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

Anonymous Authors[1]

Abstract

Although research on privacy and fairness in machine learning (ML) has made significant progress, the fairness of privacy protection across different subgroups remains underexplored. Experimental results show that existing ML and differentially private machine learning (DPML) algorithms exhibit unfair privacy leakage risks among various groups, highlighting the need for further research in this area. Current studies primarily assess the average-case privacy risk of individual data records, which can overestimate protections for certain data points and underestimate disparities between groups. To address this, we introduce a membership inference game that efficiently audits the worst-case privacy risk associated with individual data points, allowing for more effective measurement of unfairness by evaluating privacy risk parity across groups. Based on our assessment of group privacy risk parity, we propose a novel technique to mitigate inequities in privacy protection for DPML. Inspired by canaries in differential privacy auditing, we enhance existing DPML algorithms with an adaptive group-specific gradient clipping strategy. Extensive experiments demonstrate that our approach maintains the same privacy guarantees as traditional DPML algorithms while effectively reducing disparities in group privacy risks, and promoting the ethical and equitable deployment of AI systems.

1. Introduction

Artificial intelligence (AI), particularly machine learning (ML), is widely adopted across various sectors, augmenting or replacing human decision-making. However, its growing integration in critical domains like healthcare, finance, and judiciary has raised concerns, including data privacy breaches, biases, lack of explainability, security vulnerabilities etc. (Huang et al., 2023). Among these ethical issues and risks, privacy and fairness are pivotal for establishing trust in AI systems (Ekstrand et al., 2018). Regulations like the General Data Protection Regulation (GDPR) (Voigt & Von dem Bussche, 2017) underscore the vital need to protect individual privacy. Among the various privacy-preserving techniques, differential privacy (DP) (Dwork et al., 2006b) has become one of the most widely adopted methods in ML (Ponomareva et al., 2023). In parallel, fairness, supported by laws such as the Equal Credit Opportunity Act (ECOA) (Act, 2018), ensures that algorithms avoid bias against protected attributes like gender or race, thereby fostering trust, promoting justice, and ensuring non-discrimination in AI decision-making.

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 (Xu et al., 2019; Jagielski et al., 2019; Ding et al., 2020; Tran et al., 2021b; Lowy et al., 2023). Other works investigate how privacy mechanisms impact algorithm prediction fairness. For example, prior work has shown that incorporating DP into ML models can exacerbate disparities in accuracy across groups (Bagdasaryan et al., 2019). As a result, many studies have explored the reasons behind this privacy-induced unfairness and proposed methods to mitigate the disparities introduced by DP (Xu et al., 2021; Tran et al., 2021a; Esipova et al., 2023). However, whether AI systems can provide fair and equitable privacy to different subgroups remains an underexplored issue. As highlighted in (Ekstrand et al., 2018), this raises an essential yet under-investigated question:

Do AI systems offer fair or equitable privacy leakage risks across subgroups?

This question raises both ethical and practical concerns, as certain groups may face disproportionately higher privacy leakage risks, violating principles of fairness and equality. While some studies have empirically explored whether ML algorithms offer equal privacy protections across subgroups,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

their evaluation of subgroup privacy risks typically relies on averaging responses across multiple data points under membership inference attacks (MIAs) (Chang & Shokri, 2021; Yaghini et al., 2022). In other words, their attacks are formulated based on the membership inference game (MIG) introduced in (Yeom et al., 2018), which captures the average behavior across all data points. However, such evaluation methods may obscure the heterogeneity of individual privacy risks within subgroups, potentially overestimating the system's ability to protect certain data points. This overestimation of privacy protections for individual data records may, in turn, lead to an underestimation of differences in privacy risks between groups. Consequently, the issue of privacy inequality may not be adequately uncovered.

Therefore, to answer the question more rigorously, we aim to tightly audit the empirical privacy risk of individual data points within the training dataset and then analyze the differences in privacy risk across groups. By auditing such a worst-case privacy risk of individual data records, we strive to achieve a more precise identification of the privacy risks associated with specific data points, thereby encouraging greater caution in the deployment of ML models. To obtain the worst-case privacy leakage risk for each data point, we can use the worst-case MIG proposed by (Ye et al., 2022). However, the computational cost of the attacks formulated by this game, such as Leave-One-Out Attack (LOOA), is prohibitively high, making its practical implementation nearly infeasible. To address this challenge, we propose an approximate worst-case MIG to efficiently audit the 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; 2) The individual privacy risk evaluated by our method can be reliable for assessing group privacy risk.

After obtaining the worst-case individual privacy risk, we define a fairness metric to evaluate privacy risk equality across groups. Through experiments, we clearly demonstrate that our attack method has significant advantages over traditional average-case attacks under the same conditions. Specifically, our method reveals greater group privacy risk and more effectively captures privacy inequalities. Consistent with prior research (Chang & Shokri, 2021; Yaghini et al., 2022), 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.

