in each iteration of training (i.e. stochastic optimization). Thus, existing DP fair learning methods cannot be used in such settings since they require computations on the full training data set in every iteration. See Appendix A for a more comprehensive discussion of related work.

Our Contributions: In this work, we propose a novel algorithmic framework for DP fair learning. Our approach builds on the non-private fair learning method of Lowy et al. (2022). We consider a regularized empirical risk minimization (ERM) problem where the regularizer penalizes fairness violations, as measured by the *Exponential Rényi Mutual Information*. Using a result from Lowy et al. (2022), we reformulate this fair ERM problem as a min-max optimization problem. Then, we use an efficient differentially private variation of stochastic gradient descent-ascent (DP-SGDA) to solve this fair ERM min-max objective. The main features of our algorithm are:

1. *Guaranteed convergence* for any privacy and fairness level, *even when mini-batches of data are used* in each iteration of training (i.e. stochastic optimization setting). As discussed, stochastic optimization is essential in large-scale machine learning scenarios. Our algorithm is the first provably convergent stochastic DP fair learning method.

2. Flexibility to handle *non-binary* classification with *multiple (non-binary) sensitive attributes* (e.g. race and gender) under *different fairness notions* such as demographic parity or equalized odds. In each of these cases, our algorithm converges.

Empirically, we show that our method *outperforms the previous state-of-the-art methods* in terms of fairness vs. accuracy trade-off across all privacy levels. Moreover, our algorithm is capable of training with mini-batch updates and can handle *non-binary target and non-*

binary sensitive attributes. By contrast, existing DP fairness algorithms could not converge in our stochastic/non-binary experiment.

A byproduct of our algorithmic developments and analyses is *the first DP convergent algorithm for nonconvex min-max optimization*: namely, we provide an upper bound on the stationarity gap of DP-SGDA for solving problems of the form $\min_\theta \max_W F(\theta, W)$, where $F(\cdot, W)$ is non-convex. We expect this result to be of independent interest to the DP optimization community. Prior works that provide convergence results for DP min-max problems have assumed that $F(\cdot, W)$ is either (strongly) convex (Boob and Guzmán, 2021; Zhang et al., 2022) or satisfies a generalization of strong convexity known as the *Polyak-Łojasiewicz* (PL) condition (Yang et al., 2022).

Reference	Non-binary target?	Multiple fairness notions?	Convergence guarantee (poly. time)?	Guarantees with mini-batches?
This work	✓	✓	✓	✓
Jagielski et al. (2019) (post-proc.)*	✗	✗	N/A	✗
Jagielski et al. (2019) (in-proc.)	✗	✗	✗	✗
Xu et al. (2019)	✗	✗	✗	✗
Ding et al. (2020)	✗	✓	✗	✗
Mozannar et al. (2020)	✗	✓	✓	✗
Tran et al. (2021a)	✓	✗	✗	✗
Tran et al. (2021b)	✓	✓	✗	✗
Tran et al. (2022)	✓	✓	✗	✗

Figure 1: Comparison with existing works. "Guarantee" refers to *provable* guarantee. N/A: the post-processing method of Jagielski et al. (2019) is not an iterative algorithm. *Method requires access to the sensitive data at test time. The in-processing method of Jagielski et al. (2019) is inefficient. The work of Mozannar et al. (2020) specializes to equalized odds, but most of their analysis seems to be extendable to other fairness notions.

2. Preliminaries

Let $Z = \{z_i = (x_i, s_i, y_i)\}_{i=1}^n$ be a data set with non-sensitive features $x_i \in \mathcal{X}$, discrete sensitive attributes (e.g. race, gender) $s_i \in [k] \triangleq \{1, \ldots, k\}$, and labels $y_i \in [l]$. 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

$$\min_\theta \left\{ \widehat{\mathcal{L}}(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, x_i, y_i) \right\} \qquad (1)$$

in a fair manner, while maintaining the differential privacy of the sensitive data $\{s_i\}_{i=1}^n$. We consider two different notions of fairness in this work:[1]

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 predictions $\mathcal{A}(Z)$ are conditionally independent of sensitive attributes given $Y = y$ for all y.*

1. 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.

