Towards Private and Fair Machine Learning: Group-Specific Differentially Private Stochastic Gradient Descent with Threshold Optimization

Zhi Yang[1][0009-0000-8725-1125], Changwu Huang[1][0000-0003-3685-2822](✉), and Xin Yao[2][0000-0001-8837-4442]

[1] Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation,
Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China
[2] School of Data Science, Lingnan University, Hong Kong, China
huangcw3@sustech.edu.cn

Abstract. As machine learning (ML) algorithms become increasingly prevalent in daily life applications, addressing privacy and fairness concerns is imperative and crucial from both ethical and legal perspectives. Establishing private and fair ML models stands as a critical task in cultivating trustworthy ML practices. Recent research has delved into the challenges of merging differential privacy (DP) with group fairness. One aspect focuses on mitigating the amplified accuracy disparity among sensitive groups caused by DP, while another emphasizes maintaining outcome fairness in private models trained by DP methods. However, these dual research objectives often present conflicting demands, with existing methods typically tackling them independently. To bridge this gap, we introduce a novel approach that combines a group-specific DP stochastic gradient descent training mechanism with classification threshold optimization to address these intertwined challenges. Extensive experiments demonstrate the effectiveness of our method in concurrently reducing accuracy parity and demographic parity measurements. Hence, our proposed method can achieve private and fair ML models, which contribute to the development of trustworthy ML.

Keywords: Differential Privacy · Group Fairness · Trustworthy Machine Learning · Ethics of AI.

1 Introduction

Artificial intelligence (AI) technologies, particularly machine learning (ML) algorithms, are widely adopted across various sectors, augmenting or even replacing human decision-making processes in our daily lives. However, with the increasing integration of ML, especially in critical domains like healthcare, finance, and the judiciary, the potential risks associated with algorithmic decisions have come to light, including concerns such as data privacy breaches, biases or discrimination against specific groups, lack of explainability, security vulnerabilities, safety implications and etc. [11, 15]. Among these ethical issues and risks, privacy and fairness stand out as pivotal components for establishing trust in AI [9].

Data privacy protection is reinforced by regulations like the General Data Protection Regulation (GDPR) in Europe [28], spotlighting the regulatory emphasis on addressing privacy challenges within ML algorithms. In recent years, extensive research has been dedicated to addressing privacy issues in ML algorithms. Techniques like differential privacy (DP) [8], homomorphic encryption [12], and federated learning [19] have been developed. Among these, DP has emerged as the predominant choice for ensuring data privacy [25]. DP's methodologies can be categorized into four types—input perturbation, objective perturbation, gradient perturbation, and output perturbation—based on where noise is integrated into the ML pipeline [24].

Ensuring model fairness is equally critical, exemplified by laws like the Equal Credit Opportunity Act (ECOA) in the U.S., which prohibits discrimination based on protected attributes [2]. Consequently, fairness in ML has received significant attention and discussion over the past decades [15]. Many fairness-aware ML methods have been developed to address and mitigate model biases. These approaches are often classified into pre-processing, in-processing, and post-processing methods, depending on when they are implemented during the model development phase [23].

From both ethical and legal perspectives, the deployment of ML algorithms must consider fairness and privacy, as these are interconnected social issues rather than isolated concerns [4, 9]. Protecting data privacy increases individuals' willingness to share their data for model training, supporting the ongoing development of ML. Ensuring model fairness enhances public trust in using ML for real-world decision-making, facilitating its practical application. As ML advances and is applied more broadly, the feedback loop further drives the field forward. Consequently, research efforts increasingly focus on integrating DP and fairness-aware ML. However, combining them presents numerous challenges. These challenges can be broadly classified into two categories based on which aspect of fairness they aim to address. The first category deals with the amplified unfairness (i.e., the utility or accuracy parity) brought about by DP [10, 30], while the second focuses on achieving outcome fairness of the ML model [22, 27]. While existing literature demonstrates progress in separately addressing these fairness issues within DP models, a notable observation is the potential conflict between solutions targeting accuracy parity and those emphasizing outcome fairness. This disconnect implies that current methods tend to tackle these problems in isolation without considering their combined impact.

