FairDP: Certified Fairness with Differential Privacy

Khang Tran
New Jersey Institute of Technology, USA
kt36@njit.edu

Ferdinando Fioretto
Syracuse University, USA
ffiorett@syr.edu

Issa Khalil
Qatar Computing Research Institute, Qatar
ikhalil@hbku.edu.qa

My T. Thai
University of Florida, USA
mythai@cise.ufl.edu

NhatHai Phan*
New Jersey Institute of Technology, USA
phan@njit.edu

Abstract

This paper introduces **FAIRDP**, a novel mechanism designed to simultaneously ensure differential privacy (DP) and fairness. FAIRDP operates by independently training models for distinct individual groups, using group-specific clipping terms to assess and bound the disparate impacts of DP. Throughout the training process, the mechanism progressively integrates knowledge from group models to formulate a comprehensive model that balances privacy, utility, and fairness in downstream tasks. Extensive theoretical and empirical analyses validate the efficacy of FAIRDP, demonstrating improved trade-offs between model utility, privacy, and fairness compared with existing methods.

1 Introduction

The proliferation of machine learning (ML) systems in decision-making processes has brought important considerations regarding privacy, bias, and discrimination. These requirements are becoming pressing as ML systems are increasingly used to make decisions that significantly impact individuals' lives, such as in healthcare, finance, and criminal justice. These concerns underscore the need for ML algorithms that can guarantee both privacy and fairness.

Differential Privacy (DP) is an algorithmic property that helps protect the sensitive information of individuals by preventing disclosure during computations. In the context of machine learning, it enables algorithms to learn from data while ensuring they do not retain sensitive information about any specific individual in the training data. However, it has been found that DP systems may produce biased and unfair results for different groups of individuals [2, 12, 40], which can have a significant impact on their lives, particularly in areas such as finance, criminal justice, or job-hiring [11].

The issue of balancing privacy and fairness in ML systems has been the subject of much discussion in recent years. For example, [6] showed the existence of a tradeoff between differential privacy and equal opportunity, a fairness criterion that requires a classifier to have equal true positive rates for different groups. Different studies have also reported that when models are trained on data with long-tailed distributions, it is challenging to develop a private learning algorithm that has high accuracy for minority groups [33]. These findings have led to the question of whether fair models can be created while preserving sensitive information and have spurred the development of various

*Corresponding author

Preprint. Under review.

approaches [17, 24, 35, 36, 37] (see Appendix D for further discussion on related work). While these studies have contributed to a deeper understanding of the trade-offs between privacy and fairness, as well as the importance of addressing these issues in a unified manner, they all share a common limiting factor: *the inability to provide formal guarantees for both privacy and fairness simultaneously*. This aspect is essential and cannot be overstated. In many critical application contexts, such as those regulated by policy and laws, these guarantees are often required, and failure to provide them can prevent adoption or deployment [34].

This paper aims to address this gap by proposing novel mechanisms that simultaneously achieve differential privacy and provide certificates on fairness. The main challenges in developing such a mechanism are: **(1)** Designing appropriate DP algorithms that can limit the impact of privacy-preserving noise on the model bias; and **(2)** Balancing the trade-offs between model utility, privacy, and fairness, while simultaneously providing useful fairness certificates.

Contributions. The paper makes two main contributions to address these challenges. First, it introduces a novel DP training mechanism called FAIRDP, which ensures certified fairness. The mechanism controls the amount of noise injected into groups of data points classified by fairness-sensitive attributes, such as race and gender. By controlling the disparate effects of noise on model fairness through group-specific clipping terms, FAIRDP enables the derivation and tightening of certified fairness bounds. Throughout the training process, the mechanism progressively integrates knowledge from each group model, leading to improved trade-offs between model utility, privacy, and fairness. Second, it conducts extensive experiments to analyze the interplay among utility, privacy, and fairness using various benchmark datasets. The results show that FAIRDP provides a better balance between privacy and fairness compared to existing baselines, including both DP-preserving mechanisms with or without fairness constraints.

The significance of our theoretical and empirical analysis becomes apparent as it emphasizes the need to develop novel approaches for effectively combining data privacy preservation and fairness. In this context, FairDP represents an innovative solution that bridges this critical void.

2 Background and Research Goal

The paper considers datasets $D = \{(x_i, a_i, y_i)\}_{i=1}^n$ whose samples are drawn from an unknown distribution. Therein, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ is a sensitive feature vector, $a_i \in \mathcal{A} = [K]$ is a (group of) protected group attribute(s), and $y_i \in \mathcal{Y} = \{0, 1\}$ is a binary class label, similar to previous work [5, 38, 18]. For example, consider a classifier for predicting whether individuals may qualify for a loan. The data features x_i may describe the individuals' education, current job, and zip code. The protected attribute a_i may describe the individual's gender or race, and the label y_i indicates whether the individual would successfully repay a loan or not. The paper also uses notation $D_k = \{(x_i, a_i = k, y_i)\}_{i=1}^{n_k}$ to denote a non-overlapping partition over dataset D which contains exclusively the individuals belonging to a protected group k and $\cap_k D_k = \emptyset$. Although the results in this paper consider only one protected attribute, the results can be directly generalized to multiple protected attributes (see Appendix E).

Research Goal. The paper studies models $h_\theta : \mathcal{X} \to [0, 1]$ parameterized by $\theta \in \mathbb{R}^r$ and the learning task optimizes the empirical loss function:

$$\mathcal{L}(D) = \min_\theta \sum_{(x_i, a_i, y_i) \in D} \ell\left(h_\theta(x_i), y_i\right),$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a differentiable loss function. The goal is to train models satisfying three key properties: **(1)** *Privacy*: the model parameters θ are protected to prevent information leakage from the training data; **(2)** *Fairness*: the released model is unbiased towards any protected group, with theoretical guarantees; and **(3)** *Utility*: at the same time, the model's utility is maximized.

The paper uses h_θ and h_{θ_k} to denote, respectively, the models minimizing the empirical loss $\mathcal{L}(D)$ over the entire dataset and that minimizing $\mathcal{L}(D_k)$ using data from the corresponding group k.

Differential Privacy. Differential privacy (DP) [9] is a strong privacy concept ensuring that the likelihood of any outcome does not change significantly when a record is added or removed from a dataset. An adjacent dataset (D') of D is created by adding or removing a record from D. Such a relation is denoted by $D \sim D'$. t

Definition 2.1 ([9]). *A mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-differential privacy, if, for any two adjacent inputs $D \sim D' \in \mathcal{D}$, and any subset of outputs $R \subseteq \mathcal{R}$:*

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

Parameter $\epsilon > 0$ describes the *privacy loss* of the algorithm, with values close to 0 denoting strong privacy, while parameter $\delta \in [0, 1)$ captures the probability of failure of the algorithm to satisfy ϵ-DP. The global sensitivity Δ_f of a real-valued function $f : \mathcal{D} \to \mathbb{R}$ is defined as the maximum amount by which f changes in two adjacent inputs: $\Delta_f = \max_{D \sim D'} \|f(D) - f(D')\|$. In particular, the Gaussian mechanism, defined by $\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2 \mathbf{I})$, where $\mathcal{N}(0, \sigma^2)$ is the Gaussian distribution with 0 mean and standard deviation σ^2, satisfies (ϵ, δ)-DP for $\sigma = \Delta_f \sqrt{2 \log(1.25/\delta)}/\epsilon$.

Group Fairness and Certified Guarantee. This paper considers a general notion of statistical fairness metrics, defined as follows:

Definition 2.2. *General Notion of Fairness. The fairness of a given model $h_\theta(\cdot)$ is quantified by*

$$\texttt{Fair}(h_\theta) = \max_{u,v \in [K]} |Pr(\hat{y} = 1 | a = u, e) - Pr(\hat{y} = 1 | a = v, e)|, \tag{1}$$

where $\hat{y}$ is the model's prediction and e is a random event.

The fairness notion in Equation (1) captures several well-known fairness metrics, including demographic parity [23] (when $e = \emptyset$), equality of opportunity [14] (when e is the event "$y = 1$"), and equality of odd [14] (when $e = y$). If a model h_θ satisfies $\texttt{Fair}(h_\theta) \leq \tau$, for $\tau \in [0, 1]$, then we say that h_θ *achieves certification of τ-fairness*. Intuitively, as τ decreases, the model's decision becomes more independent of the protected attribute, given the random event e.