Depending on the specific problem at hand, one fairness notion may be more desirable than the other (Dwork et al., 2012; Hardt et al., 2016a). In practice, achieving exact fairness, i.e. (conditional) independence of $\widehat{Y}$ and S, is unrealistic (Cummings et al., 2019). Thus, we instead aim to design an algorithm that achieves small *fairness violation* on the given data set Z. Fairness violation can be measured in different ways: see e.g. Lowy et al. (2022) for a survey. For example, if demographic parity is the desired fairness notion, then one can measure (empirical) demographic parity violation by $\max_{\widehat{y}\in\mathcal{Y}} \max_{s\in\mathcal{S}} \left|\hat{p}_{\widehat{Y}|S}(\widehat{y}|s) - \hat{p}_{\widehat{Y}}(\widehat{y})\right|$, where $\hat{p}$ denotes an empirical probability calculated directly from $(Z, \{\widehat{y}_i\}_{i=1}^n)$.

Next, we define differential privacy (DP). Following Jagielski et al. (2019); Tran et al. (2021b, 2022), we consider a relaxation of DP, in which only the *sensitive attributes* require privacy. 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 $s_i \neq s_i'$.

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

$$\mathbb{P}(\mathcal{A}(Z) \in O) \leqslant e^{\epsilon}\mathbb{P}(\mathcal{A}(Z) \in O) + \delta, \tag{2}$$

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

Definition 2 is useful if a company wants to train a fair model, but is unable to use the sensitive attributes (which are needed to train a fair model) due to privacy concerns and laws (e.g., the E.U.'s GDPR). Definition 2 enables the company to privately use the sensitive attributes to train a fair model, while satisfying legal and ethical constraints. That being said, Definition 2 still may not prevent leakage of *non-sensitive* data. Thus, if the company is concerned with privacy of user data beyond the sensitive demographic attributes, then it should impose DP for all the features. Our algorithm and analysis readily extends to DP for all features: see Section 3.

Throughout the paper, we shall restrict attention to data sets that contain at least ρ-fraction of every sensitive attribute for some $\rho \in (0, 1)$: i.e. $\frac{1}{|Z|}\sum_{i=1}^{|Z|} \mathbb{1}_{\{s_i=r\}} \geqslant \rho$ for all $r \in [k]$. This is a reasonable assumption in practice: for example, if sex is the sensitive attribute and a data set contains all men, then training a model that is fair with respect to sex and has a non-trivial performance (better than random) seems almost impossible.

3. Private Fair ERM via Exponential Rényi Mutual Information

A standard in-processing strategy in the literature for enforcing fairness is to add a regularization term to the empirical objective that penalizes fairness violations (Zhang et al., 2018; Donini et al., 2018; Mary et al., 2019; Baharlouei et al., 2020; Cho et al., 2020; Lowy et al., 2022). We can then jointly optimize for fairness and accuracy by solving

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

where $\mathcal{D}$ is some measure of statistical (conditional) dependence between the sensitive attributes and the predictions (given Y), and $\lambda \geqslant 0$ is a scalar balancing fairness and accuracy. The choice of $\mathcal{D}$ is crucial and can lead to different fairness-accuracy profiles. Inspired by the strong empirical performance and amenability to stochastic optimization of Lowy et al. (2022), we choose $\mathcal{D}$ to be the Exponential Rényi Mutual Information (ERMI):

Definition 3 (ERMI – Exponential Rényi Mutual Information) *We define the exponential Rényi mutual information between random variables $\widehat{Y}$ and S with empirical joint distribution $\hat{p}_{\widehat{Y},S}$ and marginals $\hat{p}_{\widehat{Y}}$, $\hat{p}_S$ by:*

$$\widehat{D}_R(\widehat{Y},S) := \mathbb{E}\left\{\frac{\hat{p}_{\widehat{Y},S}(\widehat{Y},S)}{\hat{p}_{\widehat{Y}}(\widehat{Y})\hat{p}_S(S)}\right\} - 1 = \sum_{j\in[l]}\sum_{r\in[k]}\frac{\hat{p}_{\widehat{Y},S}(j,r)^2}{\hat{p}_{\widehat{Y}}(j)\hat{p}_S(r)} - 1 \tag{ERMI}$$

