
On the Fairness of Privacy Protection: Measuring and Mitigating the Disparity of Group Privacy Risks for Differentially Private Machine Learning

Zhi Yang¹* Chuangwu Huang¹† Ke Tang¹ Xin Yao²

¹Southern University of Science and Technology

²Lingnan University

Abstract

While significant progress has been made in conventional fairness-aware machine learning (ML) and differentially private ML (DPML), the fairness of privacy protection across groups remains underexplored. Existing studies have proposed methods to assess group privacy risks, but these are based on the average-case privacy risks of data records. Such approaches may underestimate the group privacy risks, thereby potentially underestimating the disparity across group privacy risks. Moreover, the current method for assessing the worst-case privacy risks of data records is time-consuming, limiting their practical applicability. To address these limitations, we introduce a novel membership inference game that can efficiently audit the approximate worst-case privacy risks of data records. Experimental results demonstrate that our method provides a more stringent measurement of group privacy risks, yielding a reliable assessment of the disparity in group privacy risks. Furthermore, to promote privacy protection fairness in DPML, we enhance the standard DP-SGD algorithm with an adaptive group-specific gradient clipping strategy, inspired by the design of canaries in differential privacy auditing studies. Extensive experiments confirm that our algorithm effectively reduces the disparity in group privacy risks, thereby enhancing the fairness of privacy protection in DPML.

1 Introduction

Artificial intelligence (AI), particularly machine learning (ML), has been widely adopted across various sectors, augmenting and even replacing human decision-making. However, its growing integration in critical domains like healthcare, finance, and judiciary has raised critical concerns, including data privacy breaches, algorithmic biases, lack of explainability, security vulnerabilities etc. [10]. Among these ethical issues and risks, privacy and fairness have emerged as two pivotal and widely discussed challenges [8], attracting substantial attention from the research community.

Research in privacy protection and fairness has made significant strides independently, and the intersection of these two critical issues has also gained considerable attention. Some studies aim to achieve both privacy protection and outcome fairness simultaneously in ML models [26, 11, 5, 25, 15]. Other works explore how privacy mechanisms affect the outcome fairness [3] and propose methods to mitigate the unfairness introduced by such mechanisms [27, 24, 9]. However, whether AI systems can provide equal privacy protections to different groups is also a noteworthy yet understudied problem at the intersection of fairness and privacy. As highlighted in [8], this raises an essential yet still underexplored and insufficiently addressed question: *Do AI systems offer fair or equitable privacy*

*Email: 12332454@mail.sustech.edu.cn

†Email: huangcw@sustech.edu.cn

protections across groups? This question raises both ethical and practical concerns, as certain groups may face disproportionately higher privacy leakage risks, violating principles of fairness and equality.

To address this issue, it is essential first to provide a rigorous answer to the question. Prior studies have empirically examined whether privacy leakage risks are evenly distributed across groups, and their auditing methods for group privacy risk rely on averaging the performance across data points within each group under membership inference attacks (MIAs) [4, 28]. The MIAs employed in these studies are formulated based on the membership inference game (MIG) introduced in [30], which captures the average behavior across all data points. However, such an average-case MIG may obscure the heterogeneity of individual privacy risks within groups and potentially underestimate the privacy leakage risks faced by certain groups. This, in turn, may lead to inaccurate and unreliable measurements of inter-group disparities in privacy risk. Consequently, the issue of privacy inequality may not be adequately uncovered.

Measuring. Therefore, we aim to provide a tighter measurement of the privacy risk of data points and then analyze the disparities in privacy risk across different groups. To this end, we can leverage the leave-one-out attack (LOOA), which is formulated based on the worst-case MIG proposed by [29] and is capable of estimating the worst-case privacy leakage risk for each data point. However, the computational cost of LOOA is prohibitively high, rendering its practical implementation nearly infeasible. To address this challenge, we propose an approximate version of the worst-case MIG to efficiently audit the approximate worst-case privacy risk of individual data points. Our experiments demonstrate that: 1) The attack simulating our proposed MIG can achieve comparable performance to LOOA as the number of attack rounds increases; 2) The individual privacy risks evaluated by our method can be reliable for assessing group privacy risk.

Then we define a fairness metric to quantify the degree of privacy unfairness (i.e., the disparity of privacy risk across groups). Through experiments, we clearly demonstrate that our auditing method significantly outperforms previous auditing methods under the same conditions. Specifically, our method reveals greater group privacy risk and more effectively captures privacy inequality. Consistent with prior research [4, 28], we find that existing ML algorithms exhibit significant unfairness in privacy risks across groups. While differentially private ML (DPML) algorithms can bind the magnitude of privacy risk disparities between groups, a certain degree of disparity still persists.

Mitigating. Upon providing a more thorough answer to the above question, we seek to alleviate this issue. Inspired by the design of canaries in DP auditing studies [18, 2, 23], we confirm that groups with larger gradient norms under training process—indicating greater contributions to model updates—are more prone to higher privacy leakage risks. Building on this insight, we enhance the existing DPML algorithm by adaptively setting group-specific gradient clipping norms. Extensive experimental results demonstrate that our algorithm effectively mitigates the disparity of group privacy risk, promoting the ethical and effective deployment of AI systems.

In summary, our main contributions are as follows:

- We propose a novel MIG to efficiently and approximately audit the worst-case privacy risks of individual data points.
- Our auditing mechanism offers a more stringent measurement of group privacy risks, enabling a tighter and more accurate assessment of disparities between groups.
- We design an enhanced DPML algorithm to reduce the group privacy risk disparities, thereby improving the fairness of privacy protection.

2 Background

2.1 Differential privacy

Differential Privacy (DP), proposed by [7], is a privacy framework designed to address privacy leakage. It has become the predominant method for ensuring algorithmic privacy [20]. In the following, we introduce the approximate (ϵ, δ) -DP definition.

Definition 1 ((ϵ, δ) -Differential Privacy [6]). An algorithm \mathcal{M} is said to satisfy approximate differential privacy if for all pairs of adjacent databases D and D' that differ on a single data record and all

possible outputs $O \subseteq \text{Range}(\mathcal{M})$, the following condition holds:

$$P[\mathcal{M}(D) \in O] \leq e^\epsilon \times P[\mathcal{M}(D') \in O] + \delta, \quad (1)$$

where e^ϵ provides an upper bound such that the adversary cannot distinguish whether the algorithm \mathcal{M} was trained on D or D' .

Differentially private stochastic gradient descent (DP-SGD). DP-SGD [1] is a widely adopted algorithm in DPML [20]. It integrates DP concepts with stochastic gradient descent (SGD). This integration ensures model privacy by employing gradient clipping and noise addition within the SGD framework, adhering to the (ϵ, δ) -DP definition. The pseudocode of DP-SGD is shown in Algo. 2.

2.2 Black-box member inference attacks

The goal of membership inference attacks (MIAs) is to determine whether a specific data record is part of the training dataset. We focus on a black-box setting, where the adversary only has access to model outputs, reflecting a more realistic scenario where the training process is inaccessible [2]. In black-box MIAs, the adversary infers membership by analyzing the model’s output behavior, typically using the sample’s output loss as an inference score or decision basis, relying on the observation that models tend to show smaller losses for training samples [30].

Different definitions of membership inference games (MIGs). MIGs conceptualize MIAs as inference games between a privacy auditor (i.e., the adversary) and a challenger. MIAs are typically carried out by simulating the MIGs through multiple rounds of random experiment. Various definitions of MIGs have been proposed, each designed to capture different aspects of privacy risk [29].

Most MIAs follow the average-case MIG framework [30] (see Def.6 in App.A)), which evaluates the vulnerability of a target model to the adversary, emphasizing the average behavior across data points [29]. A common strategy formalized under this framework is the global attack (GA), where a single inference threshold is determined based on the aggregate behavior of all data points in a given round [30, 21]. The group-based attack (GBA) extends GA by assigning a distinct threshold to each group, enabling a more fine-grained analysis of group-level privacy risks and offering deeper insight into disparities across demographic partitions [4, 28].

Nonetheless, such average-case MIAs fail to capture worst-case privacy risks for individual data records. The worst-case MIG [29] (see Def.7 in App.A) addresses this limitation by evaluating the maximum risk a single record may encounter. A concrete instance is the Leave-One-Out Attack (LOOA), which independently evaluates each data point’s worst-case exposure, aligning closely with the principles of DP. However, LOOA is computationally intensive, as it requires evaluating each record separately.

Privacy auditing. Privacy auditing is designed to proactively assess privacy risks and quantify potential leakage, typically during the model development phase. In contrast, MIAs are conducted post-deployment by adversaries aiming to exploit trained models. Privacy auditing uses MIAs for evaluation but with more background knowledge for the adversary, including access to the original dataset and knowledge of the optimal threshold. This setup simulates worst-case scenarios, enabling rigorous assessment of privacy leakage risks [4]. In this work, we focus on privacy auditing to systematically evaluate privacy vulnerabilities. Privacy auditing using the GA (PA-GA)[30], the GBA (PA-GBA) [4, 28], and the LOOA (PA-LOOA) [18, 2] for evaluating privacy leakage risks is detailed in Algos. 3, 4, and 5 of App. A, respectively.

2.3 The fairness of privacy

Few studies have explored fairness in the context of privacy, and those that do vary in their research focus. For instance, one study finds that fairness-aware algorithms can exacerbate disparities in privacy leakage across groups [4]. Meanwhile, other research highlights that ML algorithms exhibit significant group-level disparities in privacy leakage, and that DPML algorithms can help reduce these disparities [28]. These auditing mechanisms typically rely on average-case attacks, which assess privacy risks based on the average behavior of data points. While effective at capturing general trends, such approaches may overlook the nuanced privacy risks faced by individual samples, potentially concealing disparities between groups.