Inspired by the design of canaries in DP auditing studies (Nasr et al., 2023; Annamalai & Cristofaro, 2024; Steinke et al., 2023), 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 contributions are as follows:

1. We propose a novel definition of MIG to efficiently and approximately audit the worst-case privacy risks of individual data points.
2. We define the group privacy risk parity metric to measure the degree of privacy unfairness and demonstrate that both ML and DPML algorithms exhibit varying levels of privacy unfairness.
3. We improve the existing DPML algorithm to alleviate inter-group privacy risk disparities, with experimental results validating the effectiveness of our approach.

2. Background

2.1. Differential Privacy

Differential Privacy (DP), proposed by (Dwork et al., 2006b), is a privacy framework designed to address privacy leakage. It has become the predominant method for ensuring algorithmic privacy (Ponomareva et al., 2023). In the following, we introduce the approximate (ϵ, δ)-DP definition.

Definition 2.1 ((ϵ, δ)-Differential Privacy (Dwork et al., 2006a))**.** 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 (Abadi et al., 2016) is a widely adopted algorithm in DPML (Ponomareva et al., 2023). 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. In black-box MIAs, the adversary analyzes the model's output behavior, typically using the sample's output loss as a score to infer membership, based on the observation that models tend to show smaller losses for training samples (Yeom et al., 2018).

Different definitions of membership inference games (MIGs). MIGs conceptualize MIAs as inference games between a privacy auditor (i.e., the adversary) and a challenger, with various definitions capturing different aspects of privacy loss (Ye et al., 2022).

Most MIAs follow the average-case MIG (Sablayrolles et al., 2019; Yeom et al., 2018; Zarifzadeh et al., 2024) (show in Def. A.1), which evaluate the vulnerability of a target model to the adversary, emphasizing the average behavior across data points. MIAs are performed by simulating the game through multiple iterations of the random experiment. While the global attack (GA) uses a single threshold for all data points (Sablayrolles et al., 2019; Yeom et al., 2018), the group-based attack (GBA) assigns separate thresholds for each group to better analyze group privacy risks within each iteration (Chang & Shokri, 2021).

However, such an average-case MIG cannot adequately or accurately assess the worst-case privacy risks associated with individual data records. To address this, a worst-case MIG definition has been proposed (Ye et al., 2022) (show in Def.A.2). The Leave-One-Out Attack (LOOA) is an instance of this game, representing an idealized scenario in privacy auditing. It independently evaluates the worst-case privacy risk of each data record, closely aligning with DP principles(Ye et al., 2022). However, evaluating the empirical privacy leakage risk of each data point within this framework is highly time-consuming.

Privacy auditing 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. Privacy auditing using the GA, GBA, and LOOA algorithms (abbreviated as PA-GA, PA-GBA, and PA-LOOA) are presented in Algo. 3, 4,and 5 of Appendix A.3, respectively.

2.3. The Fairness of Privacy

Few studies have explored fairness in privacy protection, with notable differences in their measuring approaches and areas of focus. For instance, a study using PA-GA for evaluation highlights that fairness-aware algorithms can exacerbate privacy risk disparities across groups (Chang & Shokri, 2021) Meanwhile, research employing PA-GBA demonstrates that DPML algorithms can reduce inter-group privacy risk disparities (Yaghini et al., 2022).

In our work, we focus on more precisely measuring and then mitigating the unfairness in privacy protection capabilities. Prior studies (Yaghini et al., 2022; Chang & Shokri, 2021) rely on attacks defined in Def. A.1, which assess privacy risks based on the average-case behavior of data points. While this approach effectively captures overall trends, it fails to account for the nuanced privacy leakage risks faced by individual data samples, potentially masking disparities between groups.

3. Measuring Disparity of Group Privacy Risk with Approximate Worst-case Privacy Auditing

We propose an alternative to PA-LOOA and show it achieves comparable performance. We define a fairness metric to assess inter-group privacy disparities and demonstrate that our method effectively captures inter-group privacy inequalities to the greatest extent.

3.1. Approximate Leave-One-Out Attack

As previously discussed, the attacks under Def A.2 (i.e., LOOA) are computationally expensive. Specifically, obtaining statistically reliable results for a single sample typically requires $2R$ repeated experiments (where R is generally set to at least 50). 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. 3.1.

Definition 3.1 (Approximate Worst-case Membership Inference Game)**.** 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 and the population data pool, 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). To facilitate understanding, Fig. 1 provides a straightforward comparison of the internal mechanisms of LOOA and ALOOA.

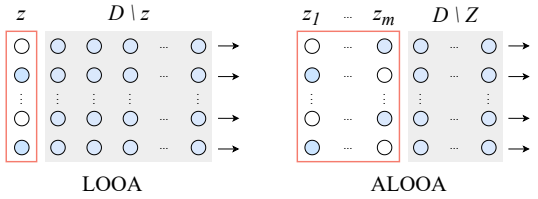


Figure 1: The left shows the LOOA mechanism for auditing a single sample, while the right depicts ALOOA for auditing m samples. Each circle represents a data point: solid circles indicate data points used for training in that iteration, while hollow circles indicate excluded points. The arrow represents the model training process using the solid-circle data points.