Definition 3 is what we would use if *demographic parity* were the desired fairness notion. If instead one wanted to encourage equalized odds, then Theorem 3 can be readily adapted to these fairness notions by substituting appropriate conditional probabilities for $\hat{p}_{\widehat{Y},S}$, $\hat{p}_{\widehat{Y}}$, and $\hat{p}_S$ in (ERMI): see Appendix B for details.[2] It can be shown that ERMI $\geqslant 0$, and is zero if and only if demographic parity (or equalized odds, for the conditional version of ERMI) is satisfied (Lowy et al., 2022). Further, any algorithm that makes ERMI small will also have small fairness violation with respect to other notions of fairness violation (Lowy et al., 2022). Lastly, (Lowy et al., 2022, Proposition 2) shows that empirical ERMI (Theorem 3) is an asymptotically unbiased estimator of "population ERMI"–which can be defined as in Definition 3, but with empirical distributions replaced by their population counterparts.

Our approach to enforcing fairness is to augment (1) with an ERMI regularizer and privately solve:

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

There are numerous ways to privately solve (FERMI obj.). For example, one could use the exponential mechanism (McSherry and Talwar, 2007), or run noisy gradient descent (GD) (Bassily et al., 2014). The problem with these approaches is that they are inefficient or require computing n gradients at every iteration, which is prohibitive for large-scale problems, as discussed earlier. We could *not* run noisy *stochastic* GD (SGD) on (FERMI obj.) because we do not (yet) have a statistically unbiased estimate of $\nabla_\theta\widehat{D}_R(\widehat{Y}_\theta(X),S)$.

Our next goal is to derive a *stochastic*, differentially private fair learning algorithm. 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 $\mathcal{F}(x,\theta)$ is differentiable in θ, has range $[0,1]^l$, 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 4 (Lowy et al. (2022)) *There are differentiable functions $\widehat{\psi}_i$ such that $\widehat{\psi}_i(\theta,\cdot)$ is strongly concave for all θ and (FERMI obj.) is equivalent to*

$$\min_\theta \max_{W\in\mathbb{R}^{k\times l}} \left\{\widehat{F}(\theta,W) := \widehat{\mathcal{L}}(\theta) + \lambda\frac{1}{n}\sum_{i=1}^n \widehat{\psi}_i(\theta,W)\right\}. \tag{3}$$

The functions $\widehat{\psi}_i$ are given explicitly in Appendix C. Theorem 4 is useful because it permits us to use *stochastic* optimization to solve (FERMI obj.): for any batch size $m \in [n]$, the gradients (with respect to θ and W) of $\frac{1}{m}\sum_{i\in\mathcal{B}}\ell(x_i,y_i;\theta) + \lambda\widehat{\psi}_i(\theta,W)$ are statistically unbiased estimators of the gradients of $\widehat{F}(\theta,W)$, if $\mathcal{B}$ is drawn uniformly from Z. However,

2. To simplify the presentation, we will assume that demographic parity is the fairness notion of interest in the remainder of this section. However, we consider both fairness notions in our numerical experiments.

when DP of the sensitive attributes is also desired, the formulation (3) presents some challenges, due to the non-convexity of $\widehat{F}(\cdot, W)$. Indeed, *there is no known DP algorithm for solving non-convex min-max problems that provably converges.* Next, we provide the first such convergence guarantee.

3.1. Noisy DP-FERMI for Stochastic Private Fair ERM

Our proposed stochastic DP algorithm for solving (FERMI obj.), is given in Algorithm 1. It is a DP variation of stochastic gradient descent ascent (SGDA) (Lin et al., 2020).