In our work, we aim to more precisely evaluate group-level privacy risks compared to prior studies. To ensure fair comparison, we adopt the same attack metrics as in [4, 28], including inference scores based on per-example loss and attack success rates.

3 Measuring the disparity of group privacy risks with approximate worst-case privacy auditing

In this section, we first propose an alternative to PA-LOOA and demonstrate that, while it improves efficiency, it achieves comparable performance as the number of repeated experiments increases. We further introduce a fairness metric to assess the disparity of inter-group privacy risks. Experimental results show that our method uncovers more pronounced group privacy risks and offers a more reliable assessment of privacy inequalities between groups compared to existing approaches, thereby exhibiting stronger auditing capability.

3.1 Approximate worst-case privacy auditing

As previously discussed, the LOOA is computationally expensive. Specifically, obtaining statistically reliable results for a single sample typically requires $2R$ repeated experiments. Consequently, auditing m samples would involve training $m \times 2R$ models, making this approach impractical for real-world applications due to the extensive time and computational resources required. To address this limitation, we propose a new MIG that allows for the simultaneous auditing of multiple samples within a single auditing process, as presented in Def. 2.

Definition 2 (Approximate Worst-case MIG). Let Ω denotes the underlying population data pool, \mathcal{M} the training algorithm, and \mathcal{A} the inference algorithm. We assume that the challenger samples n i.i.d. records from Ω to construct the training dataset D , and $Z = \{z_i\}_{i=1}^m \subseteq D$ represents the auditing samples.

- 1) The challenger flips fair choices $\{h_i\}_{i=1}^m$ randomly, where $h_i \in \{0, 1\}$, indicating whether each record z_i is included in the training or not.
- 2) The challenger samples a fixed record $z \sim Z$ along with its status h .
- 3) The challenger trains a model $f_h \leftarrow \mathcal{M}(D \setminus \{z_i \mid h_i = h\})$, and a model $f_{\sim h} \leftarrow \mathcal{M}(D \setminus \{z_i \mid h_i = \sim h\})$.
- 4) The challenger flips a fair coin $b \in \{0, 1\}$, and sends the target model and record (f_b, z) to the adversary.
- 5) The adversary, with access to the target model, outputs a guess $\hat{b} \leftarrow \mathcal{A}(f_b, z)$.
- 6) The game outputs 1 (success) if $\hat{b} = b$, and 0 otherwise.

We refer to the attack that simulates this game as the approximate leave-one-out attack (ALOOA). The process of privacy auditing using ALOOA (PA-ALOOA) to evaluate privacy risks is detailed in Algo. 6 of App. A. The auditing mechanism achieves computational efficiency by auditing multiple samples simultaneously while preserving analytical granularity by analyzing the behavior of each sample, rather than relying on aggregate statistics across multiple data points.

Comparison with PA-LOOA. The difference between the two approaches stems from their sample selection strategies. As shown in Fig. 1, PA-LOOA introduces minimal randomness into the training set, as only a single sample is randomly included or excluded in each round. In contrast, PA-ALOOA audits m samples simultaneously, with each experiment randomly choosing which samples are included in the training set, thus introducing greater variability. However, we argue that with sufficient rounds,

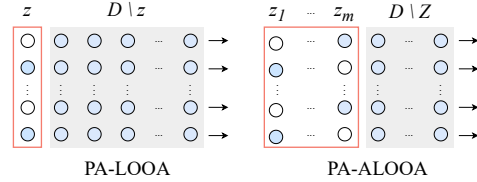
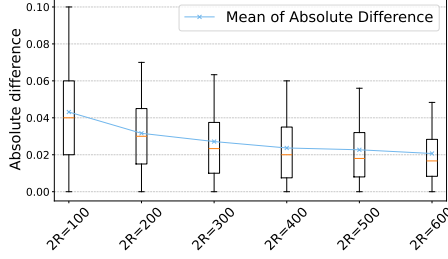


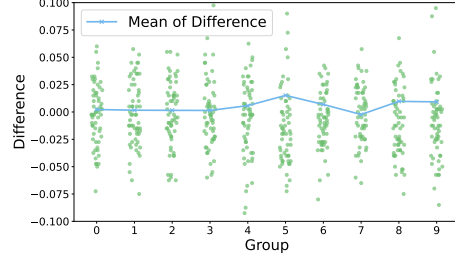
Figure 1: Left: PA-LOOA audits a single sample. Right: PA-ALOOA audits m samples. Solid circles indicate training points; hollow circles are excluded. Arrows denote model training using solid-circle data.

the random fluctuations in PA-ALOOA will average out, resulting in performance comparable to that of PA-LOOA.

We validate our hypothesis through practical experiments using the widely used MNIST dataset, training a Convolutional Neural Network (CNN) with SGD. Due to computational limitations, we randomly select 60 samples per class, totaling 600 samples to audit for PA-LOOA and PA-ALOOA. We evaluate the attacker’s performance using the accuracy metric (i.e., attack success rate) from [22], which measures the agreement between the adversary’s guesses and the actual status. Instead of evaluating overall accuracy, we compute individual accuracy for each data point. Detailed experimental settings and additional results are provided in the App. B.



(a) Comparison at the individual level.



(b) Comparison at the group level.

Figure 2: Left: The horizontal axis represents the number of random experiments for a single audit, while the vertical axis represents the absolute difference in auditing performance between the two attacks for each audited sample. Right: The horizontal axis represents different groups of the MNIST dataset, while the vertical axis indicates the performance difference between PA-LOOA and PA-ALOOA for individual data points in each group at $2R = 400$.

As shown in Fig. 2a, the experimental results indicate that as the number of random trials increases, the average absolute difference in individual accuracy between the two methods gradually decreases and eventually stabilizes. The signed differences are presented in Fig. 6 of App. B, from which we observe that the two approaches exhibit similar behavior on average, with the mean difference consistently remaining below 0.01 across all $2R$ values. Moreover, the variance of the differences further decreases as $2R$ increases. However, it is evident that the discrepancy between the two approaches still exhibits notable variance across individual data points.

Although the estimation errors for individual samples may vary significantly, we find that statistical outcomes across groups are reliable, forming a solid foundation for analyzing group privacy risk. As shown in Fig. 2b, the distribution of performance differences between PA-LOOA and PA-ALOOA within each group is highly similar. Moreover, the Kruskal-Wallis test yields p -values greater than 0.4 for all $2R$ values considered in our experiment, indicating no statistically significant differences in performance between the two methods across groups. Furthermore, the average performance difference of each group between PA-LOOA and PA-ALOOA is minimal, suggesting that extending individual-level estimates to group-level statistical analysis introduces only negligible error.

3.2 Definition of group privacy risk parity

We assess the privacy leakage risk of data points using the attacker’s membership advantage, following prior work [28, 30]. Below, we formally define the notion of individual privacy risk (IPR) as applied to a single data record in this study.

Definition 3 (Individual Privacy Risk). Let $Acc_i(\mathcal{A}, Z)$ represent the attack accuracy of a data i under privacy auditing algorithm \mathcal{A} and the auditing dataset Z . The individual privacy risk is defined as:

$$Adv_i(\mathcal{A}, Z) = 2Acc_i(\mathcal{A}, Z) - 1 \quad (2)$$

This formulation quantifies the adversary’s normalized advantage over random guessing. Building on the concept of IPR, we can extend it to define Group Privacy Risk (GPR) as in [4].

Definition 4 (Group Privacy Risk). Let D^k denote the subset of the dataset D belonging to group k . The group privacy risk is defined as:

$$Adv^k(\mathcal{A}, Z) = \mathbb{E}_{i \in D^k} [Adv_i(\mathcal{A}, Z)] \quad (3)$$

Based on GPR, we evaluate whether privacy leakage risk is fair or equitable across different groups by introducing the notion of Group Privacy Risk Parity (GPRP).

Definition 5 (Group Privacy Risk Parity). Let K represent the set of all groups. We define group privacy risk parity as:

$$\Delta = \max_{k \in K}(\text{Adv}^k(\mathcal{A}, Z)) - \min_{k \in K}(\text{Adv}^k(\mathcal{A}, Z)) \quad (4)$$

This metric provides a systematic means of quantifying the disparity in privacy risk across groups, capturing the gap between the most and least vulnerable groups.

3.3 Comparison with privacy auditing by average-case attacks

We compare the GPR and the GPRP metrics measured by our auditing method, PA-ALOOA, with those obtained from privacy auditing by average-case attacks (PA-ACAs) in previous studies: PA-GA [28] and PA-GBA [4]. To ensure a fair comparison, all three methods are configured identically, keeping the training dataset and model consistent across each repeated experiment. Specifically, for PA-ALOOA, we set the number of audit samples to $m = n$, following the same setting used in PA-ACAs. This setup ensures consistency and reflects a real-world scenario in which the privacy risk of every training sample is audited. The key distinction lies in threshold determination: PA-ACAs computes thresholds based on the aggregate behavior of multiple data points, whereas PA-ALOOA assigns a unique threshold to each sample, based on its behavior across all repeated experiments.

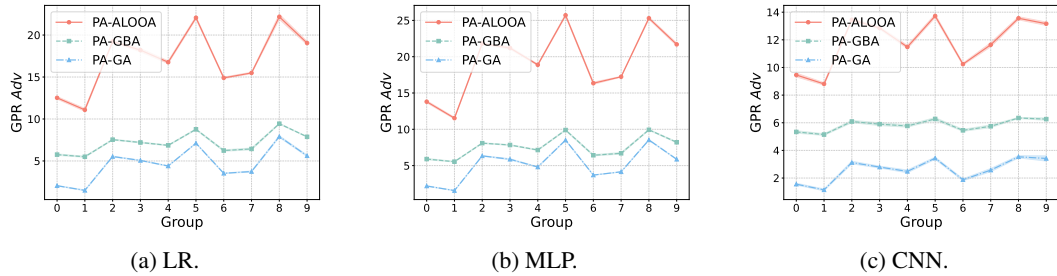


Figure 3: The comparison of GPR value across three model types—Logistic Regression (LR), Multilayer Perceptron (MLP), and CNN—trained on the MNIST dataset using SGD algorithm. The x-axis represents the groups, and the y-axis shows the corresponding GPR value at $2R = 400$.