The process of Privacy Auditing using the Approximate Leave-One-Out Attack (PA-ALOOA) is detailed in Algo. 6. The algorithm strikes an effective balance between computational efficiency and analytical granularity. By evaluating multiple samples in parallel, it significantly reduces the computational burden compared to LOOA, while maintaining the fine-grained analysis necessary to uncover the specific vulnerabilities of individual data points.

3.2. Comparison with Leave-One-Out Attack

From the perspective of the design of the two attacks, the performance difference between the two attacks stems from their sample selection strategies. In the PA-LOOA, the randomness in the training set is minimal, as only one sample is randomly included or excluded in each iteration. In contrast, the PA-ALOOA audits m samples simultaneously, with each experiment randomly choosing which samples will be included in the training set, thus introducing greater randomness. We argue that with sufficient iterations, the random fluctuations in PA-ALOOA will average out, resulting in performance that is comparable to PA-LOOA.

We validate our hypothesis through practical experiments using the widely used MNIST dataset in DP auditing studies, training a Convolutional Neural Network (CNN) with SGD. We perform uniform sampling from each subgroup in the dataset, according to the class labels, to generate a subset of $n = 10,000$ data points. Due to computational limitations, for the PA-LOOA, we randomly select 60 samples per class, totaling 600 samples to audit. For the PA-ALOOA, we consider the case where the number of audit samples m equals n, representing a more common real-world scenario where the privacy leakage risk of every training sample is audited. We evaluate the attacker's performance using the accuracy metric from (Shokri et al., 2017), 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. Experimental results (Fig. 2) show that as the number of repeated experiments increases, the difference between the two attack methods diminishes and stabilizes. Additionally, as seen in Fig. 5, the two attacks perform comparably, since the data points are evenly distributed along the line $y = x$.

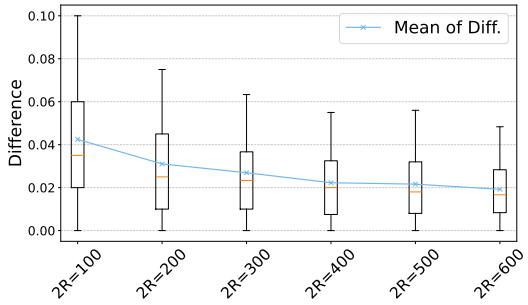


Figure 2: The horizontal axis represents the number of repeated experiments for a single audit, while the vertical axis represents the absolute difference in auditing performance between the two attacks for individual samples.

However, it is evident from Fig. 2 that the difference between the two attacks for different data points exhibits notable variance. We hypothesize that this variance may be due to two factors. First, the heterogeneity of individual data points, where the privacy leakage risk for some data points is more influenced by other training samples, causing variations in attack outcomes. Second, the inherent uncertainty in measuring privacy leakage risk for certain samples, which results in fluctuating audit results even when using the same attack method and repeating experiments. Regardless of the cause, this variance makes it unreliable to compare the privacy leakage risk of different individual samples.

Although the estimation errors for individual samples may vary significantly, we find that statistical outcomes across subgroups are reliable, forming a solid foundation for analyzing group privacy risk. As shown in Fig. 3, 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 difference in performance between the two attacks across groups. Furthermore, the average performance difference between PA-LOOA and PA-ALOOA

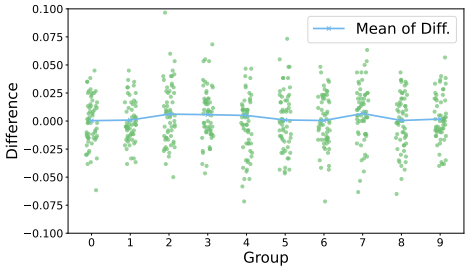


Figure 3: The horizontal axis represents different subgroups within the MNIST dataset, while the vertical axis indicates the performance difference between PA-LOOA and PA-ALOOA for individual data points at $2R = 600$.

across groups is less than 0.01, suggesting negligible error when extending individual sample estimates to group-level statistical results.

3.3. Definition of Group Privacy Risk Parity

Instead of using the membership advantage of the attacker to assess the overall dataset (Yaghini et al., 2022; Yeom et al., 2018), we evaluate the privacy leakage risk of a specific data point (Chang & Shokri, 2021).

Definition 3.2 (Individual Privacy Risk (IPR))**.** Let $Acc_i(\mathcal{A}, D)$ represent the attack accuracy of an individual i under privacy auditing algorithm $\mathcal{A}$ and the auditing dataset Z. The individual privacy risk is defined as:

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

It quantifies the normalized advantage of the adversary compared to random guessing. Building on the concept of IPR, we can extend it to define Group Privacy Risk as in (Chang & Shokri, 2021).

Definition 3.3 (Group Privacy Risk (GPR))**.** Let D^k denote the subset of the dataset D belonging to group k. The group privacy risk is defined as:

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

Building on the concept of GPR, we can assess whether the privacy leakage risk is fair and equitable across different groups.

Definition 3.4 (Group Privacy Risk Parity (GPRP))**.** Let K represent the set of all groups. We define group privacy risk parity as:

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

3.4. 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 (Yaghini et al., 2022) and PA-GBA (Chang & Shokri, 2021).