Algorithm 1 DP-FERMI Algorithm for Private Fair ERM

1: **Input**: $\theta_0 \in \mathbb{R}^{d_\theta}$, $W_0 = 0 \in \mathbb{R}^{k\times l}$, step-sizes (η_θ, η_w), fairness parameter $\lambda \geqslant 0$, iteration number T, minibatch size $|B_t| = m \in [n]$, set $\mathcal{W} \subset \mathbb{R}^{k\times l}$, noise parameters $\sigma_w^2, \sigma_\theta^2$.
2: Compute $\widehat{P}_S^{-1/2}$.
3: **for** $t = 0, 1, \ldots, T$ **do**
4: Draw a mini-batch B_t of data points $\{(x_i, s_i, y_i)\}_{i\in B_t}$
5: Set $\theta_{t+1} \leftarrow \theta_t - \frac{\eta_\theta}{|B_t|}\sum_{i\in B_t}[\nabla_\theta \ell(x_i, y_i; \theta^t) + \lambda(\nabla_\theta \widehat{\psi}_i(\theta_t, W_t) + u_t)]$, where $u_t \sim \mathcal{N}(0, \sigma_\theta^2 \mathbf{I}_{d_\theta})$.
6: Set $W_{t+1} \leftarrow \Pi_{\mathcal{W}}\Big(W_t + \eta_w\Big[\frac{\lambda}{|B_t|}\sum_{i\in B_t}\nabla_w \widehat{\psi}_i(\theta_t, W_t) + V_t\Big]\Big)$, where V_t is a $k \times l$ matrix with independent random Gaussian entries $(V_t)_{r,j} \sim \mathcal{N}(0, \sigma_w^2)$.
7: **end for**
8: Pick $\hat{t}$ uniformly at random from $\{1, \ldots, T\}$.
9: **Return:** $\hat{\theta}_T := \theta_{\hat{t}}$.

Explicit formulae for $\nabla_\theta \widehat{\psi}_i(\theta_t, W_t)$ and $\nabla_w \widehat{\psi}_i(\theta_t, W_t)$ are given in Lemma 7 (Appendix D). We provide the privacy and convergence guarantees of Algorithm 1 in Theorem 5:

Theorem 5 *Assume the loss function $\ell(\cdot, x, y)$ and $\mathcal{F}(x, \cdot)$ are Lipschitz continuous with Lipschitz gradient for all (x, y), and $\widehat{P}_S(r) \geqslant \rho > 0 \ \forall\ r \in [k]$. Then there exist algorithmic parameters such that Algorithm 1 is (ϵ, δ)-DP and*

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

For large-scale models (e.g. deep neural nets), we typically have $d_\theta \gg 1$ and $k, l = \mathcal{O}(1)$, so that the convergence rate of Algorithm 1 is essentially immune to the number of labels and sensitive attributes. In contrast, *no existing works with convergence guarantees are able to handle non-binary classification* ($l > 2$), *even with full batches* and a single binary sensitive attribute. Also, *the utility bound in Theorem 5 corresponds to DP for all of the features.*

In Theorem 9 of Appendix E, we prove more generally that noisy DP-SGDA converges to an approximate stationary point of *any* smooth nonconvex-strongly concave min-max optimization problem (not just (3)). We expect Theorem 9 to be of general interest to the DP optimization community beyond its applications to DP fair learning, since it is the first DP convergence guarantee for nonconvex min-max optimization.

4. Numerical Experiments

In this section, we evaluate the performance of our proposed approach (DP-FERMI) in terms of the fairness violation vs. test error for different privacy levels. We present our results in two parts: In Section 4.1, we assess the performance of our method in training logistic regression models on several benchmark tabular datasets. Since this is a standard setup that existing DP fairness algorithms can handle, we are able to compare our method against the state-of-the-art baselines. We carefully tuned the hyperparameters of all baselines for fair comparison. We find that *DP-FERMI consistently outperforms all state-of-the-art baselines across all data sets and all privacy levels*. These observations hold for both demographic parity and equalized odds fairness notions. To quantify the improvement of our results over the state-of-the-art 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 a performance gain of demographic parity that was 79.648 % better than Tran et al. (2021b) on average, and 65.89% better on median. The average performance gain of equalized odds was 96.65% while median percentage gain was 90.02%. In Section 4.2, we showcase the *scalability* of DP-FERMI by using it to train a deep convolutional neural network for classification on a large image dataset. In Appendix F, we give detailed descriptions of the data sets, experimental setups and training procedure, along with additional results.