Consistent with prior studies on privacy and fairness [3, 27, 9], our foundational analysis focuses on the MNIST dataset. Detailed experimental configurations and supplementary results for other datasets are included in the App. C. The results of the GPR and GPRP metrics on the MNIST dataset are presented in Fig. 3 and Tab. 1, with all values reported in percentage points for clarity. As shown in Fig. 3, the three auditing methods exhibit consistent patterns in the distribution of privacy risks across groups. Among them, PA-ALOOA consistently yields significantly higher GPR values across all model architectures compared to the other two auditing methods. Specifically, Tab. 1 shows that for the CNN model, the GPRP value obtained by PA-GA and PA-GBA suggests that DP-SGD results in higher disparity than standard SGD. This observation contradicts the conclusion in [28], which asserts that DPML algorithms should be able to bind the disparity of GPR relative to non-private counterparts. Such inconsistency implies that PA-ACAs underestimate the GPR, leading to inaccurate measurements and incorrect conclusions. In summary, PA-ALOOA offers a more rigorous means of capturing privacy risk, and thus providing more reliable evaluations of privacy unfairness across groups.

Table 1: The comparison of GPRP value computed by different privacy auditing methods for a CNN model trained on the MNIST dataset at $2R = 400$.

Method	Δ_{PA-GA}	Δ_{PA-GBA}	$\Delta_{PA-ALOOA}$
SGD	2.452	1.248	4.920
DP-SGD	2.625	1.254	3.540

4 Mitigating the disparity of group privacy risks for DP-SGD

While DPML algorithms can limit the extent of privacy risk disparities across groups, previous results demonstrate that such disparities still persist to a noticeable degree. In this section, we investigate and confirm a strong correlation between GPR and the group contribution of gradients during training. Motivated by this finding, we propose an enhanced DP-SGD algorithm designed to improve the fairness of privacy protection across different groups.

4.1 Experimental observations

In DP auditing literature, canaries are often created as mislabeled samples [18, 2, 23]. These samples generate larger gradient values during model training, which contribute more to parameter updates, thereby increasing the likelihood of being memorized by the model. Motivated by the design of canaries, we hypothesize that during training, the larger a group’s contribution to the gradient, the more likely the model is to memorize that group. Thus, groups with larger contributions are expected to face a higher privacy leakage risk compared to those with smaller contributions.

We conduct experimental analysis to validate our hypothesis. We first compute the sum of gradient vectors for a group k within a batch B and average it by dividing by the number of samples in that group $|D^k|$, i.e., $\sum_{i \in B \cap D^k} g_i / |D^k|$. The norm of this vector is then divided by the norm of the gradient used for the model update, i.e., $\sum_{i \in B} g_i / |B|$, to represent the group’s relative contribution in this iteration. We obtain group relative contribution (GRC) by averaging this ratio across all training iterations. As shown in Fig. 11 of App. D, there is a significant correlation between GRC and GPR across different models, confirming our hypothesis. Specifically, groups contributing more during training exhibit higher privacy leakage risks.

4.2 Design of mitigation algorithm for DP-SGD

Building on the previous observation, we propose an improvement to the DP-SGD algorithm to promote fair privacy protection across groups. In DP-SGD, the gradient clipping operation uses a unified clipping bound for all groups, whereas we adaptively set different clipping bounds for each group based on the GRC during training.

Algorithm 1 DP-SGD-S

Input: Training dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the parameterized model $f_w(\cdot)$, loss function ℓ , iterations T , batch size b , learning rate η , noise scale σ_1, σ_2 , clipping bound C , scale bound τ .

- 1: Initialize $w^{(0)}$ randomly.
- 2: **for** $t = 0, \dots, T - 1$ **do**
- 3: Sample a batch B from D with probability b/N .
- 4: **for** $i \in B$ **do**
- 5: $g_i \leftarrow \nabla \ell(f_{w^{(t)}}(\mathbf{x}_i), y_i)$
- 6: $\bar{g}_i \leftarrow g_i \cdot \min\left(1, \frac{1}{\|g_i\|_2}\right)$
- 7: **end for**
- 8: **for** $k \in K$ **do**
- 9: $C^k \leftarrow C \cdot \min\left(\tau, \frac{\|\frac{1}{b}(\sum_{i \in B} \bar{g}_i + \mathcal{N}(0, \sigma_1^2 \mathbf{I}))\|_2}{\|\frac{1}{|D^k|}(\sum_{i \in B \cap D^k} \bar{g}_i + \mathcal{N}(0, \sigma_1^2 \mathbf{I}))\|_2}\right)$
- 10: **end for**
- 11: **for** $i \in B$ **do**
- 12: $\tilde{g}_i \leftarrow g_i \cdot \min(1, \frac{C^k}{\|g_i\|_2})$
- 13: **end for**
- 14: $C = \max_{k \in K}(C^k)$
- 15: $\tilde{g} \leftarrow \frac{1}{b}(\sum_{i \in B} \tilde{g}_i + \mathcal{N}(0, \sigma_2^2 C^2 \mathbf{I}))$
- 16: $w^{(t+1)} \leftarrow w^{(t)} - \eta \tilde{g}$
- 17: **end for**

Return: Model $f_{w^{(T)}}(\cdot)$ and accumulated (ϵ, δ) .

As shown in Algo. 1, our proposed algorithm, DP-SGD-Scale (abbreviated as DP-SGD-S), differs from the standard DP-SGD in Lines 6–14. In each iteration, it estimates the relative contribution of each group’s samples to the overall gradient and uses this information to adaptively adjust the clipping bound for each group. To preserve privacy, we add noise to the clipped group-level gradient statistics used for computing the group-specific clipping bounds C^k . This additional privacy cost is incorporated into the overall privacy accounting via the composition theorem [1, 27]. Although different groups are assigned distinct clipping bounds, we conservatively bound the sensitivity of the final aggregated gradient by $\max_k C^k$, ensuring that the overall mechanism satisfies (ϵ, δ) -DP in the same sense as DP-SGD. Following prior work [27, 9], we set $\sigma_1 \approx 10\sigma_2$ so that the privacy cost of computing C^k is negligible relative to the total privacy budget.

In DP-SGD-S, groups with higher contributions have their clipping bounds scaled down, leading to stricter clipping operations. This adjustment limits the influence of these groups on model updates, thereby reducing the model’s memorization of these groups and mitigating their privacy leakage risks. Conversely, groups with relatively smaller contributions are assigned larger clipping bounds. The scaling factor of clipping bounds is constrained by the hyperparameter τ , as excessively large clipping norms would introduce too much noise, making the model’s performance unreliable.

5 Experimental study

In this section, we validate the effectiveness of our algorithm, DP-SGD-S, in mitigating the disparity of privacy risk across groups through extensive experiments.

5.1 Experimental setup

Full experimental details are provided in App. D.1. We conduct experiments on datasets commonly used in privacy and fairness research [2, 3, 27], including MNIST [14], as well as three fairness-related datasets: two tabular datasets, Adult and Law[13], and one image dataset, UTKFace[33]. Our study compares three training algorithms: standard SGD, DP-SGD, and our proposed DP-SGD-S. For both DP-SGD and DP-SGD-S, the default privacy parameters are set to $(\epsilon, \delta) = (10, 1e-5)$, and the default scale bound τ for DP-SGD-S is set to 2. Three model architectures are considered: LR, MLP, and CNN. To measure fairness in privacy protection across groups, we use the GPRP metric, assessed via our auditing method PA-ALOOA. The model utility is evaluated through classification accuracy. All results reported represent the average of five independent runs, with all values presented in percentage points for clarity.

5.2 Experimental results

Results across different datasets. We evaluate our proposed algorithm, DP-SGD-S, on multiple datasets to demonstrate its effectiveness in mitigating disparities in privacy leakage risks among groups. As shown in Table 2, DP-SGD-S consistently achieves the lowest Δ across all datasets. These results highlight that our enhancement to DP-SGD leads to a more equitable privacy protection mechanism. For the tabular datasets Adult and Law, the classification accuracy remains nearly unchanged between the non-private and private training algorithms. In these cases, DP-SGD-S successfully reduces Δ without compromising model utility. For the image datasets, MNIST and UTKFace, DP-SGD leads to an accuracy drop of approximately 2% compared to standard SGD, and DP-SGD-S incurs a drop of about 2% compared to DP-SGD. This indicates that DP-SGD-S incurs a slight accuracy trade-off in this scenario, but the degradation is modest and accompanied by enhanced fairness in privacy protection. The results of the other dataset are shown in App. D.2.2.

Results across different privacy guarantees. We conduct extensive experiments to compare the performance of three training algorithms under varying levels of privacy guarantees. In particular, we include $\epsilon = 1$ for each dataset when applying the differentially private training algorithms DP-SGD and DP-SGD-S. Due to space constraints, we present only the results on the MNIST dataset with DP-SGD in the main text; comprehensive results for all datasets and methods are provided in App. D.2.2. As illustrated in Fig. 4, the stronger the model’s privacy protection capability, the smaller the differences in privacy risk between groups. This is actually a rather intuitive conclusion. Imagine an extreme scenario where all data points in the model can ensure a privacy budget of 0; in this case, there would be no privacy risk differences between any points or groups. However, in practice, this is

Table 2: The results of three training algorithms under different datasets.