To ensure a fair comparison, we configure all three privacy auditing methods under identical conditions, keeping the training dataset and model consistent across each repeated experiment r. The key distinction lies in threshold determination: PA-ACAs computes thresholds based on the aggregate behavior of multiple data points, while PA-ALOOA assigns a unique threshold to each sample, derived from its behavior across all repeated experiments.

Table 1: The comparison of GPRP computed by different privacy auditing methods across three models trained on the MNIST dataset at $2R = 400$.

Model	Method	Δ_{PA-GA}	Δ_{PA-GBA}	$\Delta_{PA-ALOOA}$
LR	SGD	6.269	3.919	11.03
	DP-SGD	4.289	2.301	6.194
MLP	SGD	6.957	4.364	14.21
	DP-SGD	5.063	3.024	7.447
CNN	SGD	2.336	1.288	4.740
	DP-SGD	2.531	1.257	3.556

For clarity, both the computed GPR and GPRP metrics are expressed as percentage points. The comparison results are shown in Fig.6 and Tab.1. From these charts, it is evident that, across all model results, the GPR and GPRP obtained using PA-ALOOA are significantly higher than those from the other two methods. Specifically, the results in Tab.1 show that, for the CNN model, the GPRP measured by PA-GA indicates that the GPRP value of DP-SGD is larger than that of SGD. This finding contradicts the conclusion in(Yaghini et al., 2022), which suggests that DPML algorithms should be able to bind the disparity of GPR. This discrepancy suggests that the PA-GA method underestimates the GPR, leading to inaccurate GPRP measurements and incorrect conclusions. In conclusion, PA-ALOOA more effectively captures privacy risks, preventing GPRP underestimation and accurately identifying privacy unfairness.

4. Mitigating Disparity of Group Privacy Risk for DP-SGD

In this section, we demonstrate that the GPR is closely tied to the group norms during training. Building on this observation, we enhance the DP-SGD algorithm to improve

privacy protection equity.

4.1. Experimental Observations

From Tab. 1, we observe significant GPR disparity in both SGD and DP-SGD, with DP-SGD reducing this disparity. Our goal is to enhance DP-SGD to provide more equitable privacy protection across groups, minimizing privacy risk differences.

In DP auditing literature, canaries are often created as mislabeled samples (Nasr et al., 2023; Annamalai & Cristofaro, 2024; Steinke et al., 2023). These samples generate larger gradient values during model training, which contribute more to parameter updates, thereby increasing the likelihood of model memorization. Motivated by the design of canaries, we hypothesize that during training, the larger a subgroup'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. In our analysis, 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 calculate the group relative contribution (GRC) across all iterations by averaging. To validate our hypothesis, we analyze the relationship between GRC and GPR through experiments.

As shown in Fig. 7, 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, with this phenomenon being more pronounced in simpler model architectures.

4.2. Mitigating Disparity of Group Privacy Risk

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 GRC during training.

As shown in Algo. 1, our proposed algorithm, abbreviated DP-SGD-Scale, differs from the standard DP-SGD in lines 7–9. In each iteration, we calculate the contribution of each subgroup's samples to the overall gradient, given by $\frac{\|\sum_{i\in B\cap D^k} g_i/|D^k|\|_2}{\|\sum_{i\in B} g_i/b\|_2}$, and use this information to adaptively adjust the clipping bound for each subgroup. Subgroups with higher contributions have their clipping bounds scaled down, leading to stricter clipping operations. This adjustment limits the influence of these subgroups on model updates, thereby reducing the model's memorization of these groups and mitigating their privacy leakage risks. Conversely, subgroups with relatively smaller contributions are assigned larger clipping bounds. However, the scaling factor of the clipping bounds is constrained by the hyperparameter τ, as excessively large clipping norms would introduce too much noise, making the model's performance unreliable.

Algorithm 1 DP-SGD-Scale

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 multiplier σ, clipping bound C, scale bound τ.

1: Initialize $w^{(0)}$ randomly.
2: **for** $t = 0, ..., T-1$ **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: **end for**
7: **for** $k \in K$ **do**
8: $C^k \leftarrow C \cdot min(\tau, \frac{\|\sum_{i\in B} g_i/b\|_2}{\|\sum_{i\in B\cap D^k} g_i/|D^k|\|_2})$
9: **end for**
10: $C = \max_{k\in K}(C^k)$
11: **for** $i \in B$ **do**
12: $\bar{g}_i \leftarrow g_i \cdot min(1, \frac{C^k}{\|g_i\|_2})$
13: **end for**
14: $\tilde{g} \leftarrow \frac{1}{b}\left(\sum_{i\in B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
15: $w^{(t+1)} \leftarrow w^{(t)} - \eta\tilde{g}$
16: **end for**

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

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

Datasets. We experiment with commonly used datasets in privacy-fairness studies (Annamalai & Cristofaro, 2024; Bagdasaryan et al., 2019; Xu et al., 2021): MNIST. Additionally, we evaluate four fairness-related datasets: three tabular datasets widely employed in fairness-aware ML studies—Adult, Credit, and Law (Le Quy et al., 2022)—and an image dataset, UTKFace (Zhang & Qi, 2017).

