FairDP: Achieving Fairness Certification with Differential Privacy

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

Ferdinando Fioretto
University of Virginia
Charlottesville, Virginia, USA
fioretto@virginia.edu

Issa Khalil
Qatar Computing Research Institute (QCRI)
Doha, Qatar
ikhalil@hbku.edu.qa

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

Linh Thi Xuan Phan
University of Pennsylvania
Philadelphia, Pennsylvania, USA
linhphan@cis.upenn.edu

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

***Abstract*—This paper introduces FAIRDP, a novel training mechanism designed to provide group fairness certification for the trained model's decisions, along with a differential privacy (DP) guarantee to protect training data. The key idea of FAIRDP is to train models for distinct individual groups independently, add noise to each group's gradient for data privacy protection, and progressively integrate knowledge from group models to formulate a comprehensive model that balances privacy, utility, and fairness in downstream tasks. By doing so, FAIRDP ensures equal contribution from each group while gaining control over the amount of DP-preserving noise added to each group's contribution. To provide fairness certification, FAIRDP leverages the DP-preserving noise to statistically quantify and bound fairness metrics. An extensive theoretical and empirical analysis using benchmark datasets validates the efficacy of FAIRDP and improved trade-offs between model utility, privacy, and fairness compared with existing methods. Our empirical results indicate that FAIRDP can improve fairness metrics by more than 65% on average while attaining marginal utility drop (less than 4% on average) under a rigorous DP-preservation across benchmark datasets compared with existing baselines.**

***Index Terms*—differential privacy, fairness, machine learning**

I. Introduction

Machine learning (ML) systems are being increasingly adopted in decision processes that have a significant impact on people's lives, such as in healthcare, finance, and criminal justice [Angwin et al., 2022, Giovanola and Tiribelli, 2023]. This adoption also sparked concerns regarding how much information these systems disclose about individuals' data and how they handle bias and discrimination [Mehrabi et al., 2021, Pagano et al., 2023, Wan et al., 2023].

Differential privacy (DP) is an algorithmic property that allows the assessment and bounding of the leakage of sensitive individuals' information during computations. In the context of ML, it enables algorithms to learn from data while ensuring they do not retain sensitive information about any specific individual in the training data. However, directly applying a DP mechanism without careful calibration may aggravate the bias of the trained model's decision to a specific group of data compared to the non-DP ones, which results in unfairness for different groups of individuals [Bagdasaryan et al., 2019, Fioretto et al., 2022, Xu et al., 2021a] and incurs societal impacts for such individuals, particularly in areas including finance, criminal justice, or job-hiring [Tran et al., 2021d].

Balancing DP and group fairness while maintaining high model utility in ML systems has been the subject of much discussion in recent years. [Cummings et al., 2019] showed the existence of a trade-off between DP and equal opportunity, a fairness criterion that requires a classifier to have equal true positive rates for different groups. Different studies 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 [Sanyal et al., 2022]. 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 approaches [Jagielski et al., 2018, Mozannar et al., 2020, Tran et al., 2021a,c, 2023].

While these works have contributed to a deeper understanding of the trade-offs between DP, group fairness, and model utility, as well as the importance of addressing these issues in a unified manner, they all share a common limitation: *the inability to provide formal guarantees for DP and group fairness simultaneously while maintaining high model's utility*. The lack of a formal guarantee for these two critical aspects is essential and cannot be overstated. In many critical application contexts, such as those regulated by policy and laws [Act, 2009, Pardau, 2018, Team, 2017], these guarantees are often required, and failure to provide them can prevent adoption or deployment. For instance, the Fair Credit Reporting Act [Act, 2009] is a federal law that enforces to ensure the fairness and privacy of the information in consumer credit bureau files, which raises a concern in the finance industry around deploying and maintaining more advanced models into production [Das et al., 2021]. Conversely, a loose theoretical guarantee for fairness and privacy produces a random guess

* Corresponding Author

model, which is useless in practice.

This paper aims to address this gap by proposing a novel training mechanism that significantly improves group fairness with certificates while preserving DP without significantly degrading model utility. The key 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 FAIRDP, a novel DP training mechanism with certified fairness. FAIRDP remedies the disparate effects of DP-preserving noise on model fairness through group-wise clipping terms, enabling us to derive and tighten certified fairness bounds under DP protection. Throughout the training process, the mechanism progressively integrates knowledge from each group model, significantly improving the trade-off between model utility, privacy, and fairness with upper-bounded utility losses. Second, an extensive theoretical and empirical analysis shows that FAIRDP provides a better balance between privacy and fairness than existing baselines while maintaining high model utility, including both DP-preserving mechanisms with or without fairness constraints.

II. BACKGROUND

We consider 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 (set of) protected group attribute(s), and $y_i \in \mathcal{Y} = \{0, 1\}$ is a binary class label, similar to previous work [Celis et al., 2021, Jin et al., 2022]. 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. We also use $D_k = \{(x_i, a_i = k, y_i)\}_{i=1}^{n_k}$ to denote a non-overlapping partition over dataset D which contains exclusively the individuals belonging to a protected group k and $\cap_k D_k = \emptyset$.

To generalize for multiple protected group attributes, 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.

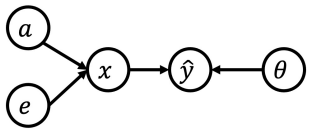

Fig. 1. Bayesian network of $h_\theta(x) = u_w(v_\phi(x))$.

We study models $h_\theta : \mathcal{X} \to [0, 1]$ parameterized by $\theta \in \mathbb{R}^r$ and the learning task optimizes the empirical loss function

$$\mathcal{L}(D) = \min_\theta \sum_{(x_i, a_i, y_i) \in D} \ell\left(h_\theta(x_i), y_i\right), \tag{1}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a differentiable loss function.

We use 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. Without loss of generality, consider h_θ as a combination of a feature extractor v_ϕ and a scoring function u_w, where ϕ and w are their corresponding parameters. Specifically, v_ϕ takes x as input and outputs an embedding vector ξ, i.e., $\xi = v_\phi(x)$. Then, the scoring function u_w, which is a linear layer, takes the embedding ξ and outputs the score z for the prediction on x, denoted as follows: $z = u_w(\xi) = \langle w, \xi \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product between two vectors. In general, we can write the combination between u_w and v_ϕ in h_θ as follows: $h_\theta(x) = u_w(v_\phi(x))$, where $\theta = \{\phi, w\}$.

This model setting is general for many structures of modern ML models in classification tasks (e.g., Neural Network, CNN, LSTM, Transformer). Finally, let ϕ_k, w_k be the weights of the feature extractor and scoring function of the group k's model ($\theta_k = \{\phi_k, w_k\}$).

Biased Model. It is worth noting that the model h_θ only observes the non-protected attributes x as input, which is similar to many practical settings where the protected attribute is hidden due to privacy and/or fairness concerns [Chen et al., 2019, Onesimu et al., 2022]. Nevertheless, x is still correlated with the protected attributes a, resulting in **a potentially biased model** affected by the protected attributes a through feature x, which leads to unfair predictions as discussed and observed in real-world applications [Corbett-Davies and Goel, 2018, Datta et al., 2014, Hajian and Domingo-Ferrer, 2012]. Therefore, it is practical to assume that $\hat{y}$ is independent of the protected attribute a and the random event e that depends on group fairness metrics (considered below), given the non-protected attribute x. Under this assumption, the predicting process can be mapped by a Bayesian network illustrated in Fig. 1

Group Fairness and Guarantees. This paper considers a class of statistical group fairness metrics as follows:

Definition 1. *The group fairness of a mechanism $\mathcal{M}$ is quantified by*

$$\mathcal{F} = \max_{u,v \in [K]} [Pr(\hat{y} = 1 | a = u, e) - Pr(\hat{y} = 1 | a = v, e)], \tag{2}$$

where $\hat{y}$ is prediction and e is a random event.

The fairness notion in Eq. (2) captures several well-known group fairness metrics, including demographic parity [Mehrabi et al., 2021] (when $e = \emptyset$), equality of opportunity [Hardt et al., 2016b] (when e is the event "$y = 1$"), and equality of odd [Hardt et al., 2016b] (when $e = y$). When $\mathcal{F} = 0$, the mechanism $\mathcal{M}$ are said to satisfy *perfect fairness* [Williamson and Menon, 2019]. However, perfect fairness cannot be achieved with DP preservation [Cummings et al., 2019]. Therefore, we focus on achieving *approximated fairness*, which allows the fairness metrics to be within a *"meaningful range."* In addition, if a mechanism satisfies $\mathcal{F} \leq \tau$ for $\tau \in [0, 1]$, then we say it *achieves certification of τ-fairness*. Intuitively, as τ decreases, the model's decision becomes more independent of the protected attribute with respect to different fairness metrics reflected through the random event e.

Differential Privacy [Dwork et al., 2014]. Differential privacy (DP) 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, denoted as $D \sim D'$.

Definition 2 (DP)**.** *A mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-DP, if, for any two adjacent inputs $D \sim D' \in \mathcal{D}$, and any subset of outputs $R \subseteq \mathcal{R}$:*

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

The parameter $\epsilon > 0$ describes the *privacy loss* of the algorithm, with smaller values denoting stronger privacy, and the parameter $\delta \in [0, 1)$ is the probability of violating ϵ-DP.

DPSGD [Abadi and et al., 2016]. DPSGD is a well-known DP-preserving algorithm to train ML models. The algorithm of DPSGD is illustrated in Algorithm 3 (Appx. A). At each updating step t, DPSGD samples a batch of data using Poisson sampling with a sampling probability q. Then, the l_2-norm of the gradient derived from each data point in the batch is clipped by a predefined upper-bound C (Line 6). The *DP-preserving Gaussian* noise with a scale σ, i.e., $\mathcal{N}(0, C^2\sigma^2\mathbf{I})$, is added to the sum of clipped gradients $\Delta\bar{g}_k$ from all data points, achieving $(q\epsilon, q\delta)$-DP at each step. DPSGD calculates the privacy loss after T steps using a Moment Accountant to track the moment of privacy loss distribution and bound the privacy budget accumulation, given q, δ, and σ.

III. Related Works

Differential privacy has been extensively used in various deep learning applications [McMahan et al., 2018, Papernot et al., 2018, Phan et al., 2020, 2016, 2017a,b, 2019]. Meanwhile, numerous efforts have been made to ensure various notions of group fairness through the use of in-processing constraints [Feldman et al., 2015], mutual information [Gupta et al., 2021], and adversarial training [Jin et al., 2022, Jovanović et al., 2022, Xu et al., 2020]. A topic of much recent discussion is the implication that DP models may inadvertently introduce or exacerbate biases and unfairness

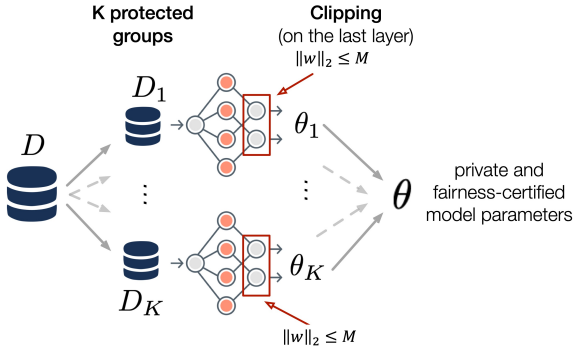

Fig. 2. A schematic overview of FairDP.

effects on the outputs of a model. For example, empirical and theoretical studies have shown that DPSGD can magnify the difference in accuracy observed across various groups, resulting in larger negative impacts for underrepresented groups [Alghamdi, 2023, Bagdasaryan et al., 2019, Islam et al., 2023, Tran et al., 2021a]. 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 and frameworks such as those presented by [Islam et al., 2023, Jagielski et al., 2018, Mozannar et al., 2020, Rafi et al., 2024, Tran et al., 2021a,c].

Despite the advancements made by these efforts, there is limited work addressing the 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. For instance, [Mangold et al., 2023] provides a guarantee that DP mechanisms have bounded disparate impact compared to the non-DP algorithms. However, it is not sufficient to establish a unified understanding of the correlation between differential privacy and group fairness. Similarly, [Makhlouf et al., 2024] studies the impact of local differential privacy on fairness but is unable to establish a unified understanding of the correlation between differential privacy and group fairness.