4.1. Standard Benchmark Experiments: Logistic Regression

We train a logistic regression model using DP-FERMI (Algorithm 1) for demographic parity and a modified DP-FERMI (described in Appendix F) for equalized odds. We compare DP-FERMI against all applicable publicly available baselines in each experiment.

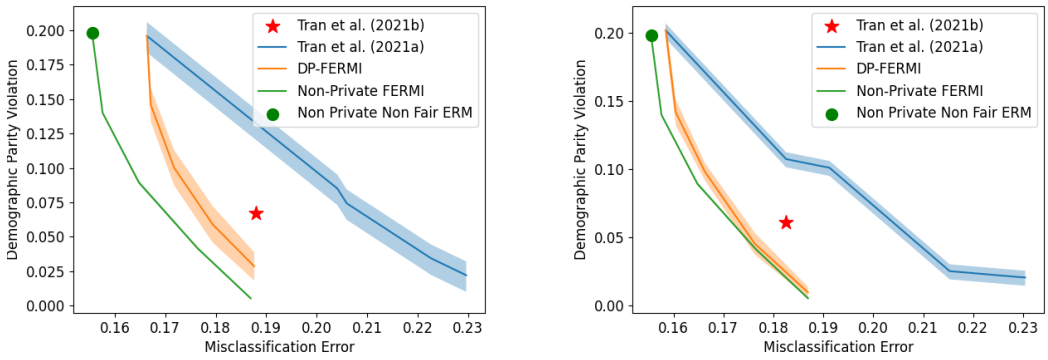
4.1.1. Demographic Parity

We use four benchmark tabular datasets: Adult Income, Retired Adult, Parkinsons, and Credit-Card dataset from the UCI machine learning repository (Dua and Graff (2017)). The predicted variables and sensitive attributes are both binary in these datasets. We analyze fairness-accuracy trade-offs with four different values of $\epsilon \in \{0.5, 1, 3, 9\}$ for each dataset (see Appendix F for complete results). We compare against state-of-the-art algorithms proposed in Tran et al. (2021a) and (the demographic parity objective of) Tran et al. (2021b). The results displayed are averages over 15 trials (random seeds) for each ϵ value.

For the Adult dataset, the task is to predict *whether the income is greater than \$50K or not* keeping *gender* as the sensitive attribute. The Retired Adult dataset is the same as the Adult dataset, but with updated data. The results for Adult and Retired Adult are shown in Figs. 2 and 7 (in Appendix F.2). DP-FERMI offers superior fairness-accuracy tradeoffs at every privacy (ϵ) level.

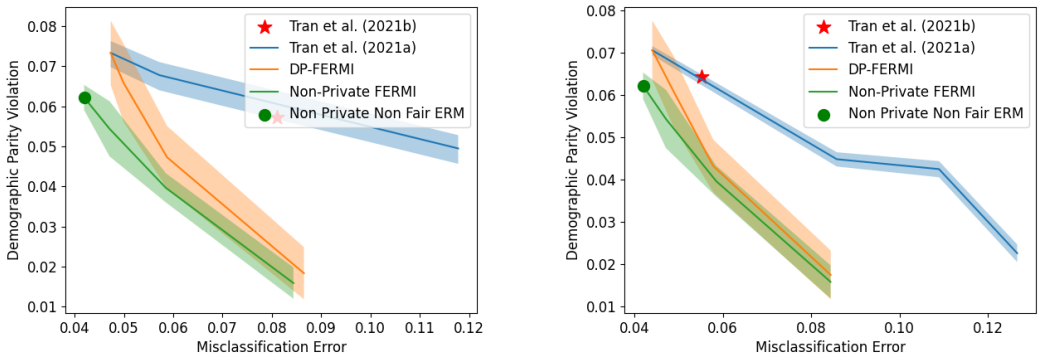
In the Parkinsons dataset, the task is to predict *whether the total UPDRS score of the patient is greater than the median or not* keeping *gender* as the sensitive attribute. Results for $\epsilon \in \{1, 3\}$ are shown in Fig. 3. Our algorithm again outperforms the baselines Tran et al. (2021a,b) for all tested privacy levels.