Models. We use three types of models: Logistic Regression (LR), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN).

Training algorithms. SGD, DP-SGD and DP-SGD-S. We implement DP-SGD using the Opacus library (Yousefpour et al., 2021). For DP-SGD and DP-SGD-S, the default privacy hyperparameters (ϵ, δ) are configured with $(10, 1\text{e}{-5})$. For DP-SGD-S, the default scale bound τ is set to 2.

Auditing setup. We set $n = m$ and $2R = 400$, which we believe is a reasonable configuration, as analyzed in Sec. 3.2. Additionally, we use the state-of-the-art DP auditing method f-DP (Nasr et al., 2023), to empirically evaluate the privacy budget $\tilde{\epsilon}$ with a 95% confidence.

Evaluation metrics. We use Δ (defined in 3.4) to evaluate the fairness of privacy protection across groups, accuracy to measure model performance, and $\tilde{\epsilon}$, obtained through the f-DP auditing method, to empirically assess the algorithm's worst-case privacy guarantees.

More detailed information is provided in Appendix A.5.

5.2. MNIST Dataset

We conduct extensive experiments to compare the privacy auditing results obtained under three training algorithms across various model types and privacy protection levels. Additionally, we examined the impact of varying scale bounds in DP-SGD-S, analyzing how different values affect both accuracy and GPRP.

Results across different model types. Fig.8 and Tab.2 present the results of the three algorithms on the GPRP metric across different models. They clearly show that both SGD and DP-SGD exhibit significant inter-group privacy risk disparities, as reflected in the relatively large Δ values, particularly in simpler model architectures LR and MLP. In more complex architecture CNN, GPRP decreases considerably. This suggests that using more complex models for training could be a practical approach when deploying on public platforms. Across all model architectures, our proposed algorithm, DP-SGD-S, consistently performs better on the GPRP metric, yielding smaller values, demonstrating that our enhancement leads to a more equitable privacy protection mechanism.

Results across different privacy budgets. For DP-SGD and DP-SGD-S, we set different theoretical privacy budgets of 10 and 100, comparing their performance on the GPRP metric under these budgets. Using the f-DP method, we also evaluate the empirical privacy budget to verify alignment with the theoretical guarantees. Results are shown in Tab. 2.

As shown in Tab. 2, a smaller ϵ consistently results in a smaller Δ, indicating that stronger privacy protection reduces GPR disparity. Under the same theoretical privacy budget, the Δ values for our method, DP-SGD-S, are consistently smaller than those of DP-SGD across all model types, demonstrating its effectiveness in mitigating GPR disparity. Additionally, the $\tilde{\epsilon}$ values for both DP-SGD and DP-SGD-S do not exceed the theoretical ϵ, validating their correct implementation. The $\tilde{\epsilon}$ values also confirm that the additional operations in DP-SGD-S do not compromise its privacy guarantees or violate the definitions of DP.

As shown in the accuracy column of Tab. 2, at $\epsilon = 100$, DP-SGD-S shows minimal differences compared to DP-SGD, with a potential 1% accuracy drop. However, at $\epsilon = 10$, DP-SGD-S may experience a larger accuracy decline of 3% to 5%. This indicates that while our method effectively reduces inter-group privacy risk disparity, it comes at the cost of model accuracy.

Table 2: The performance of three algorithms under different theoretical privacy budgets across various model types, including model prediction performance, GPRP, and empirical privacy budget results.

Model	ϵ	Method	Accuracy (↑)	Δ(↓)	$\tilde{\epsilon}$
LR	/	SGD	89.67 ± 0.0029	11.03	22.42
	100	DP-SGD	89.65 ± 0.0018	9.143	6.283
		DP-SGD-S	89.06 ± 0.0035	6.545	6.684
	10	DP-SGD	89.30 ± 0.0048	6.184	3.943
		DP-SGD-S	86.42 ± 0.0037	3.365	3.018
MLP	/	SGD	89.67 ± 0.0029	11.03	22.42
	100	DP-SGD	91.50 ± 0.0020	10.90	7.531
		DP-SGD-S	90.55 ± 0.0031	7.386	7.012
	10	DP-SGD	90.01 ± 0.0023	7.447	3.910
		DP-SGD-S	85.70 ± 0.0043	3.723	2.848
CNN	/	SGD	89.67 ± 0.0029	11.03	22.42
	100	DP-SGD	94.95 ± 0.0023	4.418	3.452
		DP-SGD-S	94.74 ± 0.0033	3.756	3.698
	10	DP-SGD	94.43 ± 0.0024	3.556	2.732
		DP-SGD-S	91.65 ± 0.0059	2.567	2.057

Results across different scale bounds. We experiment with different scale bound values τ in DP-SGD-S to evaluate their impact on accuracy and GPRP metrics. Our experimental results in Fig. 4 show that as the scale bound increases, the Δ value decreases, as larger scale bounds further limit the contribution of groups with larger norms to model updates. However, this improvement in fairness comes at the cost of accuracy, likely due to the model's diminished ability to extract optimization information from these groups. This underscores the importance of selecting an appropriate scale bound to balance privacy fairness and prediction accuracy in our method.