3 Certified Fairness with DP (FairDP)

This section introduces FAIRDP, a mechanism that addresses two key objectives: **(1)** the realization of (ϵ, δ)-differential privacy (DP) and **(2)** the provision of a provable τ-fairness guarantee. Central to this approach is the use of a stochastic gradient descent (SGD) training process. However, developing FAIRDP poses a significant challenge in balancing the disparate impact of the DP-preserving noise on specialized model predictions for different protected groups while also ensuring τ-fairness certification. Moreover, as the model parameters are updated under DP preservation during each training round, they become intricate in infinite parameter space, adding complexity to achieving τ-fairness guarantees. Finding a solution to these intertwined challenges is difficult since DP preservation and fairness can substantially reduce the model's performance, particularly without a carefully calibrated noise injection process.

3.1 FAIRDP and Privacy Guarantee

To overcome these challenges, FAIRDP relies on two key strategies: Firstly, it restricts the model parameters within a finite space, enabling us to establish a tractable boundary for the model's DP-preserving noise-influenced predictions. In a neural network, FAIRDP uses an l_2-norm clipping on the final layer weights of model h_θ, a technique also applicable to models prioritizing privacy. Second, rather than training a single model h_θ, FAIRDP trains a set of group-specific models $\{h_{\theta_k}\}_{k=1}^K$ with each θ_k being independently learned to minimize the loss $\mathcal{L}(D_k)$. This approach not only allows to preserve each group's privacy, enhancing control over noise injection per group, but also progressively aggregates group models to construct a (general) model h_θ. In doing so, FAIRDP effectively combines and propagates knowledge from each group to balance privacy, fairness, and utility.

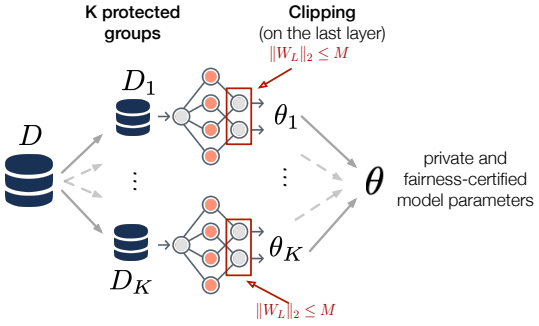

Figure 1: FAIRDP: A schematic overview.

FAIRDP. A schematic illustration of the algorithm is shown in Figure 1 and its training process is outlined in Algorithm 1. Let us consider, without loss of generality, h_θ as an L-layers neural network, where $\theta = \{W_1, \ldots, W_L\}$, W_j contains the weights at the j^{th} layer, and the activation of the last

layer is a sigmoid function for a binary classification task. In each training round t, FAIRDP clips the l_2-norm of the final layer's weights $W_L^{(t-1)} \in \theta^{(t-1)}$ by M (line 4). For each group k, FAIRDP initializes the group model parameters $\theta_k^{(t-1)}$ using the clipped model parameters $\theta^{(t-1)}$ (line 5). It then draws a batch of data points B_k from the corresponding group dataset D_k with probability q. The l_2-norm of the gradient derived from each data point in the batch is then constrained by a predefined upper-bound C (line 9). Next, Gaussian noise $\mathcal{N}(0, C^2\sigma^2\mathbf{I}_r)$ is introduced to the sum of clipped gradients $\Delta\bar{g}_k$ from all data points, ensuring DP preservation. Here, r denotes the number of model weights, $\mathbf{I}_r$ is an identity matrix of size r, and σ is a DP-preserving noise scale (line 11).

The group model's parameters θ_k are updated using DP-preserving gradients through standard SGD with a learning rate η (line 12). In order to construct the (general) model h_θ, the parameters of the group models are aggregated as: $\theta^{(t)} = \frac{\theta_1^{(t)} + \cdots + \theta_K^{(t)}}{K}$ (line 14). The aggregated model parameters $\theta^{(t)}$ are used as the parameters for every group model in the next training round (line 5). These aggregation and propagation steps (lines 5 and 14) ensure that the final model parameters $\theta^{(T)}$, where T is the number of update steps, are close to the parameters of every group, reducing bias towards any specific group and distilling knowledge from every group to improve model utility.

The parameters $\theta^{(T)}$ returned by the model satisfy (ϵ, δ)-DP.

Algorithm 1 Certified Fairness with DP (FairDP)

1: **Input**: Dataset D, sampling rate q, learning rate η, noise scale σ, gradient norm bound C, number of steps T, parameter norm bound M.
2: Initialize $\theta^{(0)} = \{W_1^{(0)}, \ldots, W_L^{(0)}\}$ randomly
3: **for** $t \in [1:T]$ **do**
4: **Clip weights**: $W_L^{(t-1)} = W_L^{(t-1)} \min\left(1, \frac{M}{\|W_L^{(t-1)}\|_2}\right)$
5: $\theta_1^{(t-1)} = \cdots = \theta_K^{(t-1)} = \theta^{(t-1)}$
6: **for** $k \in \{1, \ldots, K\}$ **do**
7: Sample B_k from D_k with sampling probability q.
8: **Compute gradient**: For $x_i \in B_k^{(t)}$, $g_i = \nabla_{\theta_k}\ell(x_i)$
9: **Clip gradient**: $\bar{g}_i = g_i \min(1, \frac{C}{\|g_i\|_2})$
10: **Compute total gradient**: $\Delta\bar{g}_k = \sum_{i \in B_k} \bar{g}_i$
11: **Add noise**: $\Delta\tilde{g}_k = \Delta\bar{g}_k + \mathcal{N}(0, C^2\sigma^2 I_r)$
12: **Update**: $\theta_k^{(t)} = \theta_k^{(t-1)} - \eta\Delta\tilde{g}_k$
13: **end for**
14: $\theta^{(t)} = \frac{\theta_1^{(t)} + \cdots + \theta_K^{(t)}}{K}$
15: **end for**
16: **Return** $\theta^{(T)}$

Theorem 3.1. *Algorithm 1 satisfies (ϵ, δ)-DP where ϵ is calculated by the moment accountant [1] given the sampling probability q, T update steps, and the noise scale σ.*

The proof of all theorems are reported in the Appendix.

3.2 Fairness Certification

To derive fairness certification, we focus on the last layer (L^{th}) of the (general) model h_θ since it directly produces the predictions. The L^{th} layer consists of an input $z_{L-1} \in \mathbb{R}^f$ and an output $z_L \in \mathbb{R}$, before the application of the sigmoid activation function. If $sigmoid(z_L) > 0.5$ (equivalent to $z_L > 0$), then the prediction of (general) model h_θ is $\hat{y} = 1$; otherwise the prediction is $\hat{y} = 0$.

Given a group model h_{θ_k}, DP-preserving noise injected into clipped gradients $\Delta\bar{g}_k$ (line 11) transforms the gradients of the last layer, denoted as μ_k, into a random variable following a multivariate Gaussian distribution $\mathcal{N}(\mu_k; \sigma^2C^2\mathbf{I}_f)$, as follows: $\tilde{\mu}_k = \mu_k + \mathcal{N}(0; \sigma^2C^2\mathbf{I}_f)$. As a result, the weights at the last layer for the group k at every step t, denoted by $W_{L,k}^{(t)}$, becomes a random variable with the following distribution $\mathcal{N}(W_{L,k}^{(t-1)} - \eta\mu_k; \eta^2\sigma^2C^2\mathbf{I}_f)$.

Notice that the weight $W_L^{(t)}$ of the (general) model h_θ is a linear combination of the K multivariate Gaussian random variables $\{W_{L,k}^{(t)}\}_{k\in[K]}$. Based on the fact that the linear combination of multivariate Gaussian random variables is also multivariate Gaussian distributed [3], the weight $W_L^{(t)}$ follows a multivariate Gaussian distribution, as follows:

$$W_L^{(t)} \sim \mathcal{N}\Big(\frac{1}{K}\sum_{k=1}^{K} W_{L,k}^{(t-1)} - \frac{\eta}{K}\sum_{k=1}^{K}\mu_k; \frac{\eta^2\sigma^2C^2}{K}\mathbf{I}_f\Big) \text{ or } W_L^{(t)} \sim \mathcal{N}\Big(W^{(t-1)} - \eta\mu; \frac{\eta^2\sigma^2C^2}{K}\mathbf{I}_f\Big),$$

where $W^{(t-1)} = \frac{1}{K}\sum_{k=1}^{K} W_{L,k}^{(t-1)}$ and $\mu = \frac{1}{K}\sum_{k=1}^{K}\mu_k$.