To bridge this gap, this paper introduces FairDP, a novel approach to establish a connection between DP preservation and certified group fairness, thereby addressing this crucial challenge. Unlike previous works, FairDP leverages the DP-preserving Gaussian noise added into the gradients of the training process to theoretically provide an upper and a lower bound for the probability that a data point is positively predicted. Then, it introduces a Monte Carlo sampling process under a rigorous DP preservation to approximate the bounds at the inference time, resulting in a tight bound on the group fairness of the model's decisions at the inference time.

IV. Certified Fairness with DP (FairDP)

This section introduces FairDP, a novel training mechanism satisfying three key objectives: **(1)** *Privacy*: the model satisfies (ϵ, δ)-DP; **(2)** *Fairness*: the decisions of released models are unbiased towards any protected group, with theoretical

Algorithm 1 FAIRDP Training

1: **Input**: Dataset D, sampling rate q, noise scale σ, norm bounds C and M, number of steps T, learning rate η, loss function ℓ.
2: Initialize $\theta^0 = \{\phi^0, w^0\}$.
3: **for** $t \in [1 : T-1]$ **do**
4: **Clip weights**: $w^{t-1} := w^{t-1} \min(1, M/\|w^{t-1}\|_2)$
5: $\theta_1^{t-1} = \cdots = \theta_K^{t-1} = \theta^{t-1}$ # model propagation
6: **for** $k \in \{1, \ldots, K\}$ **do**
7: Sample B_k^t from D_k with sampling probability q.
8: **Compute gradient**: For $x_i \in B_k^t$, $g_i^t = \nabla_{\theta_k^{t-1}} \ell(x_i)$
9: **Clip gradient**: $\bar{g}_i^t = g_i^t \min(1, \frac{C}{\|g_i^t\|_2})$
10: **Compute total gradient**: $\Delta_k = \sum_{i \in B_k} \bar{g}_i^t$
11: **Add noise**: $\tilde{\Delta}_k = \Delta_k + \mathcal{N}(0, C^2\sigma^2\mathbf{I})$
12: **Update**: $\theta_k^t = \theta_k^{t-1} - \frac{\eta}{|B_k^t|}\tilde{\Delta}_k$
13: **end for**
14: $\theta^t = (\theta_1^t + \cdots + \theta_K^t)/K$ # aggregation of groups' models
15: **end for**
16:
17: /* DP-MC for Fairness Certification */
18: **for** $k \in \{1, \ldots, K\}$ **do**
19: Sample B_k from D_k with sampling probability q.
20: $\{\theta_{k,j}^T\}_{j=1}^N \leftarrow$ DP-MC$(B_k, \eta_T, \sigma, C, M, N)$
21: **end for**
22: $\theta_j^T = \frac{1}{K}\sum_{k=1}^K \theta_{k,j}^T, \forall j \in [N]$
23: **Return** θ^T

Algorithm 2 DP-MC

1: **Input**: Set of clipped gradient $\{g_i | i \in B_k^T\}, \forall k$; learning rate η; noise scale σ; norm bounds C; number of model N; updated $\phi_k^T, \forall k$.
2: **for** $k \in [K]$ **do**
3: Update ϕ_k^T with the associated DP-preserving gradients.
4: Assemble $\Omega_k = \left\{\nabla_{w^{T-1}} \ell(x_i) \min\left(1, \frac{C}{\|g_i\|_2}\right)\right\}$ for $x_i \in B_k^T$
5: Partition Ω_k to N micro-batches $\{G_{k,j}\}_{j=1}^N$ such that $\cup_{j=1}^N G_{k,j} = \Omega_k$ and $G_{k,j} \cap G_{k,j'} = \varnothing, \forall j \neq j'$
6: **for** $j \in \{1, \ldots, N\}$ **do**
7: $\Delta_{k,j} = \sum_{i \in G_j} \nabla_{w^{T-1}} \ell(x_i) \min\left(1, \frac{C}{\|g_i\|_2}\right)$
8: $\tilde{\Delta}_{k,j} = \Delta_{k,j} + \mathcal{N}(0, C^2\sigma^2\mathbf{I})$
9: $w_{k,j}^T = w_{k,j}^{T-1} - \frac{\eta}{|G_j|}\tilde{\Delta}_{k,j}$
10: $\theta_{k,j}^T = [\phi_k^T, w_{k,j}^T]$
11: **end for**
12: **end for**
13: $\theta_j^T = \frac{1}{K}\sum_{k=1}^K \theta_{k,j}^T, \forall j \in [N]$
14: **Return** $\{\theta_j^T\}_{j=1}^N$

τ-fairness certification; and **(3)** *Utility*: the models achieve high utility for downstream ML tasks. Achieving these goals is challenging due to the intricate disparate impact incurred by DP-preserving noise and the degradation in the model utility when prioritizing fairness and privacy, particularly without a carefully calibrated noise injection.

To overcome these challenges, FAIRDP relies on two key components, including **(1)** fairness-aware DP training and **(2)** Monte Carlo approximation for fairness certification, with a schematic illustration of our training process in Fig. 2 and its pseudo-code in Algorithm 1. We specifically describe the proposed method and provide the privacy and fairness guarantee in the following sections.

A. Fairness-aware DP Training

From the training step $t = 1$ to the step $T - 1$ (Lines 4-14), FAIRDP overcomes a fundamental problem in sampling batches in DPSGD [Abadi and et al., 2016], which might create a disproportionate number of data points from different groups in the sampled batch, causing disparate contributions from each protected group to the DP-preserving model h_θ. To address this problem, FAIRDP ensures that every group will contribute to the learning process at any updating steps and upper-bounds the contribution from each group in expectation under privacy protection. This approach remedies the bias of the model's decision toward a group.

To do so, FAIRDP leverages DPSGD to train a set of group-specific models $\{h_{\theta_k}\}_{k=1}^K$, where each θ_k is independently learned to minimize the loss $\mathcal{L}(D_k)$ of the group k under DP (Lines 8-12). Gradient clipping bounds the l_2-norm of the average gradient in each group by C (Line 9); as a result, upper-bounding the contribution from each group in expectation. Also, FAIRDP clips the l_2-norm of the weights of the scoring function $w^{t-1} \in \theta^{t-1}$ by M to narrow the decision boundary in a bounded space (Line 4), which is essential to derive and tighten the fairness certification.

Given the set of group-specific models $\{h_{\theta_k}\}_{k=1}^K$ at each training step, FAIRDP aggregates groups' contributions enabling knowledge distillation from every group to better generalize the model h_θ (Line 13). Finally, the aggregated model parameters θ^t are propagated as the parameters for every group model in the next training round (Line 5). These aggregation and propagation steps ensure that the general model parameters θ^t are close to the parameters of every group, simultaneously reducing bias towards any specific group and distilling knowledge from every group to improve the model's utility.

B. DP Monte Carlo & Ensemble Inference

To provide fairness certification at the inference time, we innovate FAIRDP by executing a DP-preserving Monte Carlo (DP-MC) sampling process at the last step $t = T$ (Lines 16-20) by Algorithm 2. The DP-MC samples the DP-preserving noise, which is injected into the gradients of the scoring function to incorporate the randomness of DP-preserving noise into the prediction at inference time. To avoid extra privacy costs and obfuscating the correlation between DP and fairness, i.e., the impact of DP on fairness and vice versa[1], our DP-MC process produces a set of N (general) models $\{h_{\theta_j^T}\}_{j \in [N]}$ by partitioning the batch B_k^T of each group k into N *"disjoint"* micro-batches $\{B_{k,j}^T\}_{j \in [N]}$. For all micro-batches $j \in [N]$, the DP-MC process updates the weights $w_{k,j}$ of the scoring function with DP-preserving gradients $\tilde{\Delta}_{k,j}$ derived from the micro-batch $B_{k,j}^T$. Then, we generate the set of N (general) models $\{h_{\theta_j^T}\}_{j \in [N]}$ by aggregating these weights $\{w_{k,j}\}_{k \in [K], j \in [N]}$ in a group-wise approach, as follows:

$$\forall j \in [N] : w_j^T = \frac{1}{K}\sum_{k=1}^K w_{k,j}^T \quad \text{and} \quad \theta_j^T = \{\phi^T, w_j^T\}. \quad (3)$$

At the inference time, the prediction on a testing data point x will be the ensemble from the scores of the N

[1] Adding multiple noise to the same gradient will incur privacy accumulation [Dwork et al., 2014]; and adding separate noise to the DP-preserving gradient as smoothing will separate the impact of DP-preserving noise toward the model's decision, obfuscating the correlation between DP and fairness.

models $\{h_{\theta_j^T}\}_{j\in[N]}$. Each model will output a score z_j for x, and if the average score is greater than or equal to 0, i.e., $\frac{\sum_{j\in[N]} z_j}{N} \geq 0$, then x will be positively predicted, i.e., $\hat{y} = 1$; otherwise, $\hat{y} = 0$.

C. DP Guarantee and Fairness Certification

We provide the guarantee of DP and τ-fairness certification for the training process of FAIRDP (Algorithm 1) in the following theorems.

DP Guarantee. Similar to DPSGD, the DP-guarantee of FAIRDP is achieved by combining the gradient clipping step and the DP-preserving noise-injecting step. Furthermore, the model propagation and aggregation (Lines 5 and 13) and the weight clipping (Line 4) do not engage with dataset D; therefore, they are DP-preserving under the post-processing property of DP [Dwork et al., 2014]. Furthermore, the DP-MC separates the gradients at the last epoch to N disjoint sub-batches, which follow the parallel composition theorem [McSherry, 2009] and do not incur extra privacy risks. Finally, leveraging the Moment Accountant, we can compute the privacy budget ϵ accumulated throughout T steps.

Theorem 1. *Algorithm 1 satisfies (ϵ, δ)-DP where ϵ is calculated by the Moment Accountant [Abadi and et al., 2016] given the sampling probability q, T steps, and the noise scale σ.*

Proof. Considering one updating step t for an updating process of a particular group k, define the gradient extracting function $f(B_k)$ as follows:

$$f(B_k) = \sum_{i\in B_k} g_i \min\Big(1, \frac{C}{\|g_i\|_2}\Big), \text{ if } t \in [1, T-1] \tag{4}$$

$$f(B_k) = \Big[\sum_{i\in B_k} \nabla_\phi \ell(x_i) \min\Big(1, \frac{C}{\|g_i\|_2}\Big), \sum_{i\in G_j} \nabla_w \ell(x_i) \min\Big(1, \frac{C}{\|g_i\|_2}\Big)\Big]\Big\}_{j=1}^{N}, \text{ if } t = T \tag{5}$$

where $\{G_j\}_{j=1}^N$ is the disjoint partition of B_k.

For the intermediate step $t \in [1, T-1]$, by clipping the gradient, the l_2 sensitivity of the total gradient $f(B_k)$ is upper bounded by C. Similarly, for the last step $t = T$, for any pairs of neighboring datasets D and D', denote B_k and B_k as the batches sampled from D and D', respectively. In the worst-case, B_k and B'_k only different at one array of gradient $g_a = [\nabla_\phi \ell(x_a), \nabla_w \ell(x_a)]$ appears only in one and only one $G_j, j \in [N]$ since $G_i \cap G_i = \varnothing, \forall i, j \in [N]$. Thus, clipping the gradient also ensures the l_2 sensitivity of the total gradient $f(B_k)$ is upper-bound by C for the last step. Therefore, we achieve $(q\epsilon, q\delta)$-DP in one updating step by adding Gaussian noise scaled by C to $f(B_k)$ by the argument of Gaussian mechanism [Dwork et al., 2014]. 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 [Dwork et al., 2014]. We use the moment accountant [Abadi and et al., 2016] 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 [McSherry, 2009], we achieve (ϵ, δ)-DP for the whole dataset D where ϵ is calculated by the moment accountant. □

Fairness Certification. To derive a fairness certification, we leverage the DP-preserving noise injected to the parameter of the scoring function $z = u_w(\cdot)$ to bound the probability $Pr(\hat{y} = 1|x)$ in a range $[P_{lb}, P_{ub}]$, where P_{lb} and P_{ub} are lower and upper bounds respectively. We focus on the scoring function u_w because it is the decision boundary directly producing the prediction for a data point. Based on that, we can derive τ-fairness certification given the Bayesian network in Fig. 1, as follows:

$$\tau \leq P_{ub} - P_{lb}. \tag{6}$$

The full derivation of Eq. 6 is in the proof of Theorem 2. At step t, given a DP-preserving feature extractor v_{ϕ^t} yielding the *deterministic* process $\xi = v_{\phi^t}(x)$, the (DP-preserving) Gaussian noise added to the gradients w.r.t. the parameter w transforms the score z into a Gaussian distributed random variable given embedding ξ of an input x. Specifically, after injecting Gaussian noise into the gradient (Line 11), the model updating and the aggregating steps (Lines 12 and 13) are linear transformations of the Gaussian noise. Therefore, θ^t follows a multivariate Gaussian distribution $\mathcal{N}(w^{t-1} - \eta\mu^t; \sigma_0^2\mathbf{I})$, where m_k is the batch size of B_k, $\sigma_0^2 = \frac{\eta^2\sigma^2C^2}{K^2}\sum_{k=1}^K \frac{1}{m_k^2}$, and $\mu^t = \frac{1}{K}\sum_{k=1}^K \mu_k^t$ with μ_k^t is the total clipped gradient w.r.t. w^{t-1} incurred from B_k of group k. Also, since $z = \langle w^t, \xi\rangle$ is a linear combination of Gaussian random variables of w^t, the score z becomes a random variable: $z \sim \mathcal{N}(\langle w^{t-1} - \eta\mu^t, \xi\rangle; \|\xi\|_2^2\sigma_0^2)$.