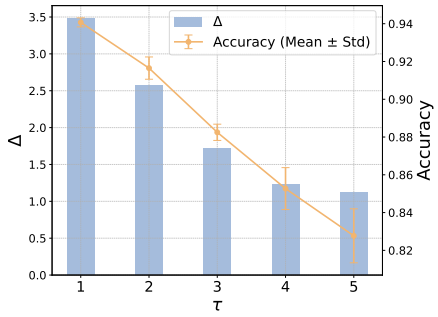


Figure 4: The performance of the DP-SGD-S algorithm under different scale bound values, showing model prediction accuracy and GPRP metrics under CNN.

5.3. Fairness-related Datasets

We perform experiments on commonly used fairness-related tabular and image datasets to demonstrate that our proposed algorithm, DP-SGD-S, can effectively mitigate the issue of unequal privacy leakage risks among sensitive groups.

As shown in Tab. 3, we first analyze the performance of three training algorithms on the tabular datasets. We observe that the Δ value is already relatively small under the SGD training algorithm. DP-SGD further bounds the disparity between GPR. Our algorithm, DP-SGD-S, achieves the smallest GPRP value while maintaining nearly the same accuracy.

For the UTKFace dataset, we observe that algorithms with privacy mechanisms, namely DP-SGD and DP-SGD-S, exhibit a significant reduction in Δ values compared to SGD. Analyzing the obtained $\tilde{\epsilon}$, it appears that models trained with SGD inherently have higher privacy risks, potentially amplifying disparities in privacy risk among groups. Furthermore, our algorithm DP-SGD-S achieves a lower Δ value compared to DP-SGD, with only a roughly 2% drop in accuracy. In terms of $\tilde{\epsilon}$, both DP-SGD and DP-SGD-S adhere to the definition of DP.

5.4. Summary

Based on the results from all the datasets used in our experiments, our method, DP-SGD-S, achieves a lower Δ value while maintaining the same privacy guarantees as DP-SGD. Additionally, increasing the scale bound for DP-SGD-S can further reduce the GPRP value. However, it is important to note that this often comes at the cost of model accuracy.

Based on the results obtained from image datasets, comparing SGD and DP-SGD under different theoretical privacy guarantees (ϵ=10 and ϵ=100), we can reach a fairly consistent conclusion: the stronger the model's privacy protection capability, the smaller the differences in privacy risk between groups. 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 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 budget compared to DP-SGD, which is highly meaningful.

Table 3: The performance of three algorithms under different fairness-related datasets, including model prediction performance, GPRP, and empirical privacy budget results.

Dataset	Method	Accuracy (↑)	Δ(↓)	$\tilde{\epsilon}$
Adult	SGD	84.84 ± 0.0006	0.516	2.737
	DP-SGD	85.00 ± 0.0003	0.280	1.426
	DP-SGD-S	84.86 ± 0.0751	0.236	2.031
Credit	SGD	81.95 ± 0.0521	0.311	3.974
	DP-SGD	81.91 ± 0.0720	0.163	1.360
	DP-SGD-S	81.92 ± 0.0901	0.081	1.882
Law	SGD	89.69 ± 0.0008	1.196	1.964
	DP-SGD	89.82 ± 0.0009	0.683	1.667
	DP-SGD-S	89.84 ± 0.0008	0.439	1.373
UTKFace	SGD	78.47 ± 0.0033	39.26	24.69
	DP-SGD	73.47 ± 0.0024	4.425	2.004
	DP-SGD-S	71.10 ± 0.0104	1.772	1.885

6. Conclusion

In this paper, we develop an efficient and precise privacy auditing method to better measure and identify inter-group privacy risk disparities, and propose a solution to mitigate these disparities. Our extensive experiments across various datasets demonstrate that our algorithm achieves a more equitable group privacy protection mechanism while maintaining the same privacy guarantees as DP-SGD.

However, many fairness issues remain to be explored and addressed. Specifically, we have not yet considered the outcome fairness issues raised in other studies on DPML algorithms. As privacy protection becomes more equitable, it is essential to investigate whether this conflicts with or complements outcome fairness. Ensuring the AI system achieves true fairness is of utmost importance.

Impact Statement

This paper aims to contribute to the advancement of Machine Learning. While our work may have various societal implications, we do not find it necessary to highlight any specific ones here.

References

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

Act, E. C. O. Equal credit opportunity act. *Women in the American Political System: An Encyclopedia of Women as Voters, Candidates, and Office Holders*, 2:129, 2018.

Annamalai, M. S. M. S. and Cristofaro, E. D. 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.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. In *Advances in neural information processing systems*, pp. 15479–15488. Curran Associates Inc., 2019.

Chang, H. and Shokri, R. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 292–303. IEEE, 2021.