Metric	Method	MNIST	Adult	Law	UTKFace
Accuracy (\uparrow)	SGD	95.89 ± 0.29	85.00 ± 0.07	89.75 ± 0.12	85.93 ± 3.79
	DP-SGD	94.46 ± 0.13	84.92 ± 0.04	89.74 ± 0.08	86.84 ± 0.48
	DP-SGD-S	92.57 ± 0.42	84.86 ± 0.08	89.60 ± 0.11	84.56 ± 0.11
GPRP Δ (\downarrow)	SGD	4.92 ± 0.18	0.42 ± 0.04	0.90 ± 0.16	1.75 ± 0.11
	DP-SGD	3.54 ± 0.13	0.27 ± 0.04	0.59 ± 0.06	1.19 ± 0.07
	DP-SGD-S	2.92 ± 0.14	0.16 ± 0.02	0.43 ± 0.03	0.74 ± 0.07

not feasible because the stricter the privacy budget, the less usable the model’s prediction accuracy becomes. Therefore, our method manages to achieve more equitable privacy protection under the same privacy guarantees compared to DP-SGD, which is meaningful.

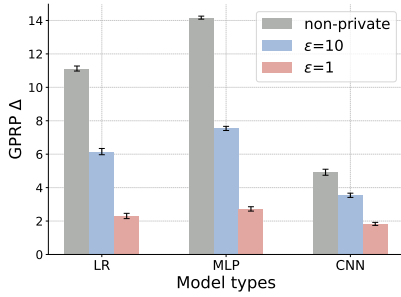


Figure 4: The results of the SGD and DP-SGD algorithms on the MNIST dataset under varying privacy guarantees and model architectures.

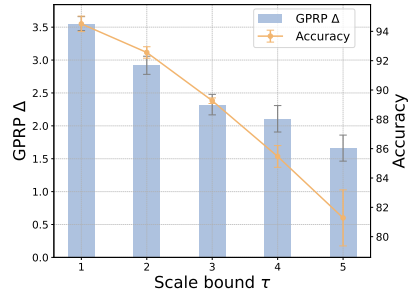


Figure 5: The results of the DP-SGD-S algorithm on the MNIST dataset using a CNN model under varying scale bounds.

Results across different scale bounds. We evaluate the impact of different scale bounds τ in DP-SGD-S on both accuracy and GPRP metrics on the MNIST dataset. As shown in Fig. 5, increasing the τ leads to a decrease in Δ , as larger τ further limits the contribution of groups with larger norms to model updates. However, this improvement in fairness comes with a trade-off in accuracy, likely due to the model’s diminished ability to extract optimization information from these groups.

5.3 Limitation and discussion

While our proposed method, DP-SGD-S, demonstrates strong effectiveness in reducing disparities in group privacy risks across diverse datasets, it also presents certain limitations that warrant further discussion. First, DP-SGD-S may introduce a slight drop in model accuracy compared to standard DP-SGD. This reflects a common trade-off where enhancing fairness in privacy protection may come at the expense of predictive performance. In practice, this trade-off is often acceptable, but it remains an important consideration in high-accuracy applications. Second, as mentioned in Sec. 4.2, DP-SGD-S requires a small portion of the overall privacy budget to protect the gradient statistics used during training.

Moreover, to provide a comprehensive understanding of group-level privacy behavior, we report detailed results in App. D.2.3 across all datasets and settings. In over 90% of the cases, DP-SGD-S does not increase the privacy risk for advantaged groups. Instead, privacy risks either decrease or remain stable for all groups, with more notable improvements in disadvantaged groups. This demonstrates that DP-SGD-S enhances privacy fairness without causing a “leveling down” effect, which is essential for real-world applications. We also examine the impact of DP-SGD-S on conventional outcome fairness metrics, such as demographic parity and accuracy parity, with the results also provided in App. D.2.4. Our findings suggest that DP-SGD-S does not exacerbate outcome unfairness. However, any outcome unfairness already present in models trained with DP-SGD still exists under DP-SGD-S. Addressing these remaining limitations and developing algorithms

that simultaneously promote both outcome fairness and privacy fairness is an important avenue for future research.

6 Conclusion

This work addresses a fundamental challenge at the intersection of fairness and privacy in AI systems: ensuring equitable privacy protection across different demographic groups. Our study makes two significant contributions to this emerging research direction. First, we develop a novel membership inference game-based privacy auditing mechanism that enables more rigorous measurement of group privacy risks. The empirical results prove that our method provides a more rigorous and reliable assessment of privacy risk disparities across groups while maintaining computational efficiency. Second, to mitigate the identified privacy protection disparities, we propose an enhanced DP-SGD algorithm that incorporates an adaptive group-specific gradient clipping strategy. Through extensive experimental evaluation across diverse datasets, we demonstrate that our algorithm successfully reduces group privacy risk disparities while preserving model utility. This research advances both empirical understanding and practical implementation of fair privacy protection in ML systems, contributing to the broader goal of responsible AI deployment.

References

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
- [2] Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. Nearly tight black-box auditing of differentially private machine learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. <https://openreview.net/forum?id=cCDMXXiamP>.
- [3] 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. Curran Associates Inc., 2019.
- [4] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.
- [5] Jiahao Ding, Xinyue Zhang, Xiaohuan Li, Junyi Wang, Rong Yu, and Miao Pan. Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 622–629, 2020.
- [6] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.
- [7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, volume 3, pages 265–284. Springer, 2006.
- [8] Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on fairness, accountability and transparency*, pages 35–47. PMLR, 2018.
- [9] Maria S. Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C Cresswell. Disparate impact in differential privacy from gradient misalignment. In *The Eleventh International Conference on Learning Representations*, 2023. <https://openreview.net/forum?id=qL0aerVteqbx>.
- [10] Changwu Huang, Zeqi Zhang, Bifei Mao, and Xin Yao. An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819, 2023.

- [11] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR, 2019.
- [12] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: how private is private sgd? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 22205–22216, 2020.
- [13] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.
- [14] Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [15] 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.
- [16] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.
- [17] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pages 866–882. IEEE, 2021.
- [18] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1631–1648, 2023.
- [19] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.
- [20] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.
- [21] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.
- [22] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [23] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 49268–49280, 2023.
- [24] Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. In *Advances in Neural Information Processing Systems*, volume 34, pages 27555–27565. Curran Associates Inc., 2021.
- [25] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9932–9939, 2021.
- [26] 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.
- [27] 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, 2021.

- [28] M Yaghini, B Kulynych, G Cherubin, M Veale, and C Troncoso. Disparate vulnerability to membership inference attacks. In *Proceedings on Privacy Enhancing Technologies*, number 1, pages 460–480, 2022.
- [29] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022.
- [30] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.
- [31] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.
- [32] Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. Bayesian estimation of differential privacy. In *International Conference on Machine Learning*, pages 40624–40636. PMLR, 2023.
- [33] 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.

A Supplementary Definitions and Algorithms

A.1 DP-SGD

In our work, we concentrate on DP-SGD to uphold model privacy. Building upon SGD, the fundamental method for training a model f with parameters w by minimizing the empirical loss function $\ell(\hat{y}, y)$ for prediction \hat{y} and label y , DP-SGD (as illustrated in Algo. 2) integrates gradient clipping and noise addition for achieving the (ϵ, δ) -DP guarantees. In Algo. 2, during each epoch, per-sample gradients g_i are computed (Line 5). Since these gradients typically have unbounded sensitivity, they are clipped to ensure their norm does not exceed the hyperparameter C (Line 6). The clipped gradients are then aggregated and Gaussian noise is added to yield \tilde{g} (Line 8). \tilde{g} is subsequently scaled by the learning rate η and utilized for parameter update (Line 9). The final accumulated (ϵ, δ) , which is calculated by Rényi differential privacy (RDP) [16] and the moment accounting mechanism proposed by [1], quantifies the privacy protection ability.

Algorithm 2 DP-SGD [1]

Input: Training dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the parameterized model $f_w(\cdot)$, loss function $\ell(\hat{y}, y)$, iterations T , batch size b , learning rate η , noise scale σ , clipping bound C .

```

1: Initialize  $w^{(0)}$  randomly.
2: for  $t = 0, \dots, T - 1$  do
3:   Sample a batch  $B$  from  $D$  with probability  $b/N$ .
4:   for  $i \in B$  do
5:      $g_i \leftarrow \nabla \ell(f_{w^{(t)}}(\mathbf{x}_i), y_i)$ 
6:      $\bar{g}_i \leftarrow g_i \cdot \min(1, \frac{C}{\|g_i\|_2})$ 
7:   end for
8:    $\tilde{g} \leftarrow \frac{1}{b} (\sum_{i \in B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}))$ 
9:    $w^{(t+1)} \leftarrow w^{(t)} - \eta \tilde{g}$ 
10: end for

```

Return: Model $f_{w^{(T)}}(\cdot)$ and accumulated (ϵ, δ) .

A.2 Definitions of Membership Inference Games

Definition 6 (Average-case Membership Inference Game [30]). Let Ω denotes the underlying population data pool, \mathcal{M} the training algorithm, and \mathcal{A} the inference algorithm. We assume that the challenger samples n i.i.d. records from Ω to construct the training dataset D .

- 1) The challenger trains a target model $f \leftarrow \mathcal{M}(D)$.
- 2) The challenger randomly selects a record $z_0 \leftarrow \Omega$ and a record $z_1 \sim D$, ensuring that $z_0 \notin D$.
- 3) The challenger flips a fair coin $b \in \{0, 1\}$, and sends the target model and target record (f, z_b) to the adversary.
- 4) The adversary, with access to the target model, outputs a guess $\hat{b} \leftarrow \mathcal{A}(f, z_b)$.
- 5) The game outputs 1 (success) if $\hat{b} = b$, and 0 otherwise.

Definition 7 (Worst-case Membership Inference Game [29]).

- 1) The challenger samples a fixed record $z \sim D$, and trains a model $f_0 \leftarrow \mathcal{M}(D \setminus z)$.
- 2) The challenger trains a model $f_1 \leftarrow \mathcal{M}(D)$.
- 3) The challenger flips a fair coin $b \in \{0, 1\}$, and sends the target model and record (f_b, z) to the adversary.
- 4) The adversary, with access to the target model, outputs a guess $\hat{b} \leftarrow \mathcal{A}(f_b, z)$.
- 5) The game outputs 1 (success) if $\hat{b} = b$, and 0 otherwise.