Since the output $z_L = W_L^{(t)\top} z_{L-1}$ is a linear combination of the Gaussian random variable, z_L is a random variable in one dimension following a Gaussian distribution:

$$z_L \sim \mathcal{N}\Big(\langle W^{(t-1)} - \eta\mu, z_{L-1}\rangle; \frac{1}{K}\|z_{L-1}\|_2^2\eta^2\sigma^2C^2\Big), \tag{2}$$

where $\langle\cdot,\cdot\rangle$ is the inner product between two vectors.

As a result of Eq. (2), the (general) model h_θ predicts z_L derived from a data point $x \in \mathbb{R}^d$ as a positive value with the probability $Pr(\hat{y} = 1|x) = Pr(z_L > 0) = 1 - Pr(z_L \leq 0)$, where the probability $Pr(z_L \leq 0)$ can be computed as follows:

$$Pr(z_L \leq 0) = \frac{1}{2}\left[1 + \texttt{erf}\Big(\frac{-\langle W^{(t-1)} - \eta\mu, z_{L-1}\rangle\sqrt{K}}{\|z_{L-1}\|_2\eta\sigma C\sqrt{2}}\Big)\right]; \texttt{erf}(\cdot) \text{ is the } \textit{error function} \tag{3}$$

Eq. (3) follows the cumulative distribution function of one-dimension Gaussian distribution up to $z_L = 0$[2] [3]. Therefore, we have the following

$$Pr(\hat{y} = 1|x) = 1 - Pr(z_L \leq 0) = \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\langle W^{(t-1)} - \eta\mu, z_{L-1}\rangle\sqrt{K}}{\|z_{L-1}\|_2\eta\sigma C\sqrt{2}}\Big) \tag{4}$$

$$= \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\langle W^{(t-1)} - \eta\mu, z_{L-1}\rangle\|W^{(t-1)} - \eta\mu\|_2\|z_{L-1}\|_2\sqrt{K}}{\|W^{(t-1)} - \eta\mu\|_2\|z_{L-1}\|_2\|z_{L-1}\|_2\eta\sigma C\sqrt{2}}\Big), \tag{5}$$

Since $\frac{\langle W^{(t-1)} - \eta\mu, z_{L-1}\rangle}{\|W^{(t-1)} - \eta\mu\|_2\|z_{L-1}\|_2} = \cos\phi$, with ϕ being the angle between vectors $(W^{(t-1)} - \eta\mu)$ and z_{L-1}, we have the following

$$Pr(\hat{y} = 1|x) = \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\|W^{(t-1)} - \eta\mu\|_2\sqrt{K}}{\eta\sigma C\sqrt{2}}\cos\phi\Big). \tag{6}$$

From Eq. (6), $\cos(\phi) \in [-1, 1]$, and the monotonicity of the error function[3], one can upper bound and lower bound the probability that $\hat{y} = 1$ given x, as follows:

$$\frac{1}{2} - \frac{1}{2}\texttt{erf}\left(\frac{\|W^{(t-1)} - \eta\mu\|_2}{\eta\sigma C\sqrt{\frac{2}{K}}}\right) \leq Pr(\hat{y} = 1|x) \leq \frac{1}{2} + \frac{1}{2}\texttt{erf}\left(\frac{\|W^{(t-1)} - \eta\mu\|_2}{\eta\sigma C\sqrt{\frac{2}{K}}}\right). \tag{7}$$

Since the weights $W^{(t-1)}$ and gradients μ are bounded by the clipping in FAIRDP (lines 4 and 9); that is $\|W^{(t-1)}\|_2 \leq M$ and $\|\mu\|_2 \leq \frac{m}{K}C$ where $m = \sum_{k=1}^{K}|B_k|$ is the total size of all training batches across protected groups $|B_k|$ in a training round, we can derive a τ-fairness certification of $\texttt{Fair}(h_\theta)$ from Eq. (7) in the following theorem:

Theorem 3.2. *A general model h_θ optimized by Algorithm 1 satisfies τ- fairness certification with,*

$$\texttt{Fair}(h_\theta) \leq \texttt{erf}\Big(\frac{(MK + \eta mC)\sqrt{K}}{K\eta\sigma C\sqrt{2}}\Big). \tag{8}$$

Remark. Theorem 3.2 provides an upper-bound on the τ-fairness certification, revealing a novel insight into the trade-off among privacy, fairness, and utility. It is worth noting that the upper bound of fairness certification τ decreases as the DP-preserving noise scale σ increases. As a result, stronger privacy (larger σ values) correspond to enhanced fairness certification (smaller τ values), due to the increased randomness influencing the model's decisions. While this theoretical impact of DP-preserving noise augments privacy and fairness assurances in our model, it could potentially diminish utility. Our theoretical observation is consistent with previous empirical studies [42, 25].

[2] If the prediction process uses a threshold other than 0.5, this probability can still be computed by an inverse of the sigmoid function to find the corresponding value for the cumulative distribution.

[3] The error function is an increasing function, i.e. if $x_1 < x_2$, then $\texttt{erf}(x_1) < \texttt{erf}(x_2)$.

3.3 Tightening Fairness Certification

While an important result, the τ-fairness certification in Theorem 3.2 lacks sufficient tightness due to the batch size m and the learning rate η included in the error function. Larger m and smaller η, which are common in typical model training, can result in a looser τ-fairness. Therefore, in our second main contribution, we derive an empirical fairness bound that substantially tightens the τ-fairness certification, enabling a pragmatic understanding of the privacy, fairness, and utility trade-offs.

By leveraging Eq. (4), the empirical fairness bound can be calculated on-the-fly (i.e., during model training). For a specific group k, at every update step t, the probability $Pr(\hat{y}=1|a=k,e)$ can be empirically computed as follows:

$$P_{emp}(\hat{y}=1|a=k,e) = \frac{1}{n_{k,e}} \sum_{x \in D_{k,e}} P_{emp}(\hat{y}=1|x) = \frac{1}{n_{k,e}} \sum_{x \in D_{k,e}} P_{emp}(z_L > 0) \tag{9}$$

$$= \frac{1}{2} + \frac{1}{2n_{k,e}} \sum_{x \in D_{k,e}} \texttt{erf}\Big(\frac{\langle W^{(t-1)} - \eta\mu, z_{L-1} \rangle \sqrt{K}}{\|z_{L-1}\|_2 \eta \sigma C \sqrt{2}} \Big), \tag{10}$$

where we use the real values of $W^{(t-1)}$, μ, and z_{L-1} at every round t, and $n_{k,e}$ is the size of $D_{k,e}$.

The empirical fairness certificate can be generalized to different fairness metrics by considering the event e. In fact, $D_{k,e} = D_k$ for ***demographic parity***, $D_{k,e}$ is the set of data point in D_k with the positive label for ***equality of opportunity***, and $D_{k,e}$ is the set of data point in D_k with the positive label when computing true positive rate or the negative label when computing false positive rate for ***equality of odd***. Finally, the empirical τ-fairness certification of the general model h_θ can be computed by $\arg\max_{u,v \in [K]} |P_{emp}(\hat{y}=1|a=u,e) - P_{emp}(\hat{y}=1|a=v,e)|$.

Proposition 3.3. *A model h_θ optimized by Algorithm 1 satisfies **empirical** τ_{emp}-fairness certification with $\tau_{emp} = \arg\max_{u,v \in [K]} |P_{emp}(\hat{y}=1|a=u,e) - P_{emp}(\hat{y}=1|a=v,e)|$.*

Utility, Privacy, and Fairness Trade-offs. FAIRDP is, to our knowledge, the first mechanism that simultaneously preserves DP and attains both theoretical and empirical certification of τ-fairness, all without sacrificing model utility, as demonstrated in the experimental results below. Additionally, Theorems 3.2 and Proposition 3.3 provide an insightful theoretical understanding of the interplay between privacy, fairness, and utility. A stronger privacy guarantee (larger noise scale σ) tends to result in better fairness certification (smaller τ value), even though it could potentially compromise model utility.