Ding, J., Zhang, X., Li, X., Wang, J., Yu, R., and Pan, M. Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 622–629, 2020.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, volume 3, pp. 265–284. Springer, 2006b.

Ekstrand, M. D., Joshaghani, R., and Mehrpouyan, H. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on fairness, accountability and transparency*, pp. 35–47. PMLR, 2018.

Esipova, M. S., Ghomi, A. A., Luo, Y., and Cresswell, J. C. Disparate impact in differential privacy from gradient misalignment. In *The Eleventh International Conference on Learning Representations*, 2023. https://openreview.net/forum?id=qLOaeRvteqbx.

Huang, C., Zhang, Z., Mao, B., and Yao, X. An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819, 2023.

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

Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: how private is private sgd? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 22205–22216, 2020.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Lowy, A., Gupta, D., and Razaviyayn, M. Stochastic differentially private and fair learning. In *Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, pp. 86–119. PMLR, 2023.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Nasr, M., Songi, S., Thakurta, A., Papernot, N., and Carlin, N. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882. IEEE, 2021.

Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., and Terzis, A. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1631–1648, 2023.

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., Vassilvitskii, S., Chien, S., and Thakurta, A. G. How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Steinke, T., Nasr, M., and Jagielski, M. Privacy auditing with one (1) training run. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 49268–49280, 2023.

Tran, C., Dinh, M., and Fioretto, F. Differentially private empirical risk minimization under the fairness lens. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27555–27565. Curran Associates Inc., 2021a.

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

Voigt, P. and Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676): 10–5555, 2017.

Xu, D., Yuan, S., and Wu, X. Achieving differential privacy and fairness in logistic regression. In *Companion proceedings of The 2019 world wide web conference*, pp. 594–599, 2019.

Xu, D., Du, W., and Wu, X. 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*, pp. 1924–1932, 2021.

Yaghini, M., Kulynych, B., Cherubin, G., Veale, M., and Troncoso, C. Disparate vulnerability to membership inference attacks. In *Proceedings on Privacy Enhancing Technologies*, number 1, pp. 460–480, 2022.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3093–3106, 2022.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Zanella-Beguelin, S., Wutschitz, L., Tople, S., Salem, A., Rühle, V., Paverd, A., Naseri, M., Köpf, B., and Jones, D. Bayesian estimation of differential privacy. In *International Conference on Machine Learning*, pp. 40624–40636. PMLR, 2023.

Zarifzadeh, S., Liu, P., and Shokri, R. Low-cost high-power membership inference attacks. In *Forty-first International Conference on Machine Learning*, pp. 58244–58282. PMLR, 2024.

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

Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818, 2017.

A. Appendix.

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) (Mironov, 2017) and the moment accounting mechanism proposed by (Abadi et al., 2016), quantifies the privacy protection ability.

Algorithm 2 DP-SGD (Abadi et al., 2016)

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** $t = 0, ..., T-1$ **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: $\tilde{g} \leftarrow \frac{1}{b}\left(\sum_{i \in B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
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. Definition of Membership Inference Games

Definition A.1 (Average-case Membership Inference Game)**.** 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 and the population data pool, 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 A.2 (Worst-case Membership Inference Game)**.**

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 and the population data pool, 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 Audting 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) (Yaghini et al., 2022), 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) (Chang & Shokri, 2021), 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 subgroups. 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 (Chang & Shokri, 2021). 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 (Yaghini et al., 2022)

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** $i = 1, \ldots, m$ **do**
4: $\quad O[i] \leftarrow \ell(f(\mathbf{x_i}), y_i)$
5: $\quad H[i] \leftarrow \begin{cases} 1 & \text{if } z_i \in D \\ 0 & \text{otherwise} \end{cases}$
6: **end for**
7: $G \leftarrow [1\{O[i] \geq \beta\} \text{ for } i = 1, \ldots, m]$.

Return: Membership status H and guesses G.

Algorithm 4 PA-GBA (Chang & Shokri, 2021)

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 $\{\beta^k\}_{i=1}^K$.

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

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 (Nasr et al., 2023; Steinke et al., 2023; Zanella-Beguelin et al., 2023; Jagielski et al., 2020; Annamalai & Cristofaro, 2024; Nasr et al., 2021). 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 (Nasr et al., 2023; Steinke et al., 2023; Zanella-Beguelin et al., 2023; Jagielski et al., 2020; Annamalai & Cristofaro, 2024; Nasr et al., 2021)

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

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

Return: Membership status H and guesses G.

A.4. Experimental Figures

In this appendix section, we include some image data related to the experiments in the main text.