Note that the weight and gradient clipping *do not affect* the fact that z is Gaussian distributed given ξ. It is because FAIRDP performs the clipping before injecting the Gaussian noise into the gradients. As a result, the probability $Pr(\hat{y} = 1|x) = Pr(z \geq 0|\xi)$ is an *integral* $\int_0^{+\infty} Pr(z|\xi)dz$ over the randomness of the DP-preserving noise added to w, which can be computed by a closed-form formula of cumulative distribution function of Gaussian distribution, as follows:

$$Pr(z \geq 0|\xi) = \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\langle w^{t-1} - \eta\mu^t, \xi\rangle}{\|\xi\|_2\sigma_0\sqrt{2}}\Big), \tag{7}$$

where $\texttt{erf}(\cdot)$ is the *error function*.

By the monotonicity of the error function[2], the property of the inner product[3], and the fact that the error function is an

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

[3]For vector a, b, we have $-\|a\|_2\|b\|_2 \leq \langle a, b\rangle \leq \|a\|_2\|b\|_2$.

odd function [4], we quantify P_{lb} and P_{ub} as follows:

$$P_{lb} = \frac{1}{2} - \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sigma_0\sqrt{2}}\Big), \tag{8}$$

$$P_{ub} = \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sigma_0\sqrt{2}}\Big). \tag{9}$$

Since the analysis is for a general step t, we can derive the fairness certification at the last updating step T. However, it is infeasible to compute the integral $\int_0^{+\infty} Pr(z|\xi)dz$ over the random variable of DP-preserving noise for testing data points since the model's parameters are given at inference time. Therefore, we innovate the DP-MC to approximate this integral at the inference time and incorporate the randomness of the DP-preserving noise through the ensemble prediction process. Nevertheless, the ensemble prediction significantly incurs the error in approximating the integral at the rate of $\mathcal{O}(\frac{1}{\sqrt{N}})$ as discussed in [Gelman et al., 1995].

Finally, it is worth noting that FAIRDP clips the weights w^{t-1} and gradients μ^t, i.e., $\|w^{t-1}\|_2 \leq M$ and $\|\mu^t\|_2 \leq C$, we derive τ-fairness certification in the following theorem:

Theorem 2. FAIRDP *satisfies τ- fairness certification, with*

$$\tau \leq \texttt{erf}\Big(\frac{MK + \eta C}{K\sigma_0\sqrt{2}}\Big) + \mathcal{O}(N^{-1/2}), \tag{10}$$

where $\sigma_0 = \frac{\eta\sigma C}{K}\sqrt{\sum_{k=1}^{K}\frac{1}{m_k^2}}$.

Proof. Recall the considered general fairness metrics:

$$\mathcal{F} = \max_{u,v\in[K]}[Pr(\hat{y} = 1|a = u, e) - Pr(\hat{y} = 1|a = v, e)]$$

Without loss of generality, assuming that $Pr(\hat{y} = 1|a = u, e) > Pr(\hat{y} = 1|a = v, e)$. In the case $Pr(\hat{y} = 1|a = u, e) < Pr(\hat{y} = 1|a = v, e)$, we just need to switch the roles of u and v. Let $\alpha_k \sim \mathcal{N}(0, \sigma^2C^2\mathbf{I})$ be the DP-preserving noise added to w_k for group k.

The high-level idea to derive a fairness certification is leveraging the DP-preserving noise added to the parameter of the decision-making function u_w to bound the probability $Pr(\hat{y} = 1|x)$ by the range $[P_{lb}, P_{ub}]$. Thus, Based on the Bayesian network in Fig. 1 which , we can derive τ-fairness certification, as follows:

$$\begin{aligned}
\tau &= \max_{u,v\in[K]}[Pr(\hat{y} = 1|a = u, e) - Pr(\hat{y} = 1|a = v, e)] \\
&\leq \max_{u,v\in[K]}\Big[\int_x Pr(\hat{y} = 1|x)Pr(x|a = u, e)dx \\
&\qquad\qquad - \int_x Pr(\hat{y} = 1|x)Pr(x|a = v, e)dx\Big] \\
&\leq \max_{u,v\in[K]}\Big[P_{ub}\int_x Pr(x|a = u, e)dx \\
&\qquad\qquad - P_{lb}\int_x Pr(x|a = v, e)dx\Big] \\
&= P_{ub} - P_{lb}
\end{aligned}$$

[4] $\forall x \in \mathbb{R} : \texttt{erf}(-x) = -\texttt{erf}(x)$

To find P_{ub} and P_{lb}, we notice that at an updating step t, given a DP-preserving updated feature extractor v_{ϕ^t} yielding the ***deterministic*** process $\xi = v_{\phi^t}(x)$, the Gaussian DP-preserving noise added to the gradients w.r.t the parameter w transforms z into a Gaussian distributed random variable given embedding ξ of the input x. Let us denote μ_k^t as the clipped gradient of w_k at the updating step t. Indeed, DP-preserving noise injected into clipped gradients $\Delta\bar{g}_k$ transforms the clipped gradients of the scoring function μ_k^t into a random variable following a multivariate Gaussian distribution $\mathcal{N}(\mu_k^t; \sigma^2C^2\mathbf{I})$. As a result, the parameter of the scoring function of group k at step t becomes a random variable with the following distribution $\mathcal{N}(w^{t-1} - \frac{\eta}{m_k}\mu_k^t; \frac{\eta^2}{m_k^2}\sigma^2C^2\mathbf{I})$ where $m_k = |B_k|$ is the batch size of group k, and $w_k^{t-1} = w^{t-1}$.

Furthermore, noticing that w^t of the (general) model is updated by a linear combination of the K multivariate Gaussian random variables $\{w_k^t\}_{k\in[K]}$. Therefore, the weight w^t follows a multivariate Gaussian distribution, as follows:

$$w^t \sim \mathcal{N}\Big(w^{t-1} - \frac{\eta}{K}\sum_{k=1}^{K}\frac{\mu_k^t}{m_k}; \frac{\eta^2\sigma^2C^2}{K^2}\sum_{k=1}^{K}\frac{1}{m_k^2}\mathbf{I}\Big).$$

Denoting $\mu^t = \frac{1}{K}\sum_{k=1}^{K}\mu_k^t$ and $\sigma_0^2 = \frac{\eta^2\sigma^2C^2}{K^2}\sum_{k=1}^{K}\frac{1}{m_k^2}$ for simpler notation. Since $z = \langle w^t, \xi\rangle$ is a linear combination of the Gaussian random variable, z is a Gaussian distributed random variable, as follows:

$$z \sim \mathcal{N}\Big(\langle w^{t-1} - \eta\mu^t, \xi\rangle; \|\xi\|_2^2\sigma_0^2\Big).$$

Moreover, under the ***deterministic*** process $\xi = v_{\phi^t}(x)$, $Pr(\hat{y} = 1|x) = Pr(\hat{y} = 1|\xi)$. As a result, given a data point x, it will be positively predicted with the probability $Pr(\hat{y} = 1|x) = Pr(z \geq 0|\xi) = 1 - Pr(z < 0|\xi)$, where the probability $Pr(z < 0|\xi)$ is an ***integral*** $\int_{-\infty}^{0} Pr(z|\xi)dz$ over the randomness of the DP-preserving noise added to w, which a closed-form formula of cumulative distribution function of Gaussian distribution can compute:

$$Pr(z < 0|\xi) = \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{-\langle w^{t-1} - \eta\mu^t, \xi\rangle}{\|\xi\|_2\sigma_0\sqrt{2}}\Big)$$

where $\texttt{erf}(\cdot)$ is the *error function*. It is worth noting that the error function is an odd function [Andrews, 1998, Yang, 2016], i.e., $\texttt{erf}(-x) = -\texttt{erf}(x)$ Thus, we have:

$$\begin{aligned}
Pr(\hat{y} = 1|x) &= Pr(z \geq 0|\xi) = 1 - Pr(z < 0|\xi) \\
&= \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\langle w^{t-1} - \eta\mu^t, \xi\rangle}{\|\xi\|_2\sigma_0\sqrt{2}}\Big)
\end{aligned} \tag{11}$$

Since $\langle w^{t-1} - \eta\mu, \xi\rangle = \|w^{t-1} - \eta\mu\|_2\|\xi\|_2\cos\varphi$, with φ being the angle between vectors $(w^{t-1} - \eta\mu)$ and ξ, $Pr(\hat{y} = 1|x)$ can be computed by:

$$Pr(\hat{y} = 1|x) = \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2\cos\varphi}{\sigma_0\sqrt{2}}\Big) \tag{12}$$

From Eq. (12), $\cos(\varphi) \in [-1, 1]$, based on the monotonicity of the error function and the fact that it is an odd function, we have that:

$$\frac{1}{2} - \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sigma_0\sqrt{2}}\Big) \leq Pr(\hat{y} = 1|x)$$
$$Pr(\hat{y} = 1|x) \leq \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2\|\xi\|_2}{\|\xi\|_2\sigma_0\sqrt{2}}\Big)$$

Therefore, we have that:

$$P_{lb} = \frac{1}{2} - \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sigma_0\sqrt{2}}\Big) \tag{13}$$
$$P_{ub} = \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sigma_0\sqrt{2}}\Big) \tag{14}$$

Now, we can leverage Eq. (13) and Eq. (14) to indicate that:

$$\begin{aligned} Pr(\hat{y} = 1|a = k, e) &\leq \mathbb{E}_{x|a,e}\Big[\frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sqrt{2}\sigma_0}\Big)\Big] \\ &= \frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sqrt{2}\sigma_0}\Big) \\ Pr(\hat{y} = 1|a = k, e) &\geq \mathbb{E}_{x|a,e}\Big[\frac{1}{2} - \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sqrt{2}\sigma_0}\Big)\Big] \\ &= \frac{1}{2} - \frac{1}{2}\texttt{erf}\Big(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sqrt{2}\sigma_0}\Big) \end{aligned}$$

As a result, we have:

$$\begin{aligned} \mathcal{F} &= \max_{u,v\in[K]}[Pr(\hat{y} = 1|a = u, e) - Pr(\hat{y} = 1|a = v, e)] \\ &\leq \texttt{erf}\left(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sqrt{2}\sigma_0}\right) \end{aligned}$$

By the monotonicity of the error function, we have

$$\mathcal{F} \leq \texttt{erf}\left(\frac{\|w^{t-1} - \eta\mu^t\|_2}{\sqrt{2}\sigma_0}\right) \leq \texttt{erf}\left(\frac{\|w^{t-1}\|_2 + \eta\|\mu^t\|_2}{\sqrt{2}\sigma_0}\right)$$