To bridge this gap, this work introduces a group-specific DP stochastic gradient descent (DP-SGD) training mechanism with classification threshold optimization, which concurrently addressing accuracy and outcome fairness issues in differentially private models. Specifically, we adopt a group-specific training mechanism to alleviate the accuracy disparity induced by DP and employ a post-processing strategy known as reject option-based classification (ROC) to promote outcome fairness. The main contributions of this work are as follows:

- We propose using a group-specific training mechanism to alleviate accuracy disparity among sensitive groups intensified by DP-SGD.
- We implement an approach that integrates the group-specific training mechanism with the post-processing method to tackle the dual challenges of integrating DP and group fairness.
- Our proposed method effectively mitigates the accuracy disparity amplified by DP and ensures outcome fairness while maintaining reasonable model performance.

The structure of this paper is as follows: Section 2 introduces DP, discusses some fairness measurements, and reviews approaches combining DP and group fairness. In Section 3, we detail our proposed method. Section 4 presents our experimental studies. Finally, Section 5 offers a brief conclusion. Our code is available at: https://github.com/ruayz/gs-dpsgd-to.

2 Related Work

In this section, we first give the problem setting and fairness measurements investigated in this work. Following that, a brief introduction to the differential privacy approach in machine learning is provided. Next, we summarize and discuss the existing research on fair and private machine learning methods that consider both differential privacy and group fairness.

2.1 Problem Setting and Fairness Measurements

Problem Setting Our work focuses on a binary classification task, which aims to learn the mapping function from user feature vectors $\mathbf{x} \in \mathbb{R}^d$ to class labels $y \in \{0, 1\}$. This task is typically accomplished by constructing a model or classifier $f : \mathbb{R}^d \mapsto \mathbb{R}$ using a training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. The goal is to enable the classifier to predict the label $\hat{y} = f(\mathbf{x})$ for a given feature vector $\mathbf{x}$. The model f is commonly parameterized with its parameters w and hence denoted as f_w to reflect its dependency on these parameters. Correspondingly, $f_w(\mathbf{x})$ denotes the prediction of model f parameterized by w on an input vector $\mathbf{x}$. Each sample $(\mathbf{x}_i, y_i)$ in the dataset D has its sensitive attribute value $s \in \mathcal{S}$ (e.g., sex, race). It's worth noting that multiple sensitive attributes can be considered, but we just focus on a single sensitive attribute case (e.g., the sex of each user $s = \{female, male\}$). We denote two distinct groups associated with the sensitive attribute as s_a (deprived group) and s_b (privileged group). Therefore, each training instance $(\mathbf{x}_i, y_i) \in D$ is associated with a specific sensitive attribute value $s_i \in \{s_a, s_b\}$. And we denote subsets of the training dataset D where $s = s_a$ and $s = s_b$ as $D_a = \{(\mathbf{x}_i, y_i) \in D | s_i = s_a\}$ and $D_b = \{(\mathbf{x}_i, y_i) \in D | s_i = s_b\}$, respectively.

Fairness Measurements In general, fairness typically refers to the just treatment of individuals or groups without bias arising from their inherent or acquired characteristics [20]. There are different types of fairness measurements, including group and individual fairness [20]. Below, we introduce two group fairness measures investigated in this work.

- Demographic Parity (DemParity) [6] requires that the difference of positive prediction rates should be small across various sensitive groups. DemParity is widely used as an indicator to measure outcome fairness, which is assessed as follows:
$$|P(\hat{y}=1|s=s_a)-P(\hat{y}=1|s=s_b)| \leq \theta, \tag{1}$$
- Accuracy Parity (AccParity) [3] requires that the difference of accuracy should be small across different sensitive groups. This is assessed as follows:
$$|P(\hat{y}=y|s=s_a)-P(\hat{y}=y|s=s_b)| \leq \theta, \tag{2}$$

where θ represents the fairness constraint that models must satisfy and can be adjusted according to specific requirements.

2.2 Differential Privacy

Differential privacy (DP) is a privacy definition proposed by Dwork [8] to address the issue of privacy leakage. In the following, we introduce the approximate (ϵ, δ)-DP definition.