A.3 Privacy Auditing by Different Attacks

Privacy auditing by average-case attacks. We introduce existing algorithms that use average-case attacks for privacy auditing (PA-ACAs). Specifically, one approach conducts privacy auditing using the global attack (PA-GA) [28], which is detailed in Algo. 3. In Algo. 3, a single threshold β is determined based on the overall behavior of all auditing samples. Another approach performs privacy auditing through the group-based attack (PA-GBA) [4], as described in Algo. 4. PA-GBA provides the adversary with background knowledge about which group g_i each data point (\mathbf{x}_i, y_i) belongs to. It then determines K thresholds β^k based on the behavior of all auditing samples within each group, where K represents the number of groups. As illustrated in Algo. 3 and Algo. 4, a single execution of the attack generates a prediction for the membership status of each data point in the auditing dataset. To evaluate the privacy risk associated with individual data points more comprehensively, researchers typically conduct multiple iterations of privacy auditing using average-case attacks [4]. Each iteration yields new results, which are aggregated to estimate the likelihood of a data point being accurately identified as either a member or a non-member.

Algorithm 3 PA-GA [28]

Input: Training dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, auditing dataset $Z = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^m$, loss function $\ell(\hat{y}, y)$, optimal threshold β .

- 1: Initialize outputs $O \leftarrow []$, membership status $H \leftarrow []$, and membership guesses $G \leftarrow []$.
- 2: $f \leftarrow \mathcal{M}(D)$
- 3: **for** $i = 1, \dots, m$ **do**
- 4: $O[i] \leftarrow \ell(f(\mathbf{x}_i), y_i)$
- 5: $H[i] \leftarrow \begin{cases} 1 & \text{if } z_i \in D \\ 0 & \text{otherwise} \end{cases}$
- 6: **end for**
- 7: $G \leftarrow [1\{O[i] \geq \beta\} \text{ for } i = 1, \dots, m]$.

Return: Membership status H and guesses G .

Algorithm 4 PA-GBA [4]

Input: Training dataset $D = \{(\mathbf{x}_i, y_i, g_i)\}_{i=1}^n$, auditing dataset $Z = \{z_i = (\mathbf{x}_i, y_i, g_i)\}_{i=1}^m$, loss function $\ell(\hat{y}, y)$, optimal threshold $\{\beta^k\}_{k=1}^K$.

- 1: Initialize outputs $O \leftarrow []$, membership status $H \leftarrow []$, and membership guesses $G \leftarrow []$.
- 2: $f \leftarrow \mathcal{M}(D)$
- 3: **for** $i = 1, \dots, m$ **do**
- 4: $O[i] \leftarrow \ell(f(\mathbf{x}_i), y_i)$
- 5: $H[i] \leftarrow \begin{cases} 1 & \text{if } z_i \in D \\ 0 & \text{otherwise} \end{cases}$
- 6: **end for**
- 7: $G \leftarrow [1\{O[i] \geq \beta^{g_i}\} \text{ for } i = 1, \dots, m]$.

Return: Membership status H and guesses G .

Privacy auditing by LOOA. In recent years, numerous studies have focused on using privacy auditing to evaluate the differential privacy (DP) guarantees of the DP-SGD algorithm [18, 23, 32, 12, 2, 17]. These studies aim to bridge the gap between theoretical guarantees and practical performance, offering empirical insights into the actual privacy leakage in real-world deployments. A common approach in these studies is Privacy Auditing via the Leave-One-Out Attack (PA-LOOA), as outlined in Algo. 5. The algorithm iteratively assesses the impact of including or excluding a specific data record z —often crafted as a worst-case scenario for auditing DP-SGD—within the training dataset D (Lines 2–8). For each repetition, the framework trains two models: f_0 , using the modified dataset $D \setminus z$, and f_1 , using the original dataset D (Lines 3–4). The outputs of these models are recorded, and the membership status of the data record is tracked (Lines 5–7). Based on these outputs and the membership status, attack scores are computed to estimate the likelihood of the record’s inclusion (Line 9). Finally, assuming an optimal adversary conducting the attack, an optimal threshold is applied to infer whether the record z was part of the training dataset (Line 10).

In our work, we focus on evaluating the empirical privacy leakage risk of each individual data record within the training dataset, rather than the worst guarantees of a mechanism in DP auditing studies.

Algorithm 5 PA-LOOA [18, 23, 32, 12, 2, 17]

Input: Training dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, auditing data record $z = (\mathbf{x}, y)$, loss function $\ell(\hat{y}, y)$, number of repetitions R , optimal threshold β .

- 1: Initialize outputs $O \leftarrow []$, membership status $H \leftarrow []$, and membership guesses $G \leftarrow []$.
- 2: **for** $r = 1, \dots, R$ **do**
- 3: $f_0 \leftarrow \mathcal{M}(D \setminus z)$
- 4: $f_1 \leftarrow \mathcal{M}(D)$
- 5: $O[2r - 1] \leftarrow \ell(f_0(\mathbf{x}), y)$
- 6: $O[2r] \leftarrow \ell(f_1(\mathbf{x}), y)$
- 7: $H \leftarrow H + [0, 1]$
- 8: **end for**
- 9: $G \leftarrow [1\{O[r] \geq \beta\} \text{ for } r = 1, \dots, 2R]$.

Return: Membership status H and guesses G .

Privacy auditing by ALOOA. In each iteration, m audit samples are randomly and independently assigned inclusion or exclusion statuses for training (Line 3). Based on this membership status set, the training dataset $D \setminus \{z_i \mid h_i = 1\}$ is constructed, and a model f_1 is trained. Similarly, a model f_0 is trained using the inverse of this state set (Lines 4–5). The membership states for each audit record, indicating whether it was used in training, are then recorded (Lines 6–7). Subsequently, the output for each audit sample is logged (Lines 8–11). Based on these outputs and the true membership statuses, membership states are inferred through the outputs and an optimal threshold (Line 13–15).

Algorithm 6 PA-ALOOA

Input: Training dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, Auditing dataset $Z = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^m$, loss function $\ell(\hat{y}, y)$, number of repetitions R , optimal thresholds $\{\beta_i\}_{i=1}^m$.

- 1: Initialize outputs $O \leftarrow []$, membership status $H \leftarrow []$, and membership guesses $G \leftarrow []$.
- 2: **for** $r = 1, \dots, R$ **do**
- 3: Randomly generate membership statuses $\{h_i\}_{i=1}^m$, where $h_i \in \{0, 1\}$ for each z_i .
- 4: $f_0 \leftarrow \mathcal{M}(D \setminus \{z_i \mid h_i = 0\})$
- 5: $f_1 \leftarrow \mathcal{M}(D \setminus \{z_i \mid h_i = 1\})$
- 6: $H[2r - 1] \leftarrow \{h_i\}_{i=1}^m$
- 7: $H[2r] \leftarrow \{\sim h_i\}_{i=1}^m$
- 8: **for** $i = 1, \dots, m$ **do**
- 9: $O[2r - 1][i] \leftarrow \ell(f_0(\mathbf{x}_i), y_i)$
- 10: $O[2r][i] \leftarrow \ell(f_1(\mathbf{x}_i), y_i)$
- 11: **end for**
- 12: **end for**
- 13: **for** $i = 1, \dots, m$ **do**
- 14: $G[i] \leftarrow [1\{O[r][i] \geq \beta_i\} \text{ for } r = 1, \dots, 2R]$.
- 15: **end for**

Return: Membership status H and guesses G .

B Comparison with PA-LOOA

We conduct experiments using the MNIST dataset and CNN models trained with the SGD optimizer. Detailed hyperparameter settings are provided in App. D.1. In the main paper, we set $m = 600$ for both PA-LOOA and PA-ALOOA. Here, we further evaluate a different configuration: $m = 600$ for PA-LOOA and $m = n$ for PA-ALOOA, where $m = n$ better reflects realistic auditing scenarios.

As shown in Fig. 7, the results under this setting are consistent with those reported in the main paper. PA-ALOOA maintains comparable auditing performance while significantly reducing computational cost. Moreover, it remains effective and reliable for measuring group-level privacy risks, further supporting its applicability in real-world deployments.

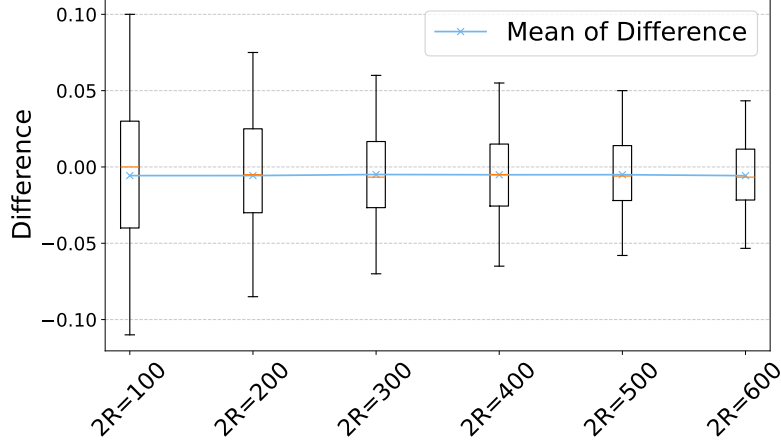
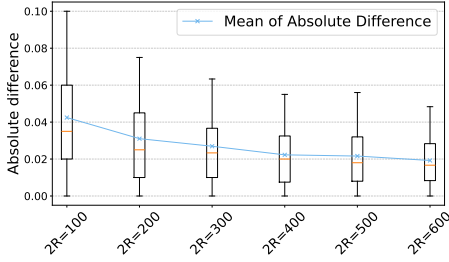
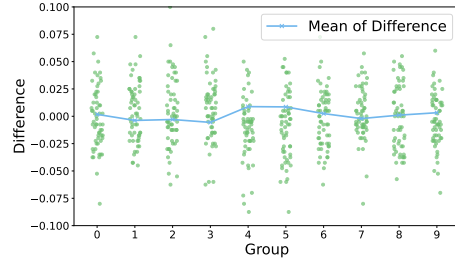


Figure 6: The horizontal axis represents the number of random experiments for a single audit, while the vertical axis represents the signed difference in auditing performance between the two attacks for each audited sample.