Furthermore, by the clipping process in Lines 6 and 11 of Algorithm 1, we have

$$\begin{aligned} &\|w^{t-1}\|_2 \leq= M \\ &\|\mu^t\|_2 \leq \frac{1}{K}\sum_{k=1}^{K}\frac{\|\mu_k^t\|_2}{m_k} \leq \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i\in B_k}\|\bar{g}_i\|_2}{m_k} \\ &\qquad \leq \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i\in B_k} C}{m_k} = \frac{C}{K} \end{aligned}$$

Therefore, along with the MC approximation error, we have the worst-case fairness certification on the true data distribution of different groups and the DP-preserving noise distribution, as follows:

$$\mathcal{F} \leq \texttt{erf}\Big(\frac{(MK + \eta C)}{K\sigma_0\sqrt{2}}\Big) + \mathcal{O}(N^{-1/2})$$

which concludes the proof. □

Remark 1. *Theorem 2 provides an upper bound on the τ-fairness certification, revealing a novel insight into the trade-off among privacy, fairness, and utility. The upper bound of τ-fairness certification decreases as the DP-preserving noise scale σ increases. As a result, stronger privacy (larger σ) enhances fairness certification due to the increased randomness influencing the model's decisions. Our theoretical observation is consistent with previous studies [Pannekoek and Spigler, 2021, Xu et al., 2019].*

D. Tightening Fairness Certification

While an important result, larger batch sizes, and lower learning rates can result in a looser τ-fairness in Theorem 2. Furthermore, the bound in Theorem 2 is guaranteed in the worst-case scenario when the model is completely biased to a specific group (i.e., the model consistently predicting positive for a group and negative for other groups), which might not be tight for a given data domain. To overcome this issue, we derive an empirical fairness certification that substantially tightens the τ-fairness certification, enabling a better understanding of the privacy, fairness, and utility trade-offs. Specifically, noticing that the probability that a data point from a group k is positively predicted, conditioned on a random event e, can be quantified as follows:

$$Pr(\hat{y} = 1|k, e) = \int_x Pr(\hat{y} = 1|x)Pr(x|k, e)dx.$$

By leveraging Eq. (7), this probability can be quantified by the expectation $\mathbb{E}_{x\sim P(x|k,e)}\Big[\frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\langle w^{t-1} - \eta\mu^t, \xi\rangle}{\|\xi\|_2\sigma_0\sqrt{2}}\Big)\Big]$, taken over by the data distribution of group k. Assuming that the training data and the testing data at the inference time for any group k are from the same distribution (i.e., marginal or no distribution shift), which is practical by the stability of DP mechanisms [Kulynych et al., 2022], one can approximate this expectation using the training data and provide a bound within a confidence interval, as follows:

$$\hat{\mathbb{E}}_{k,e} = \frac{1}{2} + \frac{1}{2n_{k,e}}\sum_{x\in D_{k,e}}\texttt{erf}\Big(\frac{\langle w^{t-1} - \eta\mu, v_{\phi^t}(x)\rangle}{\|v_{\phi^t}(x)\|_2\sigma_0\sqrt{2}}\Big),$$

where $D_{k,e}$ is the subset of D_k satisfying the random event e, and $n_{k,e}$ is the size of $D_{k,e}$. For instance, $D_{k,e} = D_k$ 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 can be computed by $\max_{u,v\in[K]}[\hat{\mathbb{E}}_{u,e}^{ub} - \hat{\mathbb{E}}_{v,e}^{lb}]$, where $\hat{\mathbb{E}}_{u,e}^{ub}, \hat{\mathbb{E}}_{v,e}^{lb}$ are computed using a tail-bound (e.g., Hoeffding inequality [Hoeffding, 1994]) from the $\hat{\mathbb{E}}_{u,e}, \hat{\mathbb{E}}_{v,e}$ with a confidence interval α.

Proposition 3. *A model h_{θ^T} optimized by Algorithm 1 satisfies empirical τ_{emp}-fairness certification with $\tau_{emp} = \max_{u,v\in[K]}[\hat{\mathbb{E}}_{u,e}^{ub} - \hat{\mathbb{E}}_{v,e}^{lb}] + \mathcal{O}(N^{-1/2})$ with a broken probability $(1 - \alpha)$.*

The proof of Proposition 3 is provided in Appx. B. In our experiments, we use the Hoeffding inequality with $\alpha = 0.95$.

Remark 2. The Proposition 3 can be used as a signal for the service provider to monitor the training process. Specifically, the service provider can train the model with FAIRDP until

the empirical fairness certification is larger than a targeted threshold. Then, the service provider can halt the training process. In case the service providers publish the model with the fairness certification, they can leverage the Laplace mechanism [Dwork et al., 2014] to add Laplace noise $Lap(\frac{1}{\epsilon'})$ to $\hat{\mathbb{E}}_{k,e}$ with extra privacy budget ϵ' to provide protection for releasing the fairness certification.

V. Convergence Analysis and Utility Loss Bound

This section focuses on deriving a utility loss bound in terms of convergence analysis for FairDP to provide guidelines on applying the mechanism to downstream tasks. The key observation is that, given a fixed noise scale σ and a careful decay of the learning rate η, FairDP will converge to the global minima for a convex and β-Lipschitz empirical loss function $\mathcal{L}(D)$ with the rate of $\mathcal{O}(\log T/\sqrt{T})$. To derive such guarantees, we consider the two following assumptions:

Assumption 4. *$\mathcal{L}(D,\theta)$ is a convex function with respect to θ, i.e., $\forall\theta,\theta' : \mathcal{L}(D,\theta) - \mathcal{L}(D,\theta') \leq \langle\nabla_\theta\mathcal{L}(D,\theta), \theta-\theta'\rangle$.*

Assumption 5. *$\mathcal{L}(D,\theta)$ is a β-Lipschitz function with respect to θ, i.e., $\forall\theta,\theta' : |\mathcal{L}(D,\theta) - \mathcal{L}(D,\theta')| \leq \beta\|\theta-\theta'\|_2$.*

These two assumptions are generally considered in previous works for private stochastic gradient descent [Bassily et al., 2014, 2019, Feldman et al., 2020]. Moreover, they are practically common for many ML models such as Linear Regression, Logistic Regression, and simple neural networks [Pilanci and Ergen, 2020].

Given these assumptions, we bound the expected empirical risk of a model h_{θ^T} trained by FairDP as follows:

Theorem 6. *Let θ^T be the output of Algorithm 1. If C is chosen such that $C \geq \beta$, $\mathcal{L}$ satisfies the considered assumptions, and the learning rate $\eta(t) = \mathcal{O}\Big(\frac{K}{\sqrt{t(\beta^2+K\sigma^2C^2r)}}\Big)$, then the excessive risk $\mathbb{E}[\mathcal{L}(D,\theta^T)] - \mathcal{L}(D,\theta^*)$ is bounded by:*

$$\mathbb{E}[\mathcal{L}(D,\theta^T)] - \mathcal{L}(D,\theta^*) \leq \mathcal{O}\Big(\frac{\sigma C\sqrt{r}\log(T)}{\sqrt{TK}}\Big) \tag{15}$$

where $\theta^* = \arg\min_\theta \mathcal{L}(D,\theta)$, r is the number of parameters in θ, and the expectation is over the randomness of FairDP.

Proof. Considering an updating step $t+1$ of FairDP

$$\forall i \in [K] : \theta_i^{t+1} = \theta^t - \eta_t\tilde{g}_i^t, \text{ where } \tilde{g}_i^t = \bar{g}_i^t + \sigma C\mathcal{N}(0,I_r)$$

$$\theta^{t+1} = \frac{1}{K}\sum_{i=1}^K \theta_i^{t+1} = \theta^t - \frac{\eta_t}{K}\sum_{i=1}^K \bar{g}_i^t + \frac{\eta_t}{\sqrt{K}}\sigma C\mathcal{N}(0,I_r)$$

Denote $\tilde{G}_t = \sum_{i=1}^K \bar{g}_i^t + \sqrt{K}\sigma C\mathcal{N}(0,I_r)$ and $\bar{G}_t = \sum_{i=1}^K \bar{g}_i^t$, then $\mathbb{E}_{\sim\mathcal{N}(0,I_r)}(\tilde{G}_t) = \bar{G}_t$. Furthermore, if C is chosen such that $C \geq \beta$, then $\mathbb{E}(\bar{G}_t) = \nabla_{\theta^t}\mathcal{L}(D,\theta^t)$ where the randomness is over the sampling process. Moreover, we can upper bound the following expectation:

$$\begin{aligned}
\mathbb{E}(\|\tilde{G}_t\|_2^2) &= \mathbb{E}(\|\bar{G}_t + \sqrt{K}\sigma C\mathcal{N}(0,I_r)\|_2^2) \\
&= \mathbb{E}(\|\bar{G}_t\|_2^2) + 2\mathbb{E}(\langle\tilde{G}_t, \sqrt{K}\sigma C\mathcal{N}(0,I_r)\rangle) \\
&\quad + K\sigma^2C^2\mathbb{E}(\|\mathcal{N}(0,I_r)\|_2^2) \\
&\leq m^2\beta^2 + K\sigma^2C^2r = G^2
\end{aligned} \tag{16}$$

Then, we can leverage the result from Theorem 2 from [Shamir and Zhang, 2013] which declares as follows:

Theorem 7 (Theorem 2 of [Shamir and Zhang, 2013])**.** *For a convex function $\mathcal{L}(\theta)$, let $\theta \in \Theta$ such that $\|\Theta\|_2 \leq Q$, $\theta^* = \arg\min_\theta \mathcal{L}(\theta)$, and θ^0 is an arbitrary point in Θ. Consider a stochastic gradient descent with $\mathbb{E}(\tilde{G}_t) = \nabla_\theta^t\mathcal{L}$ and the learning rate $\eta_t = \frac{Q}{G\sqrt{t}}$. Then for any $T > 1$, the following is true*

$$\mathbb{E}(\mathcal{L}(D,\theta^T)) - \mathcal{L}(D,\theta^*) \leq \mathcal{O}(\frac{QG\log(T)}{\sqrt{T}}) \tag{17}$$

Using the chosen form of η_t, the bound in Eq. (16) with Theorem 7, we can derive the guarantee in Theorem 6 which concludes the proof. □

Theorem 6 provides guidance on choosing the hyper-parameters C to optimize the convergence rate of FairDP. A larger value of the gradient clipping C will preserve the direction of the clean gradients but also increase the variance, which requires more updating steps to reach convergence. In addition, for simple ML models whose Lipschitz constant β can be calculated or approximated, the practitioners can choose $C = \beta$ to maintain the gradient's direction under the clipping process while avoiding excessive DP-preserving noise.

Practitioners can leverage our results to better balance the trade-offs among privacy, fairness, and utility by adaptively adjusting the training process of FairDP. For example, applying optimizers like Adam [Kingma and Ba, 2014] at the onset of training may enhance model utility and convergence rate under the same 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, such that the smaller M, the fairer the model is. However, small M could degrade model utility since it constrains the decision boundary in smaller parameter space.

To our knowledge, FairDP is the first mechanism that achieves τ-fairness certification while preserving DP, without undue sacrificing model utility, as demonstrated in experimental results below. Theorem 2 and Proposition 3 provide an insightful understanding of the interplay between privacy and fairness. A stronger privacy guarantee (larger noise scale σ) tends to result in better fairness certification (smaller τ). In addition, another application of Proposition 3 is to train a model achieving desirable privacy and fairness guarantees (ϵ,τ), which can be predefined by practitioners, by training the model until privacy and empirical fairness estimate aligns with the predetermined thresholds and halting the training if one of the guarantees is breached.

VI. Experimental Results

We conducted a comprehensive evaluation of FairDP and baseline methods on various benchmark datasets, primarily focusing on two aspects: **(1)** Assessing the accuracy and tightness of the fairness certification by comparing it with empirical results obtained from multiple statistical fairness metrics; **(2)** Examining the trade-off between model utility, privacy, and fairness; and **(3)** Exploring the contribution of each component of FairDP to the overall utility and fairness of the models through extensive ablation studies.