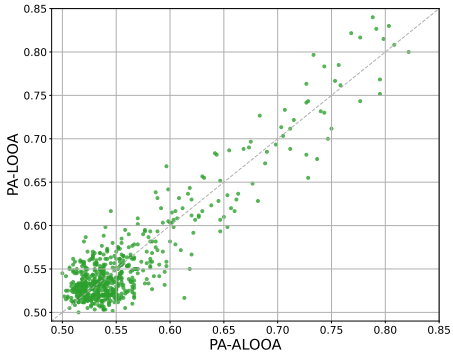
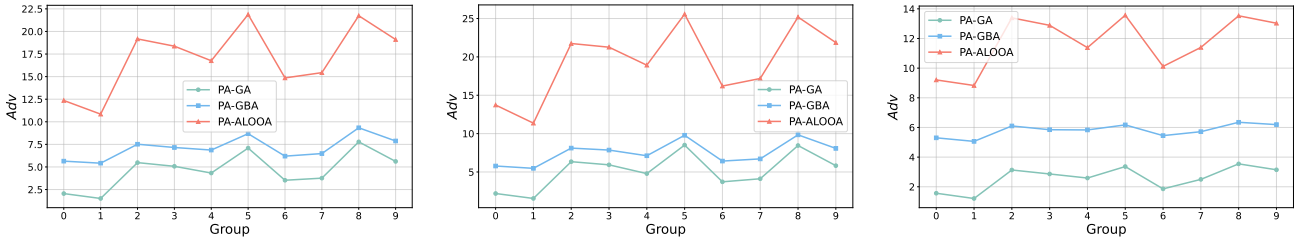


Figure 5: The horizontal axis represents the performance of PA-ALOOA, while the vertical axis represents the performance of PA-LOOA, both at $2R = 600$.



(a) LR. (b) MLP. (c) CNN.

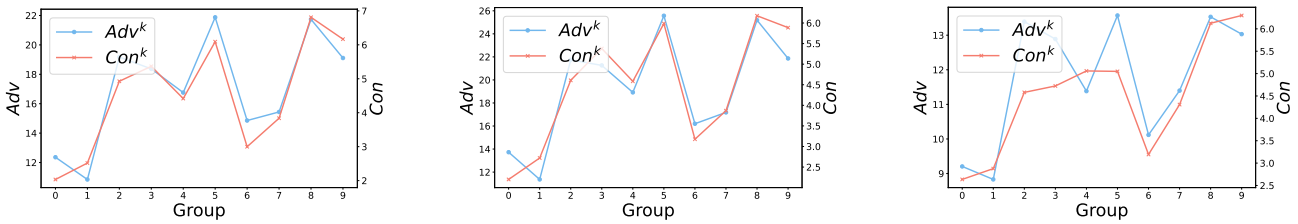
Figure 6: The comparison of GPR across three models—Logistic Regression (LR), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN)—trained on the MNIST dataset. Each subfigure depicts the GPR for different groups, with the y-axis representing GPR at $2R = 400$ and the x-axis representing groups.

A.5. Experimental Setup

Datasets. For the MNIST dataset, It contains 60,000 training and 10,000 testing samples of 28x28 grayscale images representing handwritten digits across ten classes. Due to computational limitations and to improve training efficiency, we randomly select 1,000 samples per class from the original dataset, forming a dataset of 10,000 samples for our experiments.

For the tabular fairness-related datasets, Adult, Credit, and Law (Le Quy et al., 2022; Esipova et al., 2023), containing 45,222, 30,000, and 20,798 records, respectively. We designate sensitive attributes as sex, race, and sex, respectively, all of which are binary. During the experiments, we group the data based on the sensitive attribute.

For the image fairness-related dataset–UTKFace dataset (Zhang et al., 2017) in our evaluation. After data cleaning and



(a) LR. (b) MLP. (c) CNN.

Figure 7: The relationship between GPR value and GRC value across three models trained on MNIST dataset. In each subfigure, the left vertical axis represents GPR value at $2R = 400$, while the right vertical axis represents GRC value.

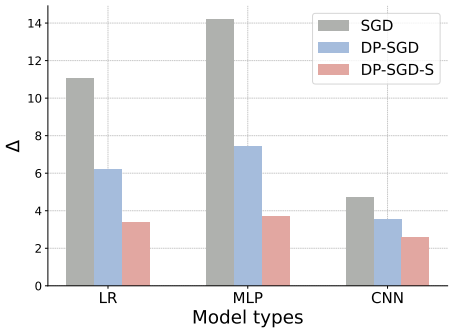


Figure 8: The comparison of the performance of three algorithms—SGD, DP-SGD, and DP-SGD-S—on the GPRP metric (Δ) across three different model types.

preprocessing[1], the dataset consists of 27,305 samples of 1x48x48 images. For this dataset, we use ethnicity as the protected attribute and the prediction label, which has five categories.

Models. For the MNIST dataset, we use Logistic Regression (LR), Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN) for training. These models are trained with the SGD optimizer using a learning rate of 0.1. For tabular fairness-related datasets, we train models using LR with the same optimizer and learning rate. For image fairness-related datasets, we use CNN models trained with the Adam optimizer and a learning rate of 0.001. In all experiments, we set the batch size to 256 and train for 20 epochs.

Auditing setup. Specifically, we identify the data point with the highest individual privacy risk and apply the f-DP method to estimated $\tilde{\epsilon}$.

Evaluation metrics. Specifically, the accuracy of the training algorithms is computed by splitting the datasets into 80% for training and 20% for testing, with results reported based on five independent repetitions of the experiment.

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