Remark. Practitioners can leverage our results to more effectively balance the trade-offs among privacy, fairness, and utility by adaptively adjusting the training process of FAIRDP. For example, the application of optimizers like Adam [20] at the training onset may lead to enhanced model utility and convergence rate under identical DP protection. As the model nears convergence, practitioners can transition to SGD to secure fairness certification, enabling us to overcome tight constraints on the weights of the last layer. Also, practitioners can adjust the hyper-parameter M to achieve better fairness. As in Theorem 3.2, the lower value of M, the fairer the model is. However, small M could degrade model utility since it constrains the decision boundary in a smaller parameter space (see Figure 13, Appendix F, for details).

4 Experimental Results

In this section, a comprehensive evaluation of FAIRDP and several baseline methods are conducted on various benchmark datasets. The evaluation primarily focuses on two aspects: **(1)** Examining the trade-off between model utility, privacy, and fairness, and **(2)** Assessing the accuracy of the fairness certification by comparing it with empirical results obtained from multiple statistical fairness metrics.

4.1 Datasets, Metrics and Model Configurations

The evaluation uses four datasets: Adult and Abalone datasets from the UCI Machine Learning Repository [7], Default of Credit Card Clients (Default-CCC) dataset [44] and Law School Admissions (Lawschool) dataset [39]. Details of the datasets are presented in Table 1. Data preprocessing steps are strictly followed as outlined in previous works such as [16, 32, 35]. On the Lawschool,

Table 1: Evaluation Datasets.

Data set	# data points	# features	Protected Attribute	# positive label	Size of minor group
Lawschool	86,022	33	Race	23,243	15,311
Default-CCC	30,000	89	Gender	6,636	11,460
Adult	48,842	41	Gender	11,687	16,192
Abalone	1,418	7	Gender	915	654

Adult, and Abalone[4] datasets, the model's performance is evaluated by *accuracy* as in previous studies [10, 13, 18, 41]. In contrast, the Default-CCC dataset evaluates it by *precision* due to its heavy imbalance. A *higher* accuracy/precision indicates *better* performance. *Demographic parity* [8], *equality of opportunity*, and *equality of odds* [15] are used as primary fairness metrics.

We employ a multi-layer perceptron (MLP) with ReLU activation on hidden layers and Sigmoid activation on the last layer for binary classification tasks. The baseline models use Adam optimizer [20] during the complete training process, while FAIRDP uses Adam for the first 90% of the training steps and then switches to vanilla SGD for the remaining steps. For FAIRDP, we set the weight clipping hyper-parameter $M \in [0.1, 1.0]$ and initialize the learning rate $\eta = 0.02$ when using Adam, and then reduce it to $\eta = 0.005$ when switching to SGD.

4.2 Baselines

To thoroughly evaluate FAIRDP, we consider a variety of DP-preserving mechanisms, fairness training algorithms, and combinations of these as our baselines. This results in eight baselines, including a standard MLP, four existing mechanisms that either preserve DP or promote fairness, one adapted mechanism that achieves both DP and fairness, and two variants of FAIRDP.

Established Baselines. We consider **DPSGD** [1], functional mechanism **(FM)** [45], **DPSGDF** [41], and **FairSmooth** [18] as baselines. Both **DPSGD** and **FM** are well-established DP mechanisms with many applications in DP research [29, 31, 27]. **DPSGDF** is designed to alleviate the disparate impact of DPSGD by focusing on accuracy parity. **FairSmooth** is a state-of-the-art mechanism that assures group fairness by transforming the model h_θ into a smooth classifier as $\hat{h}_\theta = \mathbb{E}_\nu[h_{(\theta+\nu)}]$ where $\nu \sim \mathcal{N}(0, \bar{\sigma}^2)$ in the inference process, where $\bar{\sigma}$ is the standard deviation of the Gaussian noise. Moreover, we introduce a new baseline, **DPSGD-Smooth**, by applying **FairSmooth** to a logistic regression model trained by **DPSGD**. This gives rise to the only baseline offering both DP and fairness guarantee, which we employ for comparison against FAIRDP.

Variants of FAIRDP. To examine how different features of FAIRDP affect the model performance and fairness, we introduce two FAIRDP variants, called **FairFM**, and **FairFM-Smooth**. **FairFM** (refer to Appendix A) distinguishes itself from FAIRDP by incorporating noise into the objective function as a pre-processing step to preserve DP. The mechanism trains group-specific model parameterized by θ_k, relative to dataset D_k, to optimize the objective $\theta_k^* = \arg\min_{\theta_k} \frac{1}{|D_k|} \sum_{(x_i,y_i)\in D_k} \ell(h_{\theta_k}(x_i), y_i)$. In the preprocessing step, the objective function of each group is approximated using a second-order Taylor's expansion [4], and the corresponding polynomial form $\mathcal{L}_k(\theta_k) = \theta_k^\top \lambda_k^{(2)} \theta_k + \theta_k^\top \lambda_k^{(1)} + \lambda_k^{(0)}$ is derived, where $\lambda_k^{(j)}, j = 0, 1, 2$ are the coefficients of the order j^{th} associated with group k. Then, Laplace noise [9] is added to the coefficients to derive the DP-preserving objective function $\tilde{\mathcal{L}}_k(\theta_k)$ and each group's perturbed objective function is optimized using SGD. The **FairFM-Smooth** mechanism is a variant of FairFM that applies the FairSmooth method [18] to the model trained by FairFM during the inference process.

The experiments use a range of privacy budgets across different datasets. For the Adult dataset, we set $\epsilon \in [0.1, 2.0]$; for other datasets, we use an expansive range with $\epsilon \in [0.5, 10.0]$. Although DP is celebrated for using small values of ϵ, most current deployments[5] report ϵ larger than 1 with many of them use ϵ larger than 5 and 10. Therefore, since fairness is affected by privacy loss, we believe our study is important to highlight and justify the trade-offs between privacy and fairness within this privacy loss regime. Statistical tests used are two-tailed t-tests.

[4] We present the results on the Abalone dataset, which is smaller than others, in Appendix.

[5] https://desfontain.es/privacy/real-world-differential-privacy.html

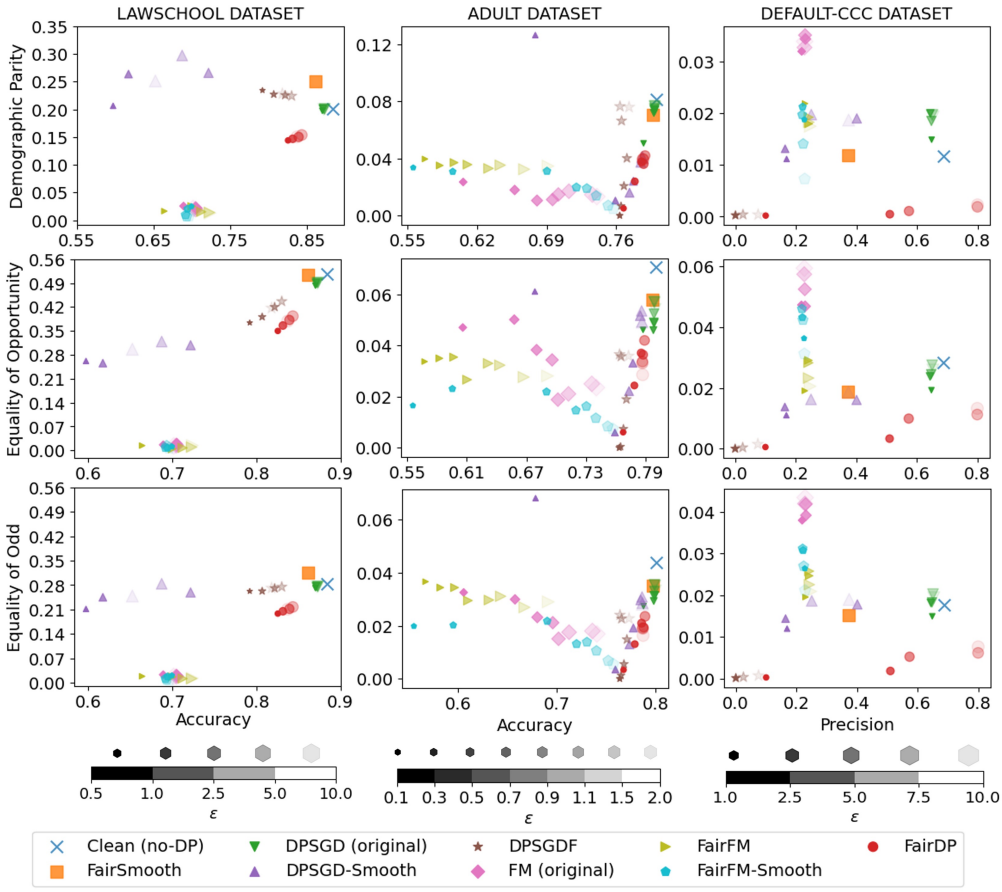

Figure 2: Trade-off among model performance, DP-preservation, and fairness.