Datasets, Metrics, and Model Configurations. The evaluation uses three datasets: the Adult dataset [Dua and Graff, 2017], the Default of Credit Card Clients (Default-CCC) dataset [Yeh and Lien, 2009], and the UTK-Face Dataset (UTK) [Zhang et al., 2017]. Details of the datasets are in Table I. These are the benchmark datasets to evaluate the fairness-aware ML algorithm [Han et al., 2023, Jin et al., 2022, Tran et al., 2022]. Data preprocessing steps are strictly followed as outlined in previous works such as [Han et al., 2023, Iofinova et al., 2021, Ruoss et al., 2020, Tran et al., 2021b]. Since the datasets are extremely imbalanced (i.e., the number of positive data is much smaller than the number of negative data), we evaluate the model's utility by using *area under the ROC curve* (ROC-AUC) and Accuracy as in previous studies. A *higher* ROC-AUC (Accuracy) indicates *better* utility. We use *demographic parity* [Dwork et al., 2012], *equality of opportunity*, and *equality of odds* [Hardt et al., 2016a] as primary fairness metrics since they are the standard metrics for fairness measurement. The lower values of these fairness metrics indicate fairer decisions. The experiments use privacy budgets in the range of $[0.5, 2.0]$ and $\delta = 1e^{-5}$ for different datasets. Although DP is celebrated for using small values of ϵ, most current deployments report ϵ larger than 1, with many of them using $\epsilon > 5$ [Des].

In Adult and Default-CCC datasets, a multi-layer perceptron (MLP) is employed with ReLU activation on hidden layers and sigmoid activation on the last layer for binary classification tasks. For the UTK dataset, a simple Convolution Neural Network (CNN) is employed since it is an image-based dataset as considered in [Tran et al., 2022]. Adam optimizer [Kingma and Ba, 2014] is employed during the complete training process. For FairDP, we set the weight clipping hyper-parameter $M \in [0.1, 1.0]$ and initialize the learning rate $\eta = 0.007$ for the first half of the training process, and then reduce it to $\eta = 0.005$ for the rest of the process. Statistical tests used are two-tailed t-tests.

Baselines. We consider a variety of DP-preserving mechanisms, fairness training algorithms, and combinations of these as baselines, resulting in six baselines, including a non-DP mechanism, three existing mechanisms that either preserve DP or promote fairness, one existing mechanism that achieves both fairness and privacy, and one adapted mechanism that achieves both DP and fairness.

Established Baselines. We consider **DP-SGD** [Abadi and et al., 2016], **DPSGDF** [Xu et al., 2021b], **FairSmooth** [Jin et al., 2022], and **DP-IS-SGD** [Kulynych et al., 2022] as baselines. **DP-SGD** is a well-established DP mechanism with many applications in DP research. **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. **DP-IS-SGD** established the connection between DP and distributional generalization and reduced the accuracy disparity through important sampling during the training process. In addition, we also consider a combination of the baselines, **DPSGD-Smooth**, by applying **FairSmooth** to models trained by **DPSGD**. Since it is the only baseline offering both DP and fairness guarantees, we employ it for comparison against FairDP. We also acknowledge a previous work DP-FERMI [Lowy et al., 2023]; however, we do not consider it as a baseline in this study since the implementation of DP-FERMI has not been finalized per indicated by the authors.

TABLE I
Evaluation Datasets.

Dataset	Default-CCC	Adult	UTK
# data	30,000	48,842	23,795
# features	89	41	$48 \times 48 \times 1$
# positive	6,636	11,687	8,608
protected attribute	Gender	Gender	Race

A. Utility, Privacy, and Fairness Trade-offs

Fig. 3 and Fig. 4 show the results of each algorithm w.r.t. model's utility, fairness, and privacy. In these figures, the points indicate the average results of an experiment of FairDP and baselines with the corresponding privacy budget. The smaller and darker points indicate experiments with lower privacy budgets and vice-versa. In Fig. 3 and Fig. 4, points positioned closer to the bottom-right corner denote superior balance among model utilty (higher accuracy/precision), privacy (strict DP protection), and fairness (lower empirical values of fairness metrics).

Comparing with Baselines. In general, FairDP significantly attains fairer decisions while maintaining high utility across datasets and fairness metrics with different privacy budgets compared with the clean model and the best baseline DP-IS-SGD.

Specifically, in the Adult dataset, under rigorous privacy protection ($\epsilon = 0.5$), FairDP gains 75% improvement in demographic parity while attaining a modest decrease in ROC-AUC and Accuracy compared to the clean model (3% in ROC-AUC, 4.3% in Accuracy on average) average across different privacy budgets ($\epsilon \in [0.5, 2]$). Specifically, at $\epsilon = 0.5$, the demographic parity reduce from 0.079 in the clean model to 0.014 of FairDP with p-value $= 3.15e^{-5}$. Fig. 8 (Appx. C) illustrates the comparison between FairDP and the clean model

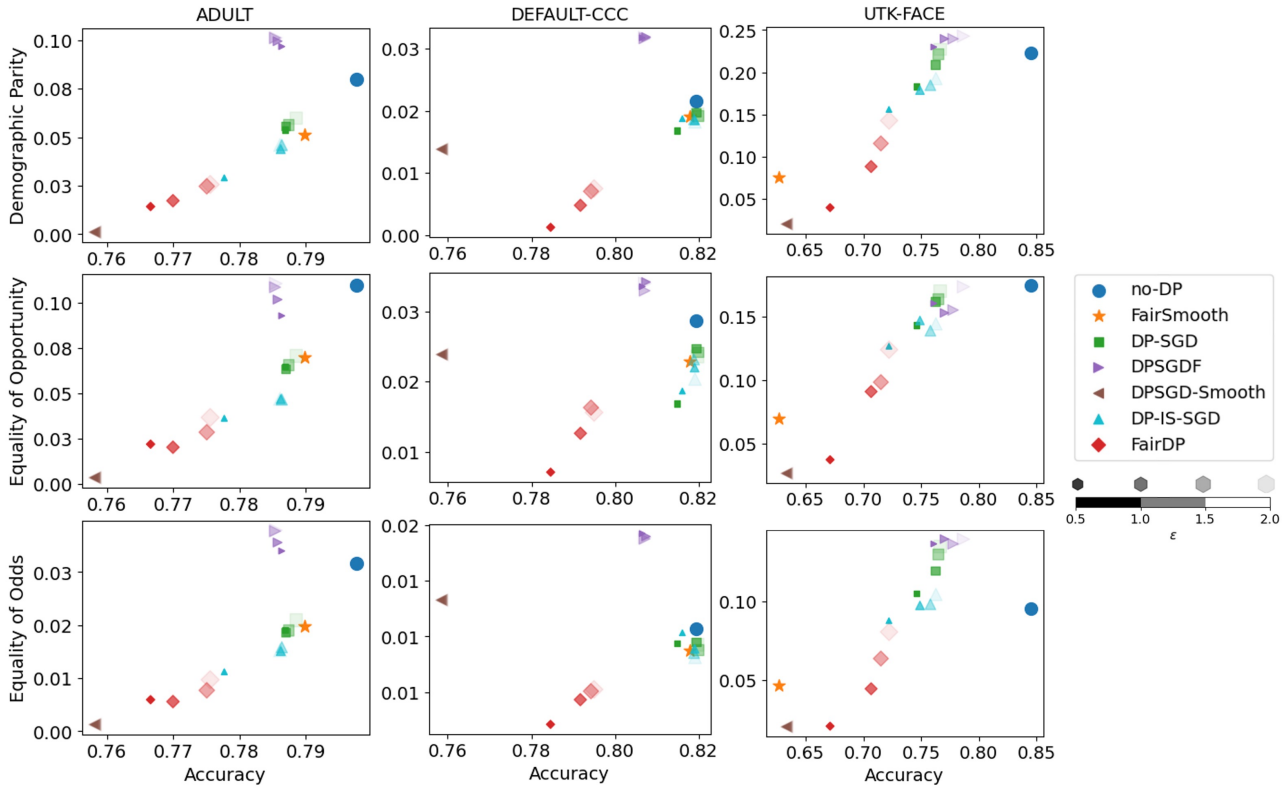

Fig. 3. Trade-off among model utility (Accuracy), DP-preservation, and fairness.

in terms of fairness gains compared to utility drops, which shows that FAIRDP achieves more than 60% improvement in terms of fairness while attaining marginal drop of utility (i.e., less than 5%) across different privacy budgets. Similar results between FAIRDP and the clean model are observed for other datasets and other fairness metrics with different privacy budgets.

Fig. 5 illustrates the comparison between FAIRDP and the best baseline (DP-IS-SGD) in terms of fairness gains and utility drops. At $\epsilon = 0.5$, FAIRDP achieves a significant gain across fairness metrics (over 45.83%) compared with DP-IS-SGD while attaining marginal utility drop (less than 2.64%) for the Adult dataset. Furthermore, FAIRDP attains 50.5% gain on average in terms of demographic parity for the Adult dataset across different privacy budgets. In addition, under looser DP protection (i.e., $\epsilon = 2.0$) where FAIRDP is less fair, FAIRDP still achieves a significant fairness gain (34%) compared with DP-IS-SGD while maintaining marginal drop of utility (less than 1.8%) on average across fairness and utility metrics. Similar results are observed in other datasets and other fairness metrics. In fact, the fairness gains in FAIRDP go up to 78.41% and 74.08% on average across fairness metrics in the Default-CCC and UTK-Face datasets given $\epsilon = 0.5$ correspondingly.

These results highlight the effectiveness of FAIRDP in addressing the trade-off among privacy, fairness, and utility for different ML tasks. *The promising results of* FAIRDP *can be attributed to its unique approach of controlling the contribution of each group to the learning process, the DP-preserving noise injected into each group, enforcing a constraint on the decision boundary, and aggregating the knowledge learned from every group at each training step.* FAIRDP *fundamentally differs from the baselines, leading to its superior performance.*

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. In the Adult dataset, DPSGDF is less fair than the clean model in terms of demographic parity (0.1 compared with 0.079 in demographic parity with $p = 3.84e^{-7}$, i.e., 25% increase). 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 a 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 performance for ML tasks.

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. 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*

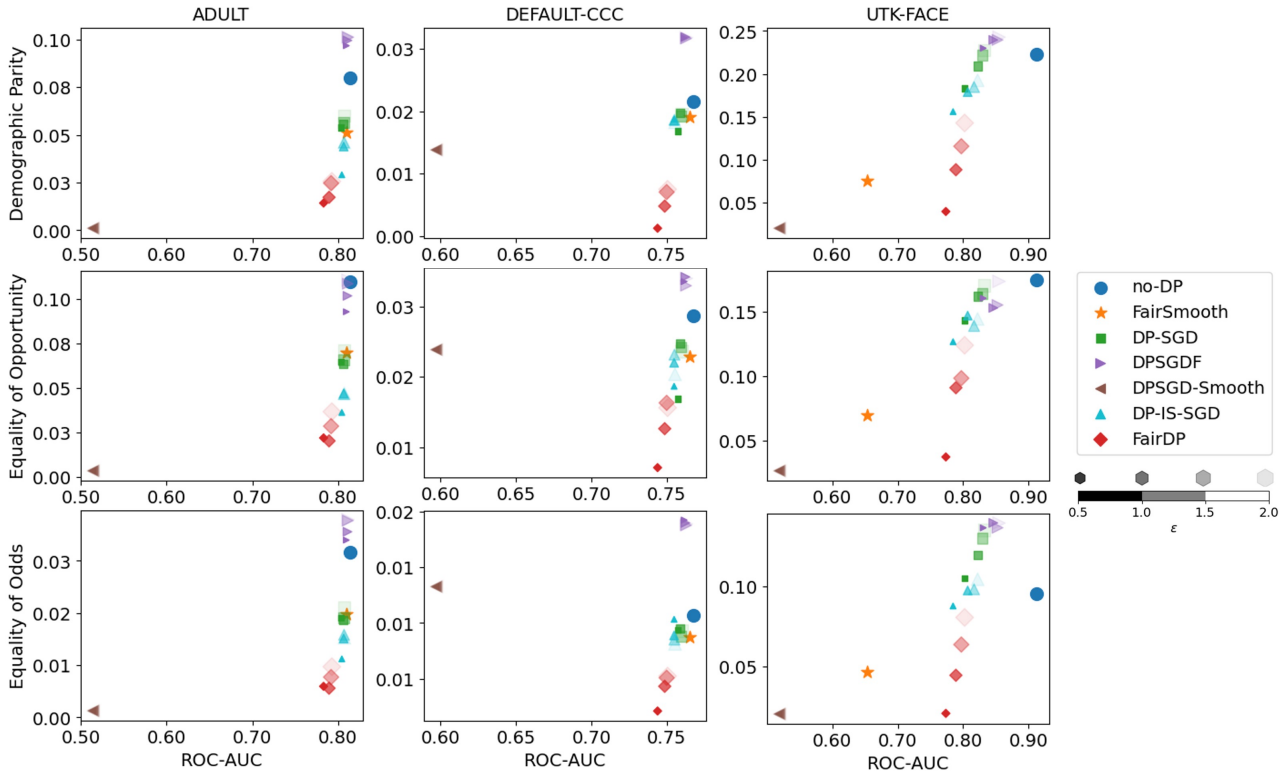

Fig. 4. Trade-off among model utility (ROC-AUC), DP-preservation, and fairness.

pioneering step in this direction.