(a) Comparison at the individual level.



(b) Comparison at the group level.

Figure 7: Left: The horizontal axis represents the number of random experiments for a single audit, while the vertical axis represents the absolute difference in auditing performance between the two attacks for each audited sample. Right: The horizontal axis represents different groups of MNIST dataset, while the vertical axis indicates the performance difference between PA-LOOA and PA-ALOOA for individual data points in each group at $2R = 400$.

C Comparison with PA-ACAs

We provide additional results on six datasets: MNIST, Adult, Bank, Credit, Law, and UTKFace. We use both standard SGD and DP-SGD training algorithms. The models employed include LR, MLP, and CNN. Detailed experimental settings are provided in App. D.1. The complete results are shown in Figs. 8, 9, and 10, respectively. As clearly shown in the figures, the conclusions remain consistent with those discussed in the main text.

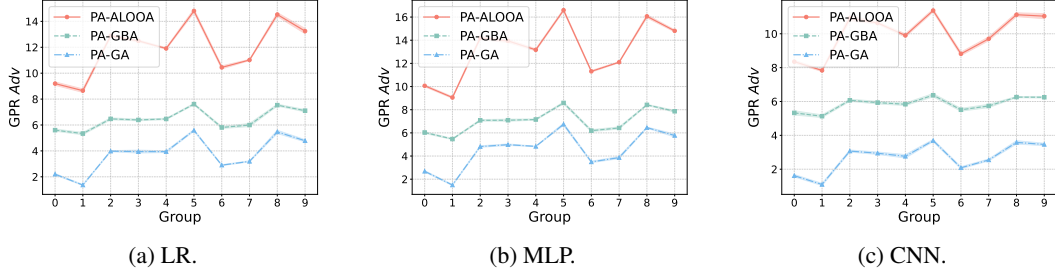


Figure 8: The comparison of GPR value across three model types—Logistic Regression (LR), Multilayer Perceptron (MLP), and CNN—trained on the MNIST dataset using DP-SGD algorithm with $\epsilon = 10$. The x-axis represents the groups, and the y-axis shows the corresponding GPR value at $2R = 400$.

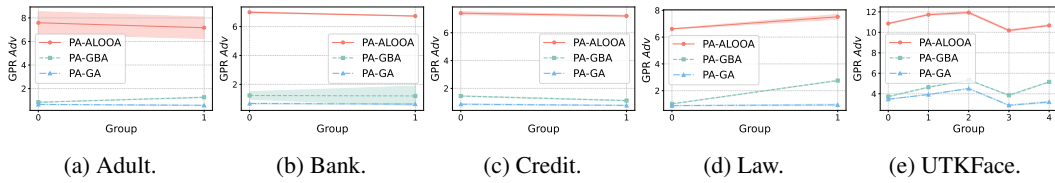


Figure 9: The comparison of GPR value across three model types—LR, MLP, and CNN—trained on the fairness-related datasets using SGD algorithm. The x-axis represents the groups, and the y-axis shows the corresponding GPR value at $2R = 400$.

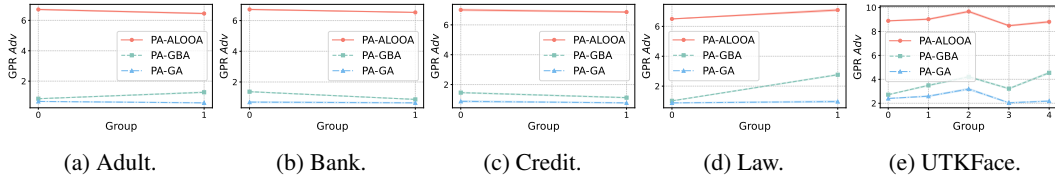


Figure 10: The comparison of GPR value across three model types—LR, MLP, and CNN—trained on the fairness-related datasets using DP-SGD algorithm with $\epsilon = 10$. The x-axis represents the groups, and the y-axis shows the corresponding GPR value at $2R = 400$.

D Experimental Details and Results

D.1 Details of Experimental Setup

Datasets. The MNIST dataset [14] comprises 60,000 training and 10,000 testing samples, with each sample being a 28×28 grayscale image of a handwritten digit from 0 to 9, spanning ten classes. Due to computational limitations and to improve training efficiency, we randomly select 1,000 samples per class from the original dataset, resulting in a balanced dataset of 10,000 samples used in our experiments. In our study, we treat the classification label (i.e., the digit class) as a proxy for the demographic group to analyze group-specific privacy risks.

For the tabular fairness-related datasets, Adult, Bank, Credit, and Law [13]³, the detailed information is shown in Tab. 3. For the image-based fairness-related dataset, UTKFace [33], we conduct evaluation after data cleaning and preprocessing.⁴ The final dataset consists of 27,305 grayscale facial images of size $1 \times 48 \times 48$. In our experiments, we treat ethnicity as the protected attribute and gender as the

³Dataset can be download from https://github.com/tailequy/fairness_dataset/tree/main/experiments/data

⁴Dataset can be download from https://www.kaggle.com/datasets/nipunarora8/age-gender-and-ethnicity-face-data-csv/data?select=age_gender.csv

prediction label. The ethnicity attribute includes five classes, with the number of samples per class being 10,078; 4,526; 3,434; 3,975; and 1,692, respectively.

Table 3: The information of experimental datasets. Here, 0 represents the advantaged group, and 1 represents the disadvantaged group.

Dataset	#Instances(cleaned)	Class ratio(0 : 1)	Sensitive attribute
Adult	45,222	2.09 : 1	Gender
Bank	40,004	2.13 : 1	Marital
Credit	30,000	1.52 : 1	Sex
Law	20,798	5.29 : 1	Race

Training algorithms. Our study compares three training algorithms: standard SGD, DP-SGD, and our proposed DP-SGD-S. DPML algorithms (i.e., DP-SGD and DP-SGD-S) are implemented using the Opacus library [31]. For the SGD algorithm, all datasets are trained using the SGD optimizer with a learning rate of 0.1. For DPML algorithms, the privacy hyperparameters (ϵ , δ) are configured with $(10, 1e-5)$ and $(1, 1e-5)$. For DP-SGD-S, the default scale bound τ is set to 2. When $\epsilon = 10$, all datasets use the SGD optimizer with a learning rate of 0.1, and the clipping bound is set to 10. When $\epsilon = 1$, the tabular datasets use the same optimizer and learning rate as above, while the image datasets are trained using the Adam optimizer with a learning rate of 0.005. In this case, the gradient clipping bound is set to 5 for all datasets. Across all experiments, we use a batch size of 256 and train for 20 epochs.

Models. For the MNIST dataset, we employ three types of models for training: Logistic Regression (LR), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN). For the tabular datasets, we use the LR model. For the UTKFace dataset, we adopt the CNN model.

The LR model consists of a single fully connected layer that directly maps the input features to the output classes, without any hidden layers or activation functions. The MLP model includes one hidden layer with 256 neurons and uses the tanh activation function. The input is first flattened and passed through the hidden layer, followed by an output layer. The CNN model comprises two convolutional layers followed by two fully connected layers. The first convolutional layer uses 16 filters of size 5×5 , followed by a 2×2 max pooling layer. The second convolutional layer has 32 filters of size 4×4 , also followed by max pooling. After flattening the resulting feature maps, the output is passed through a fully connected layer with 32 neurons and then through the final classification layer. The tanh activation function is applied throughout the network.

Evaluation metrics. We use the PA-ALOOA method to obtain the results of GPR and GPRP metrics. We set $2R = 400$ and $n = m$, which we believe is a reasonable configuration, as analyzed in Sec. 3. We use accuracy to measure model prediction performance. Specifically, the accuracy of the training algorithms is computed by splitting the datasets into 80% for training and 20% for testing.

Experimental Testbed. All our experiments are conducted on a cluster equipped with 10 NVIDIA A100 GPUs, 128 CPU cores, and 540 GB of RAM. One privacy auditing procedure with $2R = 400$ took approximately 5 hours to complete on the MNIST dataset using a CNN model under DP-SGD-S.

D.2 Supplementary Experimental Results

D.2.1 Comparison between GPR and GRC

As shown in Fig. 11, there is a significant correlation between GRC and GPR across different models. Specifically, groups contributing more during training exhibit higher privacy leakage risks, with this phenomenon being more pronounced in simpler model architectures.

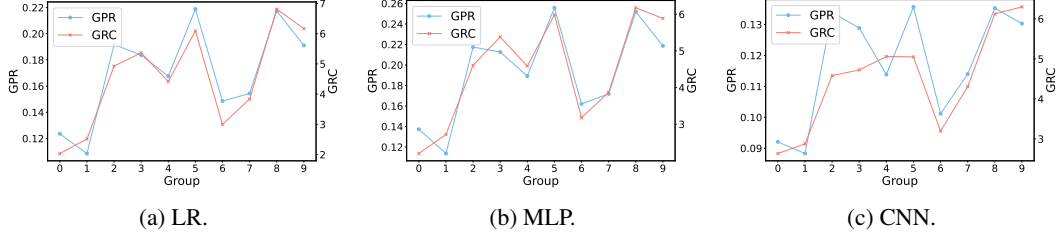


Figure 11: The values of GPR and GRC across three models—LR, MLP, and CNN—trained on the MNIST dataset using SGD training algorithm. Each subfigure shows the GPR and GRC for different groups, with the left y-axis indicating GPR, the right y-axis indicating GRC, and the x-axis representing group. All values are reported at $2R = 400$.