In the Credit Card dataset , the task is to predict *whether the user will default payment the next month* keeping *gender* as the sensitive attribute. Results are shown in Fig. 8 in Appendix F.2. DP-FERMI provides the best privacy-fairness-accuracy profile.



(a) $\epsilon = 0.5$ (b) $\epsilon = 1$

Figure 2: Private, Fair (Demographic Parity) logistic regression on Adult Dataset.



(a) $\epsilon = 1$ (b) $\epsilon = 3$

Figure 3: Private, Fair (Demogrpahic Parity) logistic regression on Parkinsons Dataset

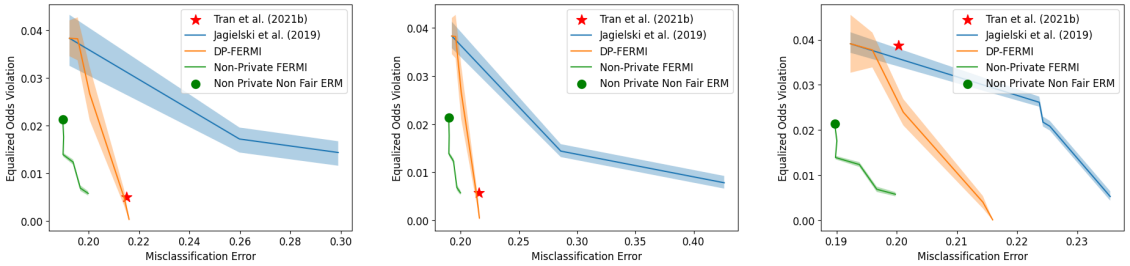
4.1.2. Equalized Odds

Next, we consider the slightly modified version of Algorithm 1, which is designed to minimize the Equalized Odds violation by replacing the absolute probabilities in the objective with class conditional probabilities: see Appendix F.2 for details.

We considered the Credit Card and Adult datasets for these experiments, using the same sensitive attributes as mentioned above. Results for Credit Card are shown in Fig. 4. Adult results are given in Fig. 10 in Appendix F.2. Compared to Jagielski et al. (2019) and the equalized odds objective in Tran et al. (2021b), our *equalized odds variant of DP-FERMI outperforms these state-of-the-art baselines at every privacy level.*

4.2. Training a Deep Convolutional Neural Network on Image Dataset

In our second set of experiments, we train a deep 9-layer VGG-like classifier (Simonyan and Zisserman, 2015) with $d \approx 1.6$ million parameters on the UTK-Face dataset (Zhang et al., 2017) using Algorithm 1. We classify the facial images into 9 age groups similar to the setup in Tran et al. (2022), while keeping *race* (containing 5 classes) as the sensitive attribute. See Appendix F.3 for more details.We analyze consider with four different privacy levels $\epsilon \in \{10, 25, 50, 100\}$. Compared to the tabular datasets, larger ϵ is needed to obtain stable results in the large-scale setting since the number of parameters d is much larger and the

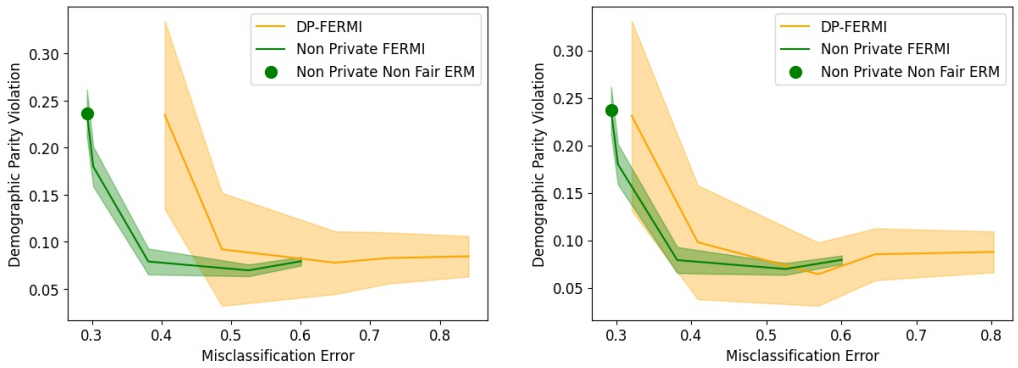


(a) $\epsilon = 0.5$ (b) $\epsilon = 1$ (c) $\epsilon = 3$

Figure 4: Private, Fair (Equalized Odds) logistic regression on Credit Card Dataset