4.3 Results

Utility, Privacy, and Fairness Trade-offs. Figures 2 and 5 (Appendix F) show the performance of each algorithm w.r.t. model utility, fairness, and privacy. In Figure 2, points positioned closer to the bottom-right corner denote superior balance among model performance (characterized by higher accuracy/precision), privacy (illustrated by strict DP protection), and fairness (represented by lower empirical values of statistical fairness metrics). Darker and smaller points indicate the application of smaller privacy budgets, translating to stricter DP protection, and the inverse holds as well.

Remarkably, our proposed FairDP consistently outperforms all baselines across all datasets, striking a balance among model utility, privacy, and fairness. For instance, in the Lawschool dataset, FairDP attains lower demographic parity (0.149 vs 0.2 in DPSGD, p-value = $2.53e^{-9}$) with small degenration in model utility (83.6% vs 87.1% in DPSGD) and similar DP protection ($\epsilon \in [0.5, 10.0]$). Despite having better Equality of Opportunity, DPSGD-Smooth suffers an 18.1% performance drop compared to FairDP (p-value $p = 4.85e^{-5}$). Compared to the best fairness algorithm, FairFM-Smooth, FairDP achieves superior accuracy (83.6% vs. 69.5%, p-value = $2.52e^{-9}$) and highly competitive demographic parity, enhancing fairness under stringent DP protection. Similar findings can be observed in the Adult, Default-CCC, and Abalone datasets (see Appendix F).

Remark. The promising results of FairDP can be attributed to its unique approach of controlling the amount of DP-preserving noise injected into each group, enforcing a constraint on the decision boundary, and fusing the knowledge learned from all groups together at each training step. This approach fundamentally differs from existing methods, leading to superior performance in FairDP.

Another noteworthy observation is that treating fairness as a constraint, as in the case of DPSGDF, does not consistently improve the trade-offs among model utility, privacy, and fairness. For instance,

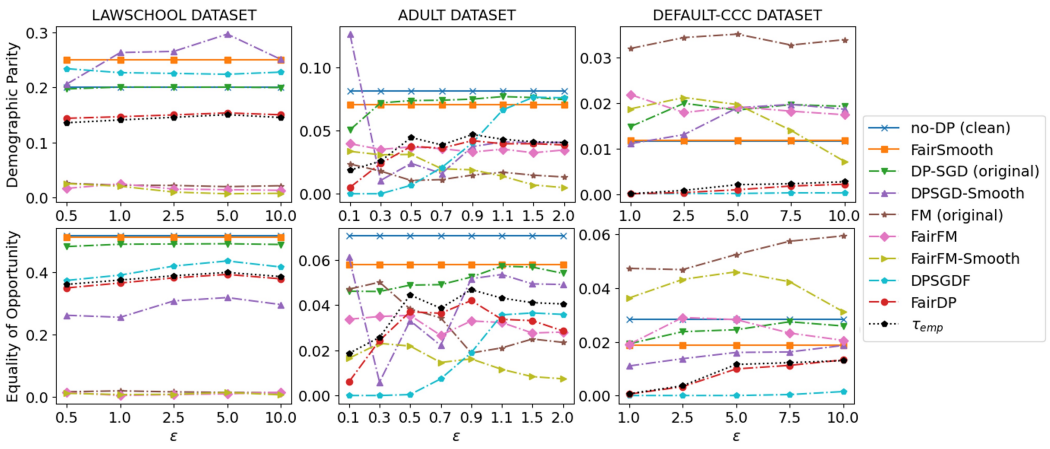

Figure 3: Tightness of fairness certification compared to empirical results of demographic parity and equality of opportunity for different privacy budgets.

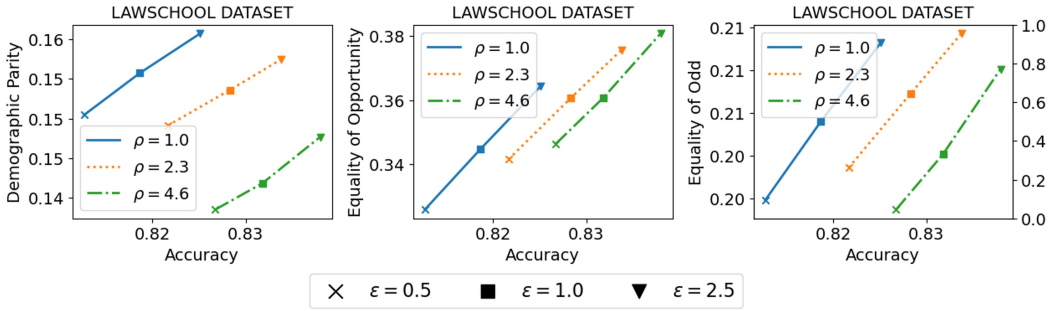

Figure 4: The trade-off among utility, privacy, and fairness for various ρ values on Lawschool dataset.

in the Lawschool dataset, DPSGDF is less fair than the original DPSGD in terms of demographic parity (0.23 compared with 0.2 in demographic parity with $p = 3.84e^{-7}$). A similar effect is observed in the Abalone dataset (Figure 5, Appendix F). This can be attributed to the fact that handling all groups simultaneously, within the noisy SGD process, can hide the information from minor groups, leading to degradation in fairness. Also, the fairness constraints, employed as penalty functions, have an impact on the optimization of the model, leading to a deterioration in its utility.

These issues can be mitigated by separating the DP-preserving training process from the methods developed to attain fairness during inference, as in the case of DPSGD-Smooth and FairFM-Smooth. These methods achieve better τ-fairness with relatively competitive model utility under equivalent DP protection. However, this approach does not effectively balance the trade-offs among model utility, privacy, and fairness as effectively as FairDP does. *These insights highlight the need to explore novel approaches to seamlessly integrate DP-preserving and fairness rather than treating them as independent (constrained) components.* FairDP *represents a pioneering step in this direction.*

Tightness of the Fairness Certification. Figure 3 and 7 to 9 (Appendix F) show the empirical fairness results and the certification value τ_{emp}. In most instances for the Lawschool, Adult, and Default-CCC datasets, our empirical certifications are substantially lower than the empirical fairness values of the baselines, particularly for DP-preserving mechanisms, without a significant drop in model performance. In particular, in the Default-CCC dataset, our empirical certifications are significantly smaller than the empirical fairness results of the state-of-the-art FairSmooth and DPSGDF ($p = 3.07e^{-8}$), while maintaining a small gap with the empirical fairness results of FairDP ($< 5\%$ of deviation). That illustrates the tightness of our certification of τ_{emp}-fairness across datasets and privacy budgets, further strengthening the advantages of FairDP in both theoretical guarantees and empirical results compared with existing baselines.

Imbalanced Protected Group. Practitioners can tune FairDP to find an appropriate setting that balances the level of DP protection with the desired level of fairness and model utility. Figures 4 and

10 through 12 (Appendix F) illustrate the effect of the ratio ρ between the size of the datasets of the minor and major groups: $\rho = (\arg\max_{a\in[K]} n_a)/(\arg\min_{b\in[K]} n_b)$. For a specific ρ, we randomly sample data points from the majority group, reducing the size of the major group training set to the desired ρ, while the test sets remain unchanged for all groups. In general, increasing ρ values lead to a greater number of data points from the majority group being utilized for training the model, thereby improving its accuracy. However, the effect on the model's fairness across different fairness metrics is not consistently observed. Nonetheless, our theoretical guarantee remains applicable across various degrees of dataset imbalance. Lower privacy budgets (indicating stronger privacy guarantees) contribute to improved fairness in the model's decisions, thereby reinforcing the theoretical assurances provided by FAIRDP.