D.2.2 Comparison between three training algorithms

The complete results are presented in Tabs. 4, 5, and 6. These tables provide a comprehensive comparison of algorithm performance across multiple datasets and privacy guarantees. As shown, the observed trends and conclusions are consistent with those discussed in the main text. However, there is one notable exception. Specifically, on the UTKFace dataset with $\epsilon = 1$, our algorithm DP-SGD-S does not successfully reduce the disparity in group privacy risks. Despite this isolated case, DP-SGD-S consistently improves privacy fairness across all other experimental settings. These results underscore the robustness of DP-SGD-S in promoting group-level privacy fairness under most conditions.

Additionally, the Tab. 6 clearly shows that simpler model architectures, such as LR and MLP, exhibit relatively large Δ values. In contrast, more complex architecture CNN leads to a significant reduction in Δ . This suggests that, when deploying models on public platforms, adopting more complex architectures may help reduce the disparity of privacy protection across groups and should be considered as a practical design choice.

Table 4: The performance of different algorithms across six datasets under $\epsilon = 10$.

Metric	Method	Adult	Bank	Credit	Law	UTKFace
Accuracy (\uparrow)	SGD	85.00 ± 0.07	89.93 ± 0.06	81.80 ± 0.04	89.75 ± 0.12	85.93 ± 3.79
	DP-SGD	84.92 ± 0.04	89.94 ± 0.04	81.83 ± 0.07	89.74 ± 0.08	86.84 ± 0.48
	DP-SGD-S	84.86 ± 0.11	89.91 ± 0.09	81.83 ± 0.16	89.56 ± 0.18	84.67 ± 0.21
GPRP Δ (\downarrow)	SGD	0.42 ± 0.04	0.26 ± 0.05	0.20 ± 0.07	0.90 ± 0.16	1.75 ± 0.11
	DP-SGD	0.27 ± 0.04	0.19 ± 0.03	0.14 ± 0.05	0.59 ± 0.06	1.19 ± 0.07
	DP-SGD-S	0.17 ± 0.02	0.13 ± 0.03	0.08 ± 0.04	0.41 ± 0.07	0.59 ± 0.07

Table 5: The performance of different algorithms across six datasets under $\epsilon = 1$.

Metric	Method	Adult	Bank	Credit	Law	UTKFace
Accuracy (\uparrow)	DP-SGD	84.84 ± 0.07	89.97 ± 0.04	81.73 ± 0.09	89.60 ± 0.15	81.60 ± 0.13
	DP-SGD-S	84.64 ± 0.18	89.81 ± 0.20	81.62 ± 0.17	89.59 ± 0.12	81.43 ± 0.91
GPRP Δ (\downarrow)	DP-SGD	0.20 ± 0.03	0.14 ± 0.03	0.08 ± 0.02	0.39 ± 0.04	0.24 ± 0.02
	DP-SGD-S	0.11 ± 0.02	0.09 ± 0.03	0.03 ± 0.02	0.21 ± 0.06	0.27 ± 0.10

Table 6: The performance of three algorithms across various model types under different theoretical privacy budgets, trained on MNIST dataset. The results include model prediction performance, GPRP value, and empirical privacy budget estimates.

Model	ϵ	Method	Accuracy (\uparrow)	GPRP $\Delta(\downarrow)$
LR	/	SGD	90.25 ± 0.23	11.13 ± 0.15
		DP-SGD	89.07 ± 0.30	6.16 ± 0.19
		DP-SGD-S	87.37 ± 0.41	4.87 ± 0.12
	10	DP-SGD	85.42 ± 0.37	2.30 ± 0.16
		DP-SGD-S	84.09 ± 0.44	2.12 ± 0.14
	1	DP-SGD	85.42 ± 0.37	2.30 ± 0.16
MLP	/	SGD	92.86 ± 0.28	14.17 ± 0.09
		DP-SGD	90.31 ± 0.41	7.54 ± 0.12
		DP-SGD-S	87.56 ± 0.41	5.89 ± 0.06
	10	DP-SGD	84.94 ± 0.94	2.73 ± 0.13
		DP-SGD-S	83.95 ± 0.80	2.30 ± 0.12
	1	DP-SGD	84.94 ± 0.94	2.73 ± 0.13
CNN	/	SGD	95.89 ± 0.29	4.92 ± 0.18
		DP-SGD	94.46 ± 0.13	3.54 ± 0.13
		DP-SGD-S	92.63 ± 0.58	2.91 ± 0.14
	10	DP-SGD	89.06 ± 0.72	1.83 ± 0.09
		DP-SGD-S	88.58 ± 0.44	1.55 ± 0.14
	1	DP-SGD	89.06 ± 0.72	1.83 ± 0.09

D.2.3 The results of each group privacy risk

We provide a comprehensive analysis of group-level privacy risks across all experimental settings, as illustrated in Figs. 12, 13, and 14. As observed from these figures, there are 16 subplots in total, each corresponding to a different experimental configuration. Among them, only one case shows that DP-SGD-S increases the privacy leakage risk for a specific group, which occurs on the UTKFace dataset with $\epsilon = 1$. This result demonstrates that improving privacy fairness does not compromise the protection of already well-protected groups. As a result, our method avoids the undesirable “leveling down” effect and ensures more responsible and practical deployment in real-world scenarios.

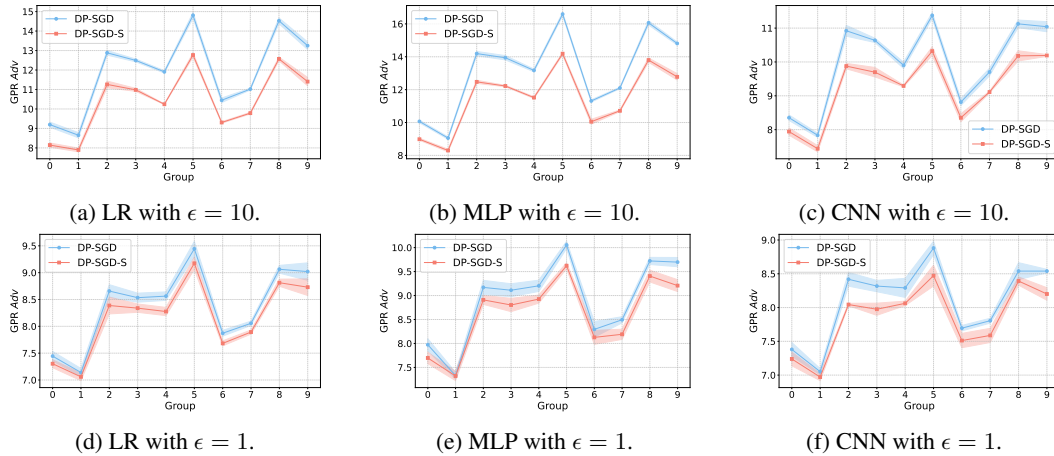


Figure 12: The GPR *Adv* of each group across three models and different privacy budgets, trained on MNIST dataset. In each subfigure, the vertical axis represents GPR value at $2R = 400$.

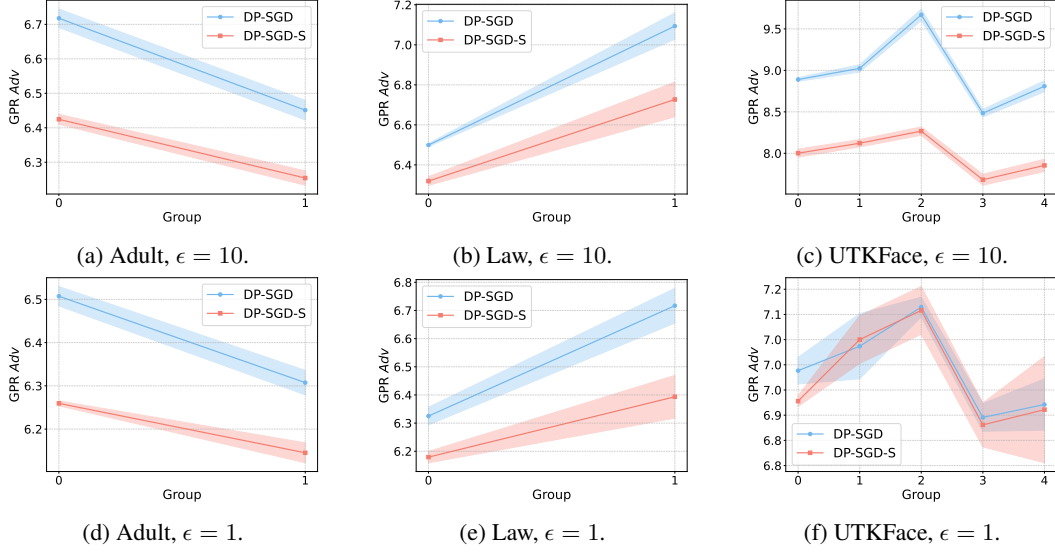


Figure 13: The GPR *Adv* of each group across different datasets and privacy budgets. In each subfigure, the vertical axis represents GPR value at $2R = 400$.