Privacy and Fairness Correlation. Our observation from the results is that *privacy enhances fairness*. Specifically, the lower the value of the privacy budget ϵ (stronger privacy protection), the lower the fairness metric (Fig. 3) and the higher the fairness gain (Fig. 5). For instance, in the Adult dataset, the demographic parity gains 44% (i.e., reducing from 0.025 to 0.014) when the privacy budget decreases from $\epsilon = 2.0$ to $\epsilon = 0.5$. Similar behavior is observed for other datasets and other fairness metrics. This observation verifies the validity of Remark 1 and our fairness certification. *These insights highlight the contribution of* FairDP *in understanding the correlation between fairness and privacy.*

B. Fairness Certification under Data Distribution Shift

Tightness of Fairness Certification. Fig. 6 and 9 (Appx. C) show the empirical fairness results and the certification τ_{emp}. It is worth noting that we inject the Laplace noise with $\epsilon' = 0.1$ into the empirical fairness certification in our experiments. Although we make an assumption that the training data and the testing data are from the same distribution, we create a shift in the two distributions quantified by γ, which measures the total variation between the joint probability $Pr(a, y)$ in the training and testing datasets. To create the distribution shift, we modify the joint probability $Pr(a, y)$ as described in [An et al., 2022]. In general, the empirical results confirm the validity of our τ_{emp}-fairness certification across different datasets and privacy budgets since the empirical value of the fairness metrics ($\gamma = 0$) is lower than the certification. Furthermore, in most instances for the Adult and UTK-Face datasets, our τ_{emp}-fairness certifications are substantially lower than the empirical fairness values of the clean model. For instance, for demographic parity in the UTK-Face dataset, our empirical certifications are significantly smaller than the empirical fairness results of the clean model ($p = 2.14e^{-5}$) while maintaining a small gap with the empirical fairness results of FairDP. Moreover, across the different values of γ, we can observe that our τ_{emp}-fairness certification still holds under some magnitude distribution shift, which highlights the robustness of the certification in real-world scenarios. That shows the correctness and tightness of our τ_{emp}-fairness certification across datasets and DP budgets, strengthening the advantages of FairDP in theoretical guarantees and empirical results compared with the baselines. Note that τ_{emp} can be marginally larger than the empirical fairness results of the clean model for some instances. It is because our certification is tailored to DP preservation in FairDP instead of the (non-DP-preserving) clean model.

DP-MC Approximation Error. Table II shows the error of the DP-MC approximation when $N = 10$. In general, the Monte Carlo error is small across the datasets compared to the τ_{emp}-fairness certification. Specifically, in the Adult dataset, the error is $4.37e^{-5}$, which is 0.1% of the τ_{emp}-fairness certification for demographic parity across different privacy budgets. Similar results are observed for other datasets and fairness metrics, which highlights the practicality of FairDP when collaborating the τ-fairness certification with the DP-

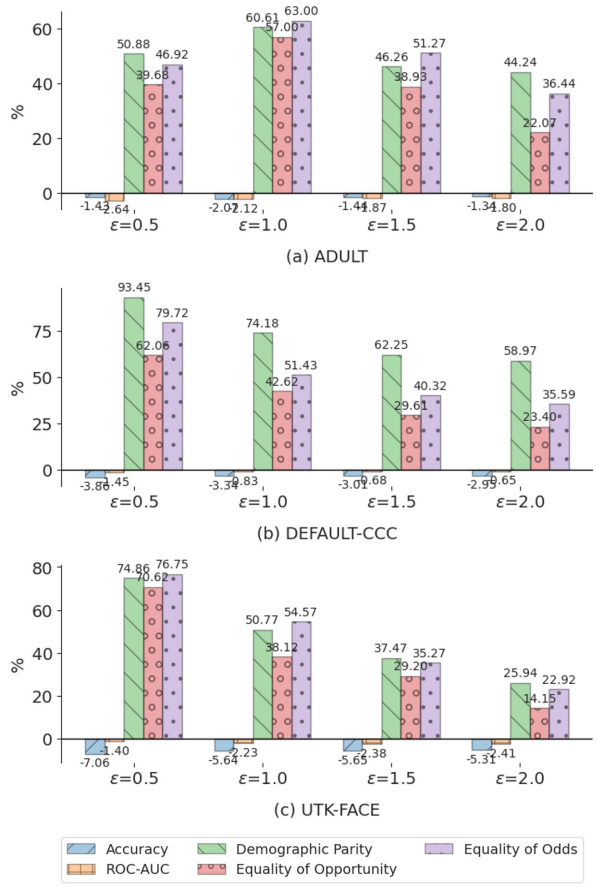

Fig. 5. Relative utility drop and fairness gain of FAIRDP compared with the best baseline - DP-IS-SGD.

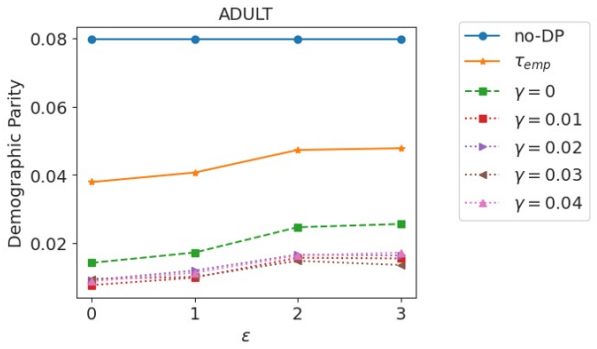

Fig. 6. Tightness of empirical fairness certification on demographic parity w.r.t different privacy budgets for the Adult dataset. The results for other fairness metrics and datasets are in Fig. 9, Appx. C

MC approximation, strengthening the advantages of FAIRDP in theoretical guarantees.

C. Ablation Study

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. Fig. 7 and Figs. 10-11 (Appx. C) illustrate the effect

TABLE II
MONTE CARLO APPROXIMATION ERROR OF FAIRNESS CERTIFICATION WITH $N = 10$ ACROSS DIFFERENT DATASETS.

Dataset	Demographic Parity	Equality of Opportunity	Equality of Odds
Adult	$4.37e^{-5}$	$3.16e^{-4}$	$3.7e^{-4}$
Default-CCC	$1.1e^{-4}$	$4.8e^{-4}$	$6.2e^{-4}$
UTK	$1.36e^{-4}$	$3.7e^{-4}$	$6.1e^{-4}$

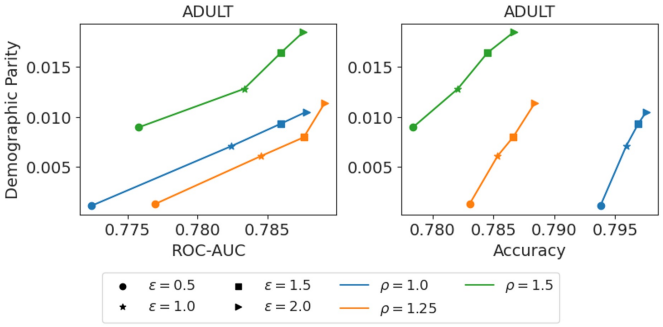

Fig. 7. Model utility, privacy, and demographic parity for various ρ values on the Adult dataset. The results for other fairness metrics and datasets are in Figs. 10-11, Appx. C

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 to the desired ρ. In general, increasing ρ values leads to more 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 guarantee remains applicable across various degrees of dataset imbalance. Lower privacy budgets (i.e., stronger privacy guarantees) contribute to improved fairness in the model's decisions, strengthening the theoretical certification of FAIRDP.

Hyper-parameter M. Table III (Appx. C) shows the effect of the hyper-parameter M toward the trade-off among privacy, fairness and utility. In general, reducing M leads to a fairer model's decision, as described in our fairness certification. However, reducing M will restrict the model's decision boundary, resulting in the degradation of model utility. In the Adult dataset, when M drops from 1.0 to 0.25, the ROC-AUC drops from 0.80 to 0.59 ($\approx 50\%$ reduction), and Accuracy drops from 0.80 to 0.75 ($\approx 4\%$ reduction), while the demographic parity is reduced from 0.027 to 0.0. Practitioners can tune this parameter to better balance the trade-off among privacy, fairness, and utility in FAIRDP.

VII. Discussion

FairDP is a robust framework for integrating differential privacy with group fairness. A fundamental limitation is its assumption that all groups have sufficient data for practical model training, which is often not the case in real-world scenarios where underrepresented groups pose challenges to fairness and utility. However, if data instances of some groups are too small, the practitioner can apply data augmentation techniques [Bao et al., 2024, He et al., 2024, Song et al., 2024] under privacy protection on top of FairDP to mitigate this limitation.

In addition, FairDP is that the framework is designed for a centralized setting, where the protected attributes are exposed to the server. In contexts such as hospitals, banks, or universities, the organization typically has access to its members' complete set of protected attributes, which it uses to make informed decisions about privacy and fairness. Nevertheless, applying FairDP to this organizational-level decision-making ensures that privacy and group fairness concerns are addressed in the context of the broader goals and policies of the organization.

Finally, regarding tuning the privacy budgets, practitioners can choose appropriate ones by carefully tailoring them to the specific requirements and sensitivity of their data domain [Dwork et al., 2019]. Then, the practitioner can leverage Theorem 2 and Proposition 3 to tune the noise scale σ and clipping values C, M to achieve the desirable fairness and utility.

VIII. Conclusion

This paper introduces FairDP, a novel mechanism that achieves certified group fairness while preserving DP and sustaining high model utility. Departing from existing mechanisms, the key ideas of FairDP are fairness-aware DP training and Monte Carlo approximation for fairness certification in the inference time, resulting in a rigorous privacy guarantee and fairness certification. Therefore, FairDP provides a comprehensive understanding of the influence of DP-preserving noise on model fairness guarantees and derives tight fairness certification by leveraging the DP-preserving noise. Our extensive experimental results showed that FairDP enhances the trade-off among model utility, privacy, and fairness, outperforming an array of baselines on benchmark datasets.

Acknowledgement

This work is partially supported by grants NSF CNS-1935928, NSF SaTC 2133169, NSF SaTC 1935923, NSF FAI 1939725, NSF SCH 2123809, and QRDI ARG01-0531-230438.

References

List of real-world uses of differential privacy. https://desfontain.es/privacy/real-world-differential-privacy.html. Accessed: 2024-01-24.

Abadi and et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Fair Credit Reporting Act. Fair credit reporting act. *Flood Disaster Protection Act and Financial Institute*, 2009.

Wael Mohammed A Alghamdi. *Estimation and Optimization of Information Measures with Applications to Fairness and Differential Privacy*. PhD thesis, Harvard University, 2023.

Bang An, Zora Che, Mucong Ding, and Furong Huang. Transferring fairness under distribution shifts via fair consistency regularization. *Advances in Neural Information Processing Systems*, 35:32582–32597, 2022.

Larry C Andrews. *Special functions of mathematics for engineers*, volume 49. Spie Press, 1998.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, pages 15479–15488, 2019.

Wenxuan Bao, Francesco Pittaluga, Vijay Kumar BG, and Vincent Bindschaedler. Dp-mix: mixup-based data augmentation for differentially private learning. *Advances in Neural Information Processing Systems*, 36, 2024.

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 Guha Thakurta. Private stochastic convex optimization with optimal rates. volume 32, 2019.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pages 1349–1361. PMLR, 2021.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

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.

Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Muhammad Bilal Zafar. Fairness measures for machine learning in finance. *The Journal of Financial Data Science*, 2021.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

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

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.

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

Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2), 2019.

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.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. In *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5470–5477, 2022. URL https://doi.org/10.24963/ijcai.2022/766.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Benedetta Giovanola and Simona Tiribelli. Beyond bias and discrimination: redefining the ai ethics principle of fairness in healthcare machine-learning algorithms. *AI & society*, 38 (2):549–563, 2023.

Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg 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.

Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25 (7):1445–1459, 2012.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. *arXiv preprint arXiv:2306.09468*, 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016b.

Tianyu He, Peiyi Han, Shaoming Duan, Zirui Wang, Wentai Wu, Chuanyi Liu, and Jianrun Han. Generative data augmentation with differential privacy for non-iid problem in decentralized clinical machine learning. *Future Generation Computer Systems*, 2024.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

Eugenia Iofinova, Nikola Konstantinov, and Christoph H Lampert. Flea: Provably fair multisource learning from unreliable training data. *arXiv preprint arXiv:2106.11732*, 2021.

Rashidul Islam, Kamrun Naher Keya, Shimei Pan, Anand D Sarwate, and James R Foulds. Differential fairness: An intersectional framework for fair ai. *Entropy*, 25(4):660, 2023.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. *arXiv preprint arXiv:1812.02696*, 2018.

Jiayin Jin, Zeru Zhang, Yang Zhou, and Lingfei Wu. Input-agnostic certified group fairness via gaussian parameter smoothing. In *International Conference on Machine Learning*, pages 10340–10361. PMLR, 2022.

Nikola Jovanović, Mislav Balunović, Dimitar I Dimitrov, and Martin Vechev. Fare: Provably fair representation learning. *arXiv preprint arXiv:2210.07213*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bogdan Kulynych, Yao-Yuan Yang, Yaodong Yu, Jarosław Błasiok, and Preetum Nakkiran. What you see is what you get: Principled deep learning via distributional generalization. *Advances in Neural Information Processing Systems*, 35:2168–2183, 2022.

Andrew Lowy, Devansh Gupta, and Meisam Razaviyayn. Stochastic differentially private and fair learning. In *Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, pages 86–119. PMLR, 2023.

Karima Makhlouf, Héber H Arcolezi, Sami Zhioua, Ghassen Ben Brahim, and Catuscia Palamidessi. On the impact of multi-dimensional local differential privacy on fairness. In *Submitted to ECML (Journal Track)*, 2024.

Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, pages 23681–23705. PMLR, 2023.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *ICLR*, 2018.

Frank 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.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness

in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021.

Hussein Mozannar, Mesrob I. Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

J Andrew Onesimu, J Karthikeyan, Jennifer Eunice, Marc Pomplun, and Hien Dang. Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access*, 10:86979–86997, 2022.

Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

Marlotte Pannekoek and Giacomo Spigler. Investigating trade-offs in utility, fairness and differential privacy in neural networks. *arXiv preprint arXiv:2102.05975*, 2021.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. 02 2018.

Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68, 2018.

Hai Phan, My T Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. Scalable differential privacy with certified robustness in adversarial learning. In *ICML*, pages 7683–7694. PMLR, 2020.

NhatHai Phan, Yue Wang, Xintao Wu, and Dejing Dou. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

NhatHai Phan, Xintao Wu, and Dejing Dou. Preserving differential privacy in convolutional deep belief networks. *Machine learning*, 106(9):1681–1704, 2017a.

NhatHai Phan, Xintao Wu, Han Hu, and Dejing Dou. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *2017 IEEE international conference on data mining (ICDM)*, pages 385–394. IEEE, 2017b.

NhatHai Phan, Minh Vu, Yang Liu, Ruoming Jin, Dejing Dou, Xintao Wu, and My T Thai. Heterogeneous gaussian mechanism: Preserving differential privacy in deep learning with provable robustness. *IJCAI*, 2019.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pages 7695–7705. PMLR, 2020.

Taki Hasan Rafi, Faiza Anan Noor, Tahmid Hussain, and Dong-Kyu Chae. Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, 105:102198, 2024.

Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. *Advances in neural information processing systems*, 33: 7584–7596, 2020.

Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning? In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 1738–1748, 2022.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*, 2024.

ITGP Privacy Team. *EU General Data Protection Regulation (GDPR)*. IT Governance Limited, 2017.

Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 27555–27565. Curran Associates, Inc., 2021a.

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, 2021b.

Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 9932–9939. AAAI Press, 2021c.

Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 560–566, 2021d.

Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 35: 17652–17664, 2022.

Cuong Tran, Keyu Zhu, Ferdinando Fioretto, and Pascal Van Hentenryck. SF-PATE: scalable, fair, and private aggregation of teacher ensembles. In *International Joint Conference on Artificial Intelligence*, pages 501–509. ijcai.org, 2023. doi: 10.24963/ijcai.2023/56. URL https://doi.org/10.24963/ijcai.2023/56.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

Robert Williamson and Aditya Menon. Fairness risk measures. In *International conference on machine learning*, pages 6786–6797. PMLR, 2019.

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.

Depeng Xu, Wei Du, and Xintao 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, 2021a. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467268. URL https://doi.org/10.1145/3447548.3467268.

Depeng Xu, Wei Du, and Xintao 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, 2021b.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. *ICLR*, 2020.

Xin-She Yang. *Engineering mathematics with examples and applications*. Academic Press, 2016.

I-Cheng Yeh and Che-hui 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.

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.

APPENDIX A
ALGORITHM

Algorithm 3 DPSGD

Input: Dataset D, sampling rate q, noise scale σ, norm bounds C and M, number of steps T.
Initialize θ^0 randomly
for $t \in [1:T]$ **do**
  Sample B^t from D with sampling probability q.
  Compute gradient: For $x_i \in B^t$, $g_i = \nabla_\theta \ell(x_i)$
  Clip gradient: $\bar{g}_i = g_i \min(1, \frac{C}{\|g_i\|_2})$
  Compute total gradient: $\Delta = \sum_{i \in B^t} \bar{g}_i$
  Add noise: $\tilde{\Delta} = \Delta + \mathcal{N}(0, C^2\sigma^2 I_r)$
  Update: $\theta^t = \theta^{t-1} - \frac{\eta_t}{|B^t|}\tilde{\Delta}$
end for
Return θ^T

APPENDIX B
PROOFS OF FAIRNESS CERTIFICATION

A. Proof of Proposition 3

Proof. From the considered fairness metric in Eq. (2), by leveraging Eq. (7), we have that:

$$\tau = \max_{u,v} Pr(\hat{y}=1|u,e) - Pr(\hat{y}=1|v,e) = \max_{u,v} \mathbb{E}_{u,e} - \mathbb{E}_{v,e} \tag{18}$$

where $\mathbb{E}_{k,e} = \mathbb{E}_{x\sim P(x|k,e)}\Big[\frac{1}{2} + \frac{1}{2}\texttt{erf}\Big(\frac{\langle w^{t-1} - \eta\mu^t, \xi\rangle}{\|\xi\|_2\sigma_0\sqrt{2}}\Big)\Big]$. Denote $\hat{\mathbb{E}}_{k,e}$ as the sample mean computed on the training set of group k. By leveraging the Central Limit Theorem, there exists an upper-bound $\hat{\mathbb{E}}^{ub}_{u,e} \geq \mathbb{E}_{u,e}$ and a lower-bound $\hat{\mathbb{E}}^{lb}_{v,e} \leq \mathbb{E}_{v,e}$ with the confident interval of $(1-\alpha)$ which can be computed using a tail-bound (e.g., Hoeffding inequality [Hoeffding, 1994]). Therefore, along with the MC approximation error, we have that

$$\tau = \max_{u,v} \mathbb{E}_{u,e} - \mathbb{E}_{v,e} \leq \max_{u,v} \hat{\mathbb{E}}^{ub}_{u,e} - \hat{\mathbb{E}}^{lb}_{v,e} + \mathcal{O}(N^{-1/2}) \tag{19}$$

with a confident interval of $(1-\alpha)$, which concludes the proof for Proposition 3. □

B. Discussion on MC error

As in [Gelman et al., 1995], the Monte Carlo approximation will be estimated to an error of approximately $\sqrt{\frac{Var\Big(\frac{1}{n_{k,e}}\sum_{x\in D_{k,e}} \mathbb{I}(h_{\theta_j^T}(x)>0)\Big)}{N}}$, where $\mathbb{I}$ is the indicator function and we also have:

$$Var\Big(\frac{1}{n_{k,e}}\sum_{x\in D_{k,e}} \mathbb{I}(h_{\theta_j^T}(x)>0)\Big) = \frac{1}{n^2_{k,e}}\sum_{x\in D_{k,e}} Var\big(\mathbb{I}(h_{\theta_j^T}(x)>0)\big) \tag{20}$$

with $\mathbb{I}(h_{\theta_j^T}(x)>0)$ is a random variable distributed by Bernoulli distribution $Bern\Big(Pr(z>0|\xi)\Big)$.

Therefore, $Var(\mathbb{I}(h_{\theta_j^T}(x)>0)) = Pr(z>0|\xi)[1-Pr(z>0|\xi)] \leq 1/4$. As a result, the Monte Carlo approximation error is approximately $\frac{1}{2n_{k,e}\sqrt{N}}$.

APPENDIX C
SUPPLEMENTAL RESULTS

TABLE III
IMPACT OF HYPER-PARAMETER M ON FAIRDP UNDER $\epsilon = 1.0$.

Dataset	M	AUC	Acc	DP	EOpp	EOdd
Adult	0.25	0.59	0.75	0.0	0.0	0.0
	0.50	0.77	0.76	0.001	0.002	0.001
	0.75	0.79	0.79	0.019	0.021	0.005
	1.00	0.80	0.80	0.027	0.038	0.011
Default-CCC	0.25	0.73	0.78	0.0	0.0	0.0
	0.50	0.75	0.80	0.011	0.026	0.009
	0.75	0.76	0.81	0.013	0.018	0.009
	1.00	0.76	0.81	0.016	0.016	0.009
UTK-Face	0.25	0.73	0.64	0.0	0.0	0.0
	0.50	0.79	0.71	0.087	0.090	0.044
	0.75	0.80	0.73	0.115	0.104	0.061
	1.00	0.80	0.75	0.130	0.115	0.068

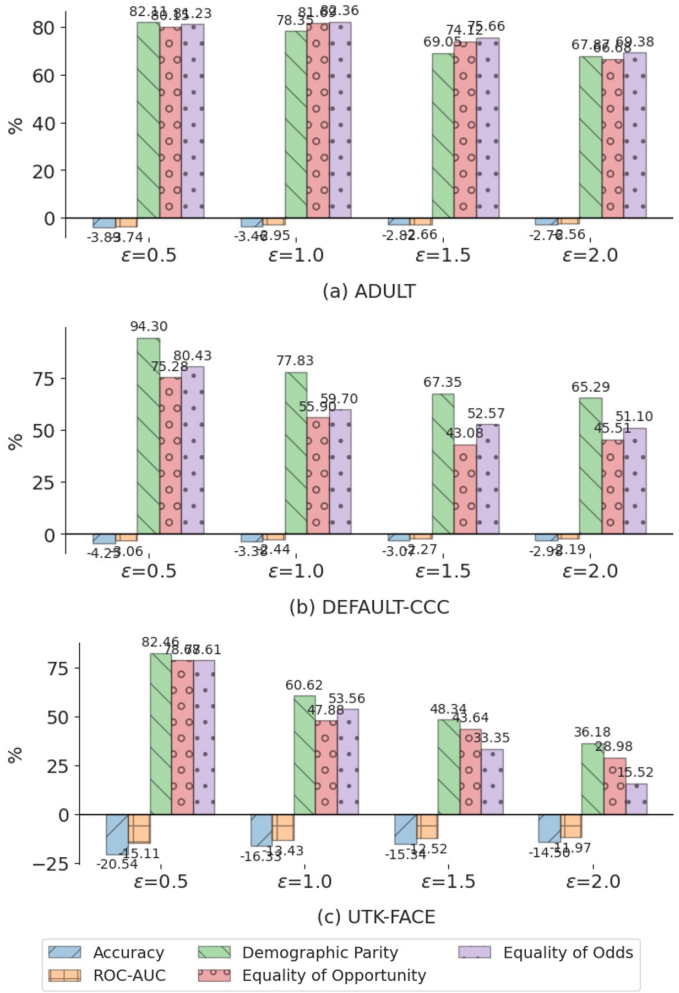

Fig. 8. Relative utility drop and fairness gain of FAIRDP compared with the clean model.