5 Conclusion

This paper introduced FAIRDP, a novel mechanism that, for the first time, ensures both differential privacy and certified group fairness, while sustaining superior model performance. FAIRDP provides a comprehensive understanding of the influence of noise on model fairness. Besides the theoretical analysis, the paper examined the empirical certification bounds and showed that FAIRDP offers enhanced trade-offs among model utility, privacy, and fairness, outperforming an array of baselines.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, pages 15479–15488, 2019.

[3] W. Bryc. *The normal distribution: characterizations with applications*, volume 100. Springer Science & Business Media, 2012.

[4] C. Canuto, A. Tabacco, C. Canuto, and A. Tabacco. Taylor expansions and applications. *Mathematical Analysis I*, pages 225–257, 2015.

[5] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pages 1349–1361. PMLR, 2021.

[6] R. Cummings, V. Gupta, D. Kimpara, and J. 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.

[7] D. Dua and C. Graff. UCI machine learning repository, 2017.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

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

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. 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.

[11] F. Fioretto, C. Tran, and P. V. Hentenryck. Decision making with differential privacy under a fairness lens, 2021.

[12] F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. In *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5470–5477, 2022.

[13] U. Gupta, A. M. Ferber, B. Dilkina, and G. Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619, 2021.

[14] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

[15] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[16] E. Iofinova, N. Konstantinov, and C. H. Lampert. Flea: Provably fair multisource learning from unreliable training data. *arXiv preprint arXiv:2106.11732*, 2021.

[17] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR, 2019.

[18] J. Jin, Z. Zhang, Y. Zhou, and L. Wu. Input-agnostic certified group fairness via gaussian parameter smoothing. In *International Conference on Machine Learning*, pages 10340–10361. PMLR, 2022.

[19] N. Jovanović, M. Balunović, D. I. Dimitrov, and M. Vechev. Fare: Provably fair representation learning. *arXiv preprint arXiv:2210.07213*, 2022.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. *ICLR*, 2018.

[22] F. D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

[23] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[24] H. Mozannar, M. I. Ohannessian, and N. Srebro. Fair learning with private demographic data. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[25] M. Pannekoek and G. Spigler. Investigating trade-offs in utility, fairness and differential privacy in neural networks. *arXiv preprint arXiv:2102.05975*, 2021.

[26] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[27] H. Phan, M. T. Thai, H. Hu, R. Jin, T. Sun, and D. Dou. Scalable differential privacy with certified robustness in adversarial learning. In *ICML*, pages 7683–7694. PMLR, 2020.

[28] N. Phan, M. Vu, Y. Liu, R. Jin, D. Dou, X. Wu, and M. T. Thai. Heterogeneous gaussian mechanism: Preserving differential privacy in deep learning with provable robustness. *IJCAI*, 2019.

[29] N. Phan, Y. Wang, X. Wu, and D. Dou. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[30] N. Phan, X. Wu, and D. Dou. Preserving differential privacy in convolutional deep belief networks. *Machine learning*, 106(9):1681–1704, 2017.

[31] N. Phan, X. Wu, H. Hu, and D. Dou. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *2017 IEEE international conference on data mining (ICDM)*, pages 385–394. IEEE, 2017.

[32] A. Ruoss, M. Balunovic, M. Fischer, and M. Vechev. Learning certified individually fair representations. *Advances in neural information processing systems*, 33:7584–7596, 2020.

[33] A. Sanyal, Y. Hu, and F. Yang. How unfair is private learning? In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 1738–1748, 2022.

[34] I. P. Team. *EU General Data Protection Regulation (GDPR).* IT Governance Limited, 2017.

[35] C. Tran, M. H. Dinh, and F. Fioretto. Differentially private deep learning under the fairness lens, 2021.

[36] C. Tran, F. Fioretto, and P. V. Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 9932–9939. AAAI Press, 2021.

[37] C. Tran, K. Zhu, F. Fioretto, and P. Van Hentenryck. Sf-pate: scalable, fair, and private aggregation of teacher ensembles. *arXiv preprint arXiv:2204.05157*, 2022.

[38] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan. Robust optimization for fairness with noisy protected groups, 2020.

[39] L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[40] D. Xu, W. Du, and X. Wu. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1924–1932, New York, NY, USA, 2021. Association for Computing Machinery.

[41] D. Xu, W. Du, and X. Wu. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1924–1932, 2021.

[42] D. Xu, S. Yuan, and X. Wu. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 594–599, 2019.

[43] Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon. A theory of usable information under computational constraints. *ICLR*, 2020.

[44] I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

[45] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: regression analysis under differential privacy. *VLDB*, 2012.

A Algorithms

Algorithm 2 FairFM for Logistic Regression

Input: Dataset D, privacy budget ϵ, number of steps T.
Output: Model's parameter $\theta^{(T)}$
Set $\Delta = \frac{d^2}{4} + 3d$
for $k \in \{1, \dots, K\}$ **do**
  for each $j \in \{0, 1, 2\}$ **do**
    Add noise: $\tilde{\lambda}_k^{(j)} = \lambda_k^{(j)} + Lap(\frac{\Delta}{\epsilon})$
  end for
end for
Initialize $\theta^{(0)}$ randomly
for $t \in [1 : T]$ **do**
  $\theta_1^{(t-1)} = \cdots = \theta_K^{(t-1)} = \theta^{(t-1)}$
  for $k \in \{1, \dots, K\}$ **do**
    Compute gradient: Set $\nabla_k = 2\tilde{\lambda}_k^{(2)}\theta_k^{(t-1)} + \tilde{\lambda}_k^{(1)}$
    Update: $\theta_k^{(t)} = \theta_k^{(t-1)} - \eta\nabla_k$
  end for
  $\theta^{(t)} = \frac{\theta_1^{(t)} + \cdots + \theta_K^{(t)}}{K}$
end for
return $\theta^{(T)}$

B Proof of Theorem 3.1

Proof. For an updating process of a particular group k, by clipping the gradient, the clipping bound C is the maximum impact a single data point in the dataset D_k could have on one updating step t. Therefore, by adding Gaussian noise scaled by C, we achieve DP in one updating step. The model parameter fusing $\theta^{(t)} = \frac{\theta_1^{(t)} + \cdots + \theta_K^{(t)}}{K}$ does not introduce any extra privacy risk at each updating step t following the post-processing property in DP [9]. We use the moment accountant [1] to calculate the privacy loss for each dataset D_k after T updating steps given the sampling probability q, the broken probability δ, and the noise scale σ. Finally, since the datasets $\{D_k\}_{k=1}^K$ are disjoint ($D_a \cap D_b = \varnothing, \forall a \neq b \in [1, K]$), by the parallel composition theorem [22], we achieve (ϵ, δ)-DP for the whole dataset D where ϵ is calculated by the moment accountant. □

C Proof of Theorem 3.2

Proof. Recall the considered general fairness metrics:

$$\texttt{Fair}(h_\theta) = \max_{u,v\in[K]} |Pr(\hat{y} = 1|a = u, e) - Pr(\hat{y} = 1|a = v, e)| \tag{11}$$

$$= \max_{u,v\in[K]} \Big| \int_x Pr(\hat{y} = 1|x)Pr(x|a = u, e)dx - \int_x Pr(\hat{y} = 1|x)Pr(x|a = v, e)dx \Big| \tag{12}$$

$$\leq \max_{u,v\in[K]} \Big| \max Pr(\hat{y} = 1|x) \int_x Pr(x|a = u, e)dx - \min Pr(\hat{y} = 1|x) \int_x Pr(x|a = v, e)dx \Big| \tag{13}$$

$$= |\max Pr(\hat{y} = 1|x) - \min Pr(\hat{y} = 1|x)| \tag{14}$$

By Eq. (7) and the monotonicity of error function, we have

$$\texttt{Fair}(h_\theta) \leq \texttt{erf}\left(\frac{\|W^{(t-1)} - \eta\mu\|_2}{\eta\sigma C\sqrt{\frac{2}{K}}}\right) \leq \texttt{erf}\left(\frac{\|W^{(t-1)}\|_2 + \eta\|\mu\|_2}{\eta\sigma C\sqrt{\frac{2}{K}}}\right) \tag{15}$$