cost of privacy increases with d (see Theorem 5). Larger values of $\epsilon > 100$ were used in the baseline Jagielski et al. (2019) for smaller scale experiments.

The results in Fig. 5 empirically verify our main theoretical result: *DP-FERMI converges even for non-binary classification with small batch size and non-binary sensitive attributes.* We took Tran et al. (2021a,b) as our baselines and attempted to adapt them to this non-binary large-scale task. We observed that the baselines were very unstable while training and mostly gave degenerate results. By contrast, our method was able to obtain stable and meaningful tradeoff curves. Also, while Tran et al. (2022) reported results on UTK-Face, their code is not publicly available and we were unable to reproduce their results.



(a) $\epsilon = 25$ (b) $\epsilon = 50$

Figure 5: DP-FERMI on a Deep CNN for Image Classification on UTK-Face

5. Concluding Remarks

Motivated by pressing legal, ethical, and social considerations, we studied the challenging problem of learning fair models with private demographic data. We observed that existing approaches require full batches of data in each iteration (and/or exponential runtime) in order to provide convergence/accuracy guarantees. We addressed this limitation by deriving a DP stochastic optimization algorithm for fair learning, and rigorously proved the convergence of our method. Finally, we evaluated our method in extensive experiments and found that it significantly outperforms the previous state-of-the-art models, in terms of fairness-accuracy tradeoff. Potential societal impacts of our work are discussed in Appendix G.

Acknowledgments

This work was supported in part with funding from the NSF CAREER award 2144985, from the YIP AFOSR award, from a gift from the USC-Meta Center for Research and Education in AI & Learning, and from a gift from the USC-Amazon Center on Secure & Trusted Machine Learning.

References

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`.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

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

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, 2019.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. Equal opportunity in online classification with partial feedback. *arXiv preprint arXiv:1902.02242*, 2019.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

Digvijay Boob and Cristóbal Guzmán. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *arXiv preprint arXiv:2104.02988*, 2021.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf`.

Jiahao Ding, Xinyue Zhang, Xiaohuan Li, Junyi Wang, Rong Yu, and Miao Pan. Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 622–629, 2020.

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*, pages 2791–2801, 2018.

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

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

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*, pages 214–226, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 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, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016b. PMLR. URL `http://proceedings.mlr.press/v48/hardt16.html`.

Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 148–162, 2019.

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*, pages 3000–3008. PMLR, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR, 2018.

Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 277–287. PMLR, 26–28 Aug 2020.

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*, pages 2345–2355, 2017.

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

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

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR, 2019.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

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

Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Mengkai Song, Zhibo Wang, Zhifei Zhang, Yang Song, Qian Wang, Ju Ren, and Hairong Qi. Analyzing user-level privacy attack against federated learning. *IEEE Journal on Selected Areas in Communications*, 38(10):2430–2444, 2020. doi: 10.1109/JSAC.2020.3000372.

Latanya Sweeney. Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*, 2013.

Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34:27555–27565, 2021a.

Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9932–9939, 2021b.

Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In *IJCAI*, pages 560–566, 2021c.

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.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

Depeng Xu, Shuhan Yuan, and Xintao Wu. Achieving differential privacy and fairness in logistic regression. In *Companion proceedings of The 2019 world wide web conference*, pages 594–599, 2019.

Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Varshney, Siwei Lyu, and Yiming Ying. Differentially private sgda for minimax problems. *arXiv preprint arXiv:2201.09046*, 2022.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 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*, pages 335–340, 2018.

Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *arXiv preprint arXiv:2206.00363*, 2022.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.