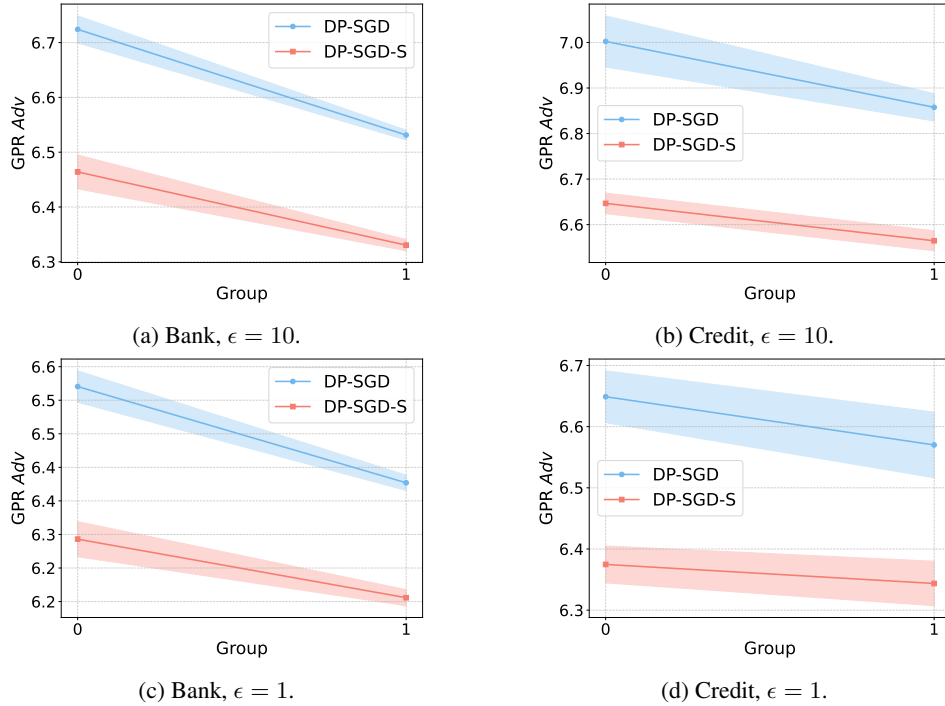


Figure 14: The GPR *Adv* of each group across different datasets and privacy budgets. In each subfigure, the vertical axis represents GPR value at $2R = 400$.

D.2.4 The results of outcome fairness

Traditional definitions of outcome fairness are primarily designed for binary classification tasks where the sensitive attribute has only two groups (e.g., male vs. female) [19]. Under this setting, fairness metrics are typically computed using the absolute difference between the values of the two groups. However, in the case of the UTKFace dataset, the sensitive attribute is multi-class, containing more than two categories. To accommodate this, we adopt a natural extension of the conventional

definitions: instead of calculating the absolute difference between two groups, we compute the range across all subgroups, i.e., the difference between the maximum and minimum values of the fairness metric across the group set. This allows us to quantify disparities in outcomes among multiple demographic groups.

The exact formulas used to compute outcome fairness measurements are as follows. Here, s represents the sensitive attribute, and K denotes the set of all subgroups:

- Accuracy Parity (AP): the maximum disparity in prediction accuracy among different groups:

$$\text{AP} = \max_{k \in K} (P(\hat{y} = y \mid s = s_k)) - \min_{k \in K} (P(\hat{y} = y \mid s = s_k))$$

- Demographic Parity (DmP): the maximum disparity in the rate of positive predictions across groups:

$$\text{DmP} = \max_{k \in K} (P(\hat{y} = 1 \mid s = s_k)) - \min_{k \in K} (P(\hat{y} = 1 \mid s = s_k))$$

- Equal Opportunity (EOp): the maximum gap in true positive rates among groups:

$$\text{EOp} = \max_{k \in K} (P(\hat{y} = 1 \mid s = s_k, y = 1)) - \min_{k \in K} (P(\hat{y} = 1 \mid s = s_k, y = 1))$$

- Equalized Odds (EOd): the maximum divergence in both true positive and false positive rates among groups:

$$\begin{aligned} \text{EOd} = & \max_{k \in K} (P(\hat{y} = 1 \mid s = s_k, y = 1) + P(\hat{y} = 1 \mid s = s_k, y = 0)) \\ & - \min_{k \in K} (P(\hat{y} = 1 \mid s = s_k, y = 1) + P(\hat{y} = 1 \mid s = s_k, y = 0)) \end{aligned}$$

The experimental results for outcome fairness metrics of three algorithms, evaluated under varying theoretical privacy budgets on fairness-related datasets, are presented in Tab. 7. From the table, we can observe that DP-SGD-S and DP-SGD perform similarly, with no clear superiority of one over the other across these metrics.

Table 7: The outcome fairness measurements of three algorithms under different theoretical privacy budgets, including AP, DmP, EOp, and EOd.

Dataset	ϵ	Method	AP	DmP	EOp	EOd
Adult	/	SGD	0.1148 ± 0.0012	0.1845 ± 0.0035	0.1103 ± 0.0058	0.1831 ± 0.0049
		DP-SGD	0.1175 ± 0.0015	0.1886 ± 0.0045	0.1078 ± 0.0043	0.1853 ± 0.0045
		DP-SGD-S	0.1160 ± 0.0022	0.1877 ± 0.0036	0.1045 ± 0.0138	0.1811 ± 0.0159
	10	DP-SGD	0.1171 ± 0.0013	0.1889 ± 0.0067	0.1053 ± 0.0259	0.1834 ± 0.0294
		DP-SGD-S	0.1193 ± 0.0021	0.1925 ± 0.0006	0.1294 ± 0.0097	0.2107 ± 0.0093
	1	DP-SGD	0.1171 ± 0.0013	0.1889 ± 0.0067	0.1053 ± 0.0259	0.1834 ± 0.0294
Bank	/	SGD	0.0283 ± 0.0012	0.0401 ± 0.0044	0.0881 ± 0.0090	0.1034 ± 0.0112
		DP-SGD	0.0283 ± 0.0013	0.0406 ± 0.0040	0.0927 ± 0.0175	0.1083 ± 0.0194
		DP-SGD-S	0.0272 ± 0.0022	0.0435 ± 0.0065	0.1179 ± 0.0219	0.1344 ± 0.0253
	10	DP-SGD	0.0279 ± 0.0008	0.0402 ± 0.0024	0.0900 ± 0.0106	0.1052 ± 0.0119
		DP-SGD-S	0.0300 ± 0.0032	0.0397 ± 0.0044	0.0902 ± 0.0161	0.1062 ± 0.0169
	1	DP-SGD	0.0279 ± 0.0008	0.0402 ± 0.0024	0.0900 ± 0.0106	0.1052 ± 0.0119
Credit	/	SGD	0.0181 ± 0.0009	0.0365 ± 0.0018	0.0576 ± 0.0050	0.0772 ± 0.0057
		DP-SGD	0.0185 ± 0.0011	0.0367 ± 0.0029	0.0569 ± 0.0048	0.0768 ± 0.0070
		DP-SGD-S	0.0214 ± 0.0018	0.0339 ± 0.0044	0.0454 ± 0.0096	0.0654 ± 0.0120
	10	DP-SGD	0.0213 ± 0.0012	0.0362 ± 0.0062	0.0502 ± 0.0134	0.0717 ± 0.0172
		DP-SGD-S	0.0202 ± 0.0030	0.0375 ± 0.0103	0.0570 ± 0.0226	0.0785 ± 0.0290
	1	DP-SGD	0.0213 ± 0.0012	0.0362 ± 0.0062	0.0502 ± 0.0134	0.0717 ± 0.0172
Law	/	SGD	0.1499 ± 0.0053	0.2136 ± 0.0165	0.1181 ± 0.0151	0.5262 ± 0.0265
		DP-SGD	0.1537 ± 0.0043	0.2161 ± 0.0145	0.1223 ± 0.0117	0.5306 ± 0.0329
		DP-SGD-S	0.1564 ± 0.0043	0.2248 ± 0.0153	0.1308 ± 0.0128	0.5407 ± 0.0333
	10	DP-SGD	0.1554 ± 0.0054	0.2140 ± 0.0177	0.1223 ± 0.0117	0.5231 ± 0.0413
		DP-SGD-S	0.1582 ± 0.0049	0.2252 ± 0.0202	0.1315 ± 0.0160	0.5581 ± 0.0469
	1	DP-SGD	0.1554 ± 0.0054	0.2140 ± 0.0177	0.1223 ± 0.0117	0.5231 ± 0.0413
UTKFace	/	SGD	0.0773 ± 0.0140	0.1548 ± 0.0301	0.0829 ± 0.0562	0.2008 ± 0.1091
		DP-SGD	0.0595 ± 0.0107	0.1597 ± 0.0156	0.0658 ± 0.0187	0.1528 ± 0.0378
		DP-SGD-S	0.0741 ± 0.0101	0.1565 ± 0.0129	0.0570 ± 0.0249	0.1676 ± 0.0350
	10	DP-SGD	0.0835 ± 0.0124	0.1973 ± 0.0147	0.1501 ± 0.0156	0.3408 ± 0.0400
		DP-SGD-S	0.0825 ± 0.0084	0.1868 ± 0.0223	0.1367 ± 0.0205	0.3280 ± 0.0274
	1	DP-SGD	0.0835 ± 0.0124	0.1973 ± 0.0147	0.1501 ± 0.0156	0.3408 ± 0.0400

E Broader Impact

This paper primarily aims to advance the fairness of privacy protection in machine learning. While differential privacy techniques provide formal privacy guarantees, they often overlook disparities in privacy protection across different groups. Our work addresses this gap by proposing methods that enhance the fairness of privacy protection at the group level.

By improving the equity of privacy protection, our research contributes to building more trustworthy and ethical machine learning systems, particularly in sensitive application domains where privacy concerns are paramount. This focus on privacy fairness helps mitigate potential harms caused by uneven privacy leakage, thereby supporting more responsible and inclusive AI deployment. Moreover, by rigorously evaluating group privacy risks and proposing a privacy fairness metric, this research fosters a deeper understanding of the inherent trade-offs, guiding practitioners and policymakers toward the responsible and equitable deployment of machine learning technologies.

We acknowledge that no method is without limitations, and further research is necessary to simultaneously address other aspects of fairness, such as outcome fairness. Nonetheless, we believe this research helps pave the way for the ethical and equitable deployment of AI systems.