Furthermore, by the clipping process in lines 4 and 9 of Algorithm 1, we have

$$\|W^{(t-1)}\|_2 \leq \frac{1}{K}\sum_{k=1}^{K}\|W_{L,k}^{(t-1)}\|_2 \leq \frac{1}{K}\sum_{k=1}^{K} M = M \tag{16}$$

$$\|\mu\|_2 \leq \frac{1}{K}\sum_{k=1}^{K}\|\mu_k\|_2 \leq \frac{1}{K}\sum_{k=1}^{K}\sum_{i\in B_k}\|\bar{g}_i\|_2 \leq \frac{1}{K}\sum_{k=1}^{K}\sum_{i\in B_k} C = \frac{C}{K}\sum_{k=1}^{K}|B_k| = \frac{mC}{K} \tag{17}$$

Therefore,

$$\texttt{Fair}(h_\theta) \leq \texttt{erf}\Big(\frac{(MK+\eta mC)\sqrt{K}}{K\eta\sigma C\sqrt{2}}\Big) \tag{18}$$

which concludes the proof. □

D Related Works

Differential privacy has been extensively used in various deep learning applications [29, 30, 31, 21, 26, 28, 27]. Meanwhile, numerous efforts have been made to ensure various notions of group fairness through the use of in-processing constraints [10], mutual information [13], and adversarial training [43, 19, 18]. A topic of much recent discussion is the implication that DP models may inadvertently introduce or exacerbate biases and unfairness effects on the outputs of a model. For example, empirical and theoretical studies have shown that DP-SGD can magnify the difference in accuracy observed across various groups, resulting in larger negative impacts for underrepresented groups [2, 35]. These findings have led to the question of whether it is possible to create fair models while preserving sensitive information. They have spurred the development of various approaches such as those presented by [17, 24, 36, 35].

Despite the advancements made by these efforts, there remains a critical gap in ensuring group fairness. In particular, current methods have not been able to bound the effect of the private models on the model utility in various protected groups. To bridge this gap, this paper introduces a novel approach that establishes a connection between DP preservation and certified group fairness, thereby addressing this crucial challenge.

E Extending to multiple protected attributes

First, considering the scenario of $\mathcal{K}$ protected attributes, $\mathcal{A} \subset A_1 \times \cdots \times A_{\mathcal{K}}$ and in each $A_i, i \in [\mathcal{K}]$ there are K_i categories. To apply FairDP, users can divide the dataset D into $K = \prod_{i=1}^{\mathcal{K}} K_i$ disjoint datasets categorized by the combination between the protected attributes. In a particular dataset $D_i = \{x_j, \vec{a}_j, y_j\}_{j=1}^{n_i}, i \in [K]$, each data point $(x_j, \vec{a}_j, y_j)$ will have the protected attribute as $\vec{a}_j \in \mathcal{A}$ and $D_i \cap D_j = \varnothing, \forall i, j \in [K]$. For example, consider a dataset D with the protected attributes are gender with two categories (male and female) and race with five categories (Black, White, Asian, Hispanic, and Other); dataset D can be divided into groups with the combined attributes such as Black male, Black female, Hispanic male, Hispanic female, and so on. Then, users can apply FairDP with the new separation of groups.

F Supplemental results

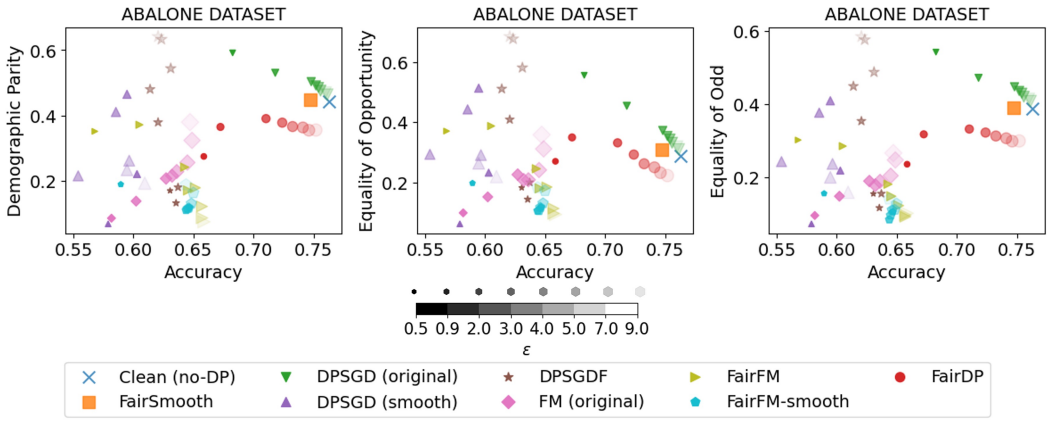

Figure 5: Trade-off among model performance, DP-preservation, and fairness on Abalone dataset.

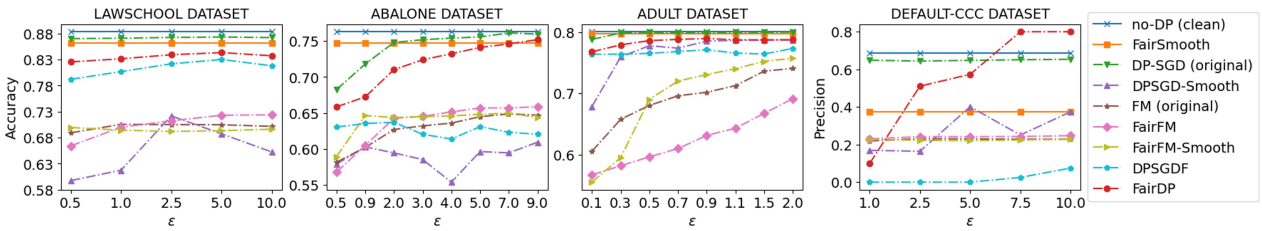

Figure 6: Model's performance across all datasets and baselines.

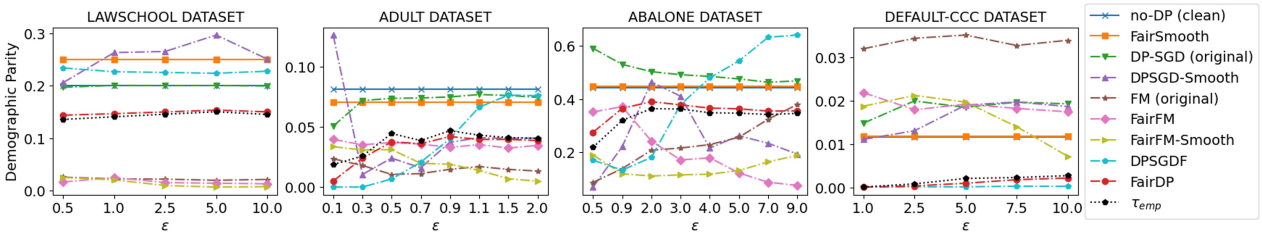

Figure 7: Demographic parity across all datasets and baselines.

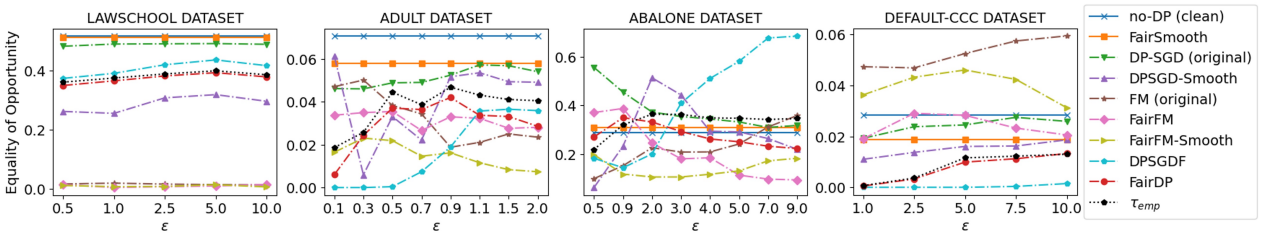

Figure 8: Equality of Opportunity across all datasets and baselines.