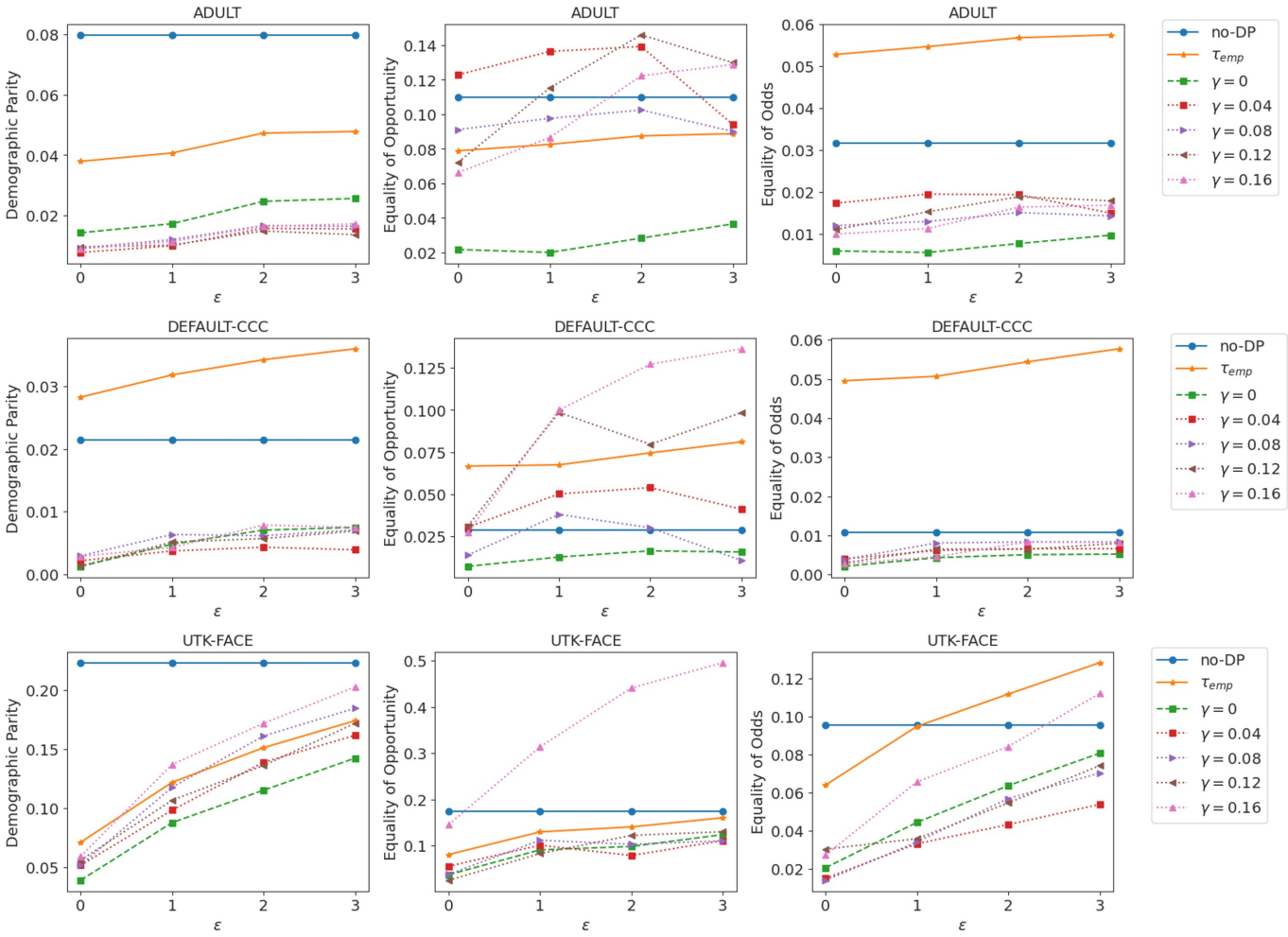

Fig. 9. Empirical fairness certification under distribution shift. γ is the total variance between $Pr(y, a)$ of the train and test sets.

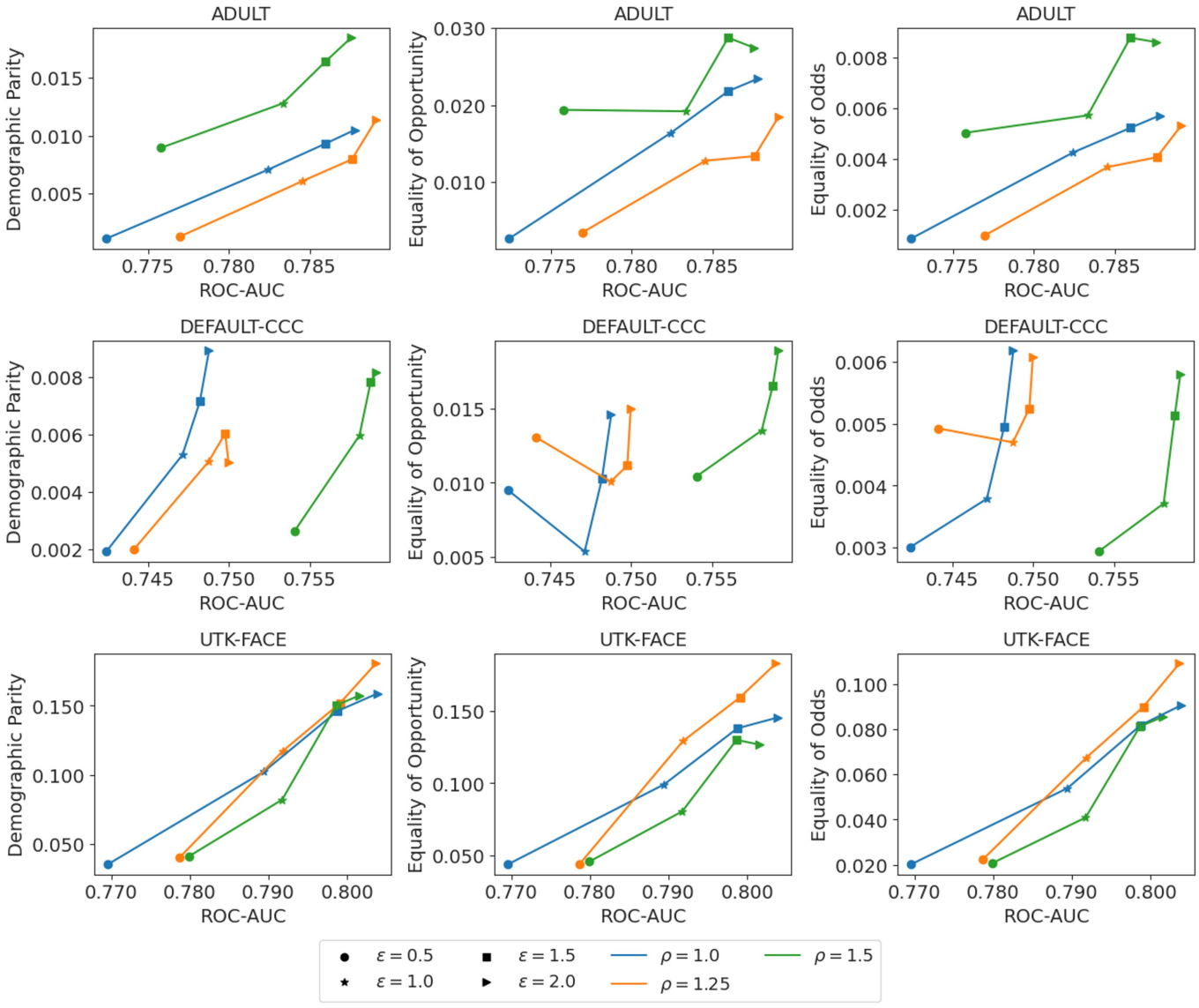

Fig. 10. Model utility (ROC-AUC), privacy, and fairness for various ρ values, which measure the ratio between the number of data points of advantage and disadvantage groups.

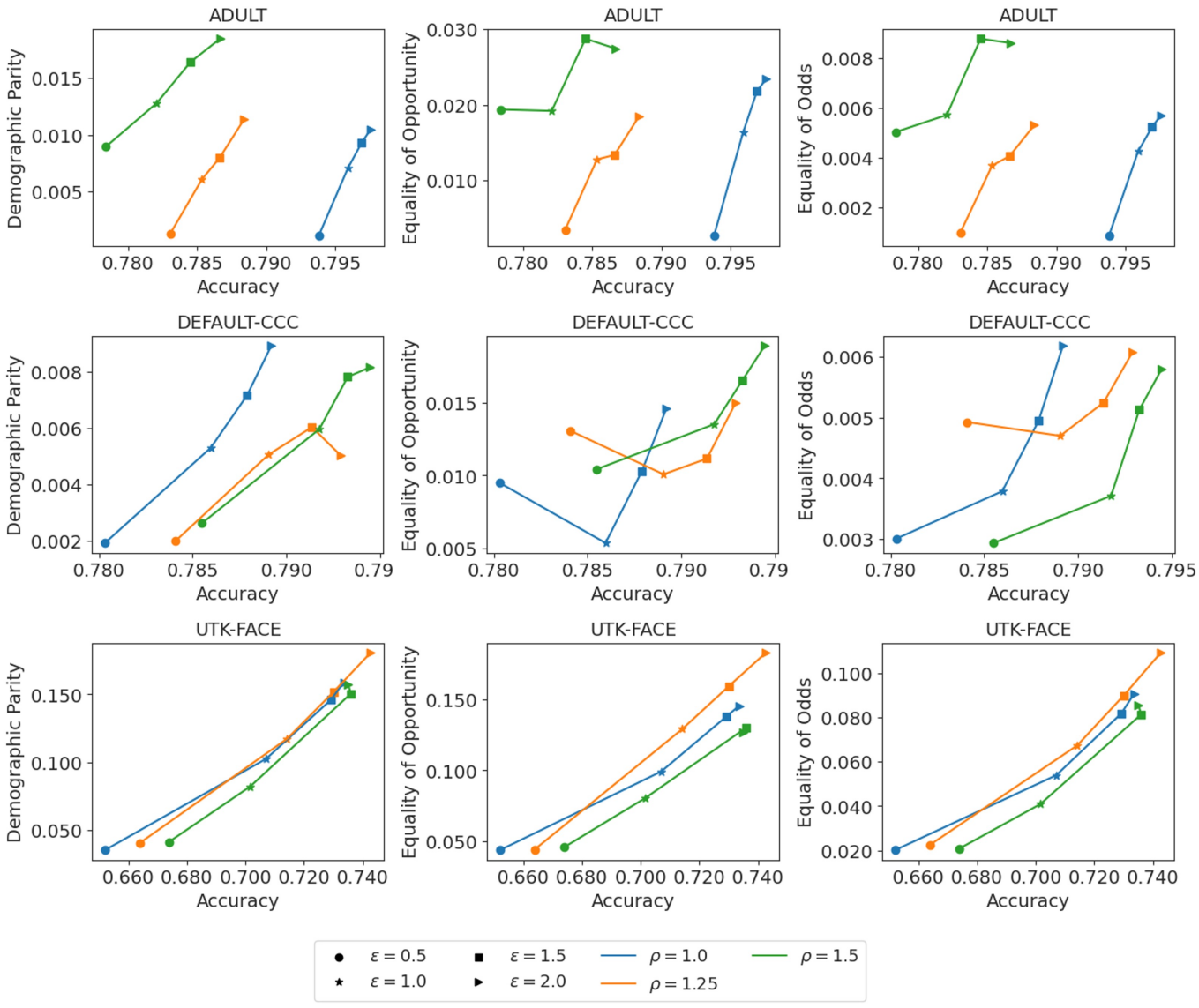

Fig. 11. Model utility (Accuracy), privacy, and fairness for various ρ values, which measure the ratio between the number of data points of advantage and disadvantage groups.