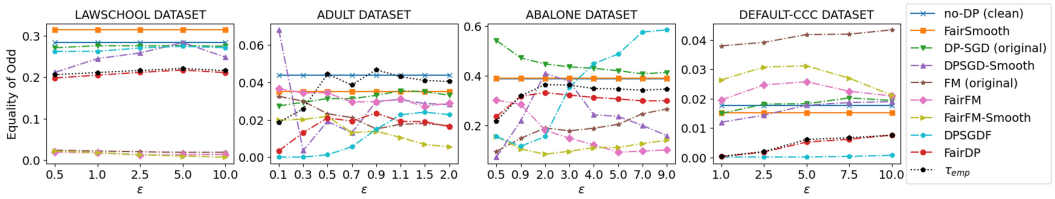

Figure 9: Equality of Odd across all datasets and baselines.

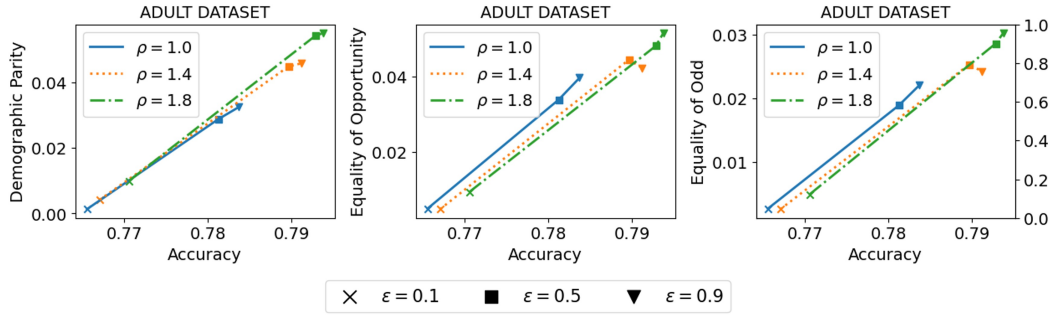

Figure 10: The trade-off among model utility, privacy, and fairness for the ratio ρ on Adult dataset.

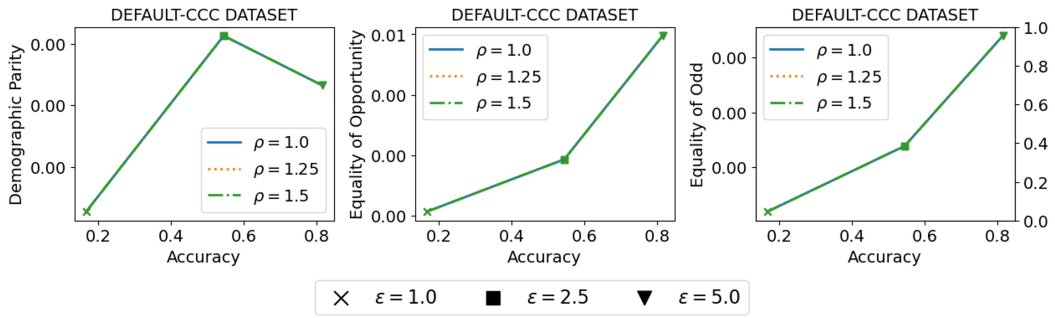

Figure 11: The trade-off among model utility, privacy and fairness for the ratio ρ on Default-CCC dataset.

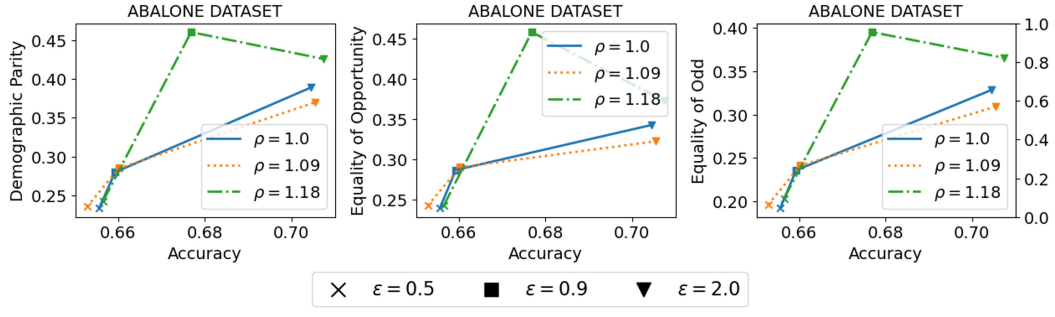

Figure 12: The trade-off among model utility, privacy and fairness for the ratio ρ on Abalone dataset.

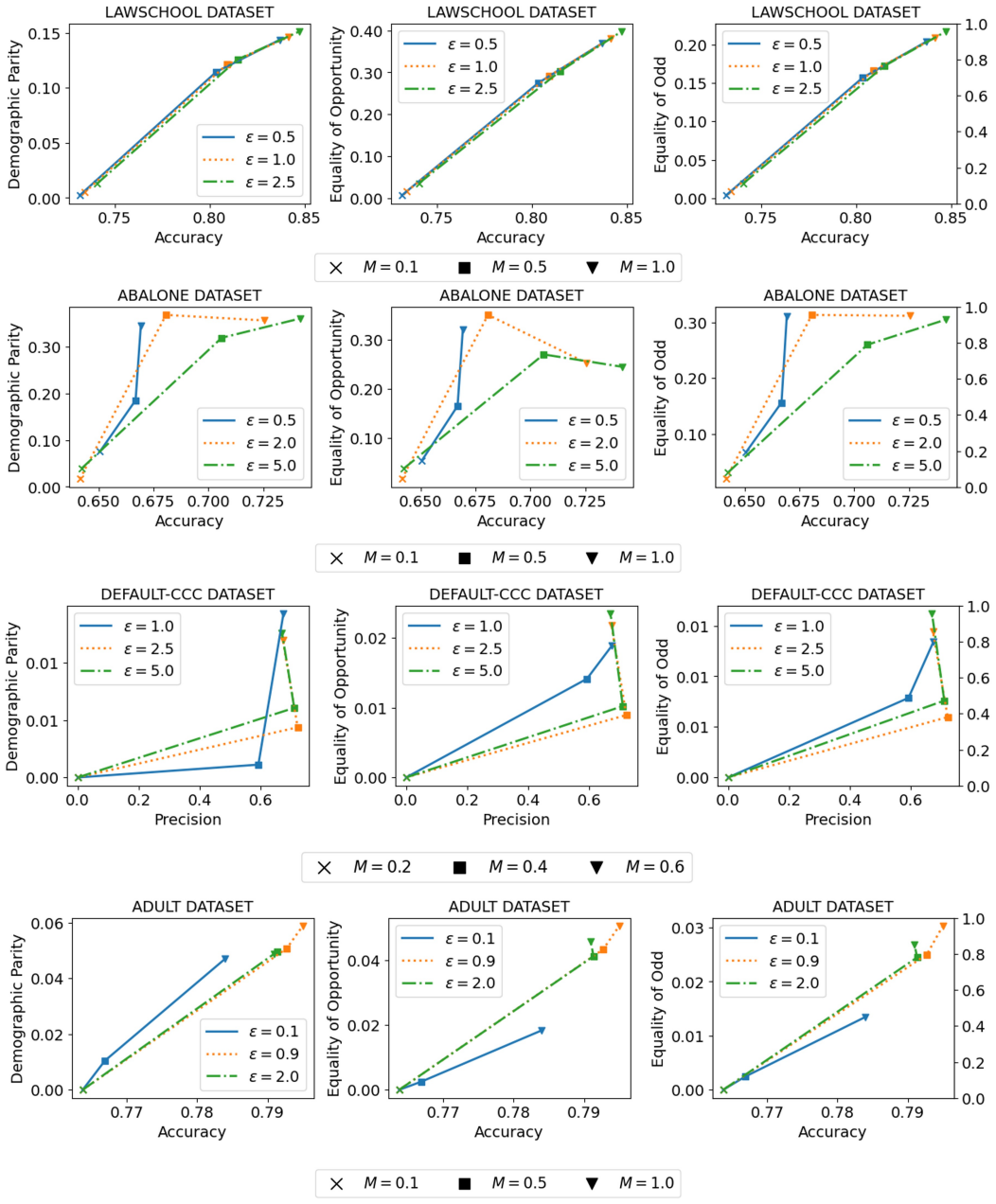

Figure 13: The trade-off among model utility, privacy and fairness for the clipping hyper-parameter M.