Definition 1 ((ϵ, δ)-Differential Privacy [7]): An algorithm $\mathcal{F}$ is said to satisfy approximate differential privacy if for all pairs of databases D and D' and all possible outputs $O \subseteq \text{Range}(\mathcal{F})$, the following condition holds:

$$P[\mathcal{F}(D) \in O] \leq \exp(\epsilon) \times P[\mathcal{F}(D') \in O] + \delta, \tag{3}$$

where ϵ in the definition of DP is generally called privacy parameter or privacy budget. A smaller value of ϵ indicates that the mechanism provides outputs that are very similar to neighboring inputs, thereby ensuring stronger privacy protection. The δ represents the probability of failure to satisfy the approximation to DP. It is typically set to a small value. In practical applications, this definition of approximate differential privacy is more commonly used.

Differentially Private Stochastic Gradient Descent (DP-SGD) DP-SGD [1] integrates differential privacy concepts with stochastic gradient descent (SGD), protecting individual data privacy by adding noise to gradients during training. It upholds model privacy via gradient clipping and noise operations within SGD, following the (ϵ, δ)-DP definition. DP-SGD is extensively adopted in deep learning, serving as a key technique for integrating DP into ML models and underpinning privacy-centric libraries [24].

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 $\ell(\hat{y}, y)$, DP-SGD (as illustrated in Algorithm 1) integrates gradient clipping and noise addition for achieving the (ϵ, δ)-DP guarantees. In Algorithm 1, 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) [21] and the moment accounting mechanism proposed by [1], quantifies the privacy protection ability.

Algorithm 1 DP-SGD [1]

```
Input: Training dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the parameterized model $f_w(\cdot)$, loss function $\ell(\hat{y}, y)$ for prediction $\hat{y}$ and label $y$, iterations $T$, batch size $b$, learning rate $\eta$, noise scale $\sigma$, gradient norm bound $C$.
1:  Initialize $w^{(0)}$ randomly.
2:  for $t = 0, 1, ..., T-1$ do
3:      Sample a batch $B^{(t)}$ from $D$ with sampling probability $b/N$ for each data point.
4:      for $i \in B^{(t)}$ 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:      $\tilde{g} \leftarrow \frac{1}{b}\left(\sum_{i \in B^{(t)}} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
9:      $w^{(t+1)} \leftarrow w^{(t)} - \eta \tilde{g}$
10: end for
Output: Model $f_{w^{(T)}}(\cdot)$ and accumulated $(\epsilon, \delta)$.
```

2.3 Differential Privacy and Group Fairness

While DP and group fairness have been extensively studied in isolation, their intersection has only recently garnered significant attention [25]. Integrating these two concepts may introduce several challenges. Research in this area can be broadly classified into two categories. The first category aims to mitigate the unfairness introduced by the application of DP mechanisms. The second category of research focuses on combining DP with outcome fairness. We will introduce each category separately below.

Mitigating the unfairness amplified by DP. Recent studies have found that incorporating DP into models can increase AccParity measurement between sensitive groups [3, 26, 30]. Various efforts have been made to address this issue [10, 14, 30]. For example, [30] introduces a heuristic removal algorithm, termed DP-SGD-F, designed to achieve equal utility loss across different groups. Similarly, [10] introduces an adaptive global scaling technique to mitigate the unfairness resulting from the application of DP, which is called DP-SGD-Global-Adapt (hereafter referred to as DP-SGD-A). Although current methods have effectively alleviated accuracy disparity among sensitive groups, they have not yet addressed the issue of outcome fairness measured by DemParity. Specifically, models may still exhibit biased decision-making towards sensitive groups.

Achieving outcome fairness in differentially private models. Several works study how to achieve outcome fairness using fairness-aware learning when enforcing differential privacy in the private model [31, 27, 22, 16, 5, 13]. For instance, [27] improves outcome fairness by clipping the weights of the final layer during model training, a method known as FairDP. This approach is supported by both theoretical proofs and experimental results that demonstrate its certified fairness. The findings indicate that FairDP surpasses existing combination methods in achieving a superior utility-fairness-privacy trade-off. Additionally, [22] employs a post-processing method, specifically the reject option based classification method proposed by [17], to achieve outcome fairness in private models. We refer to this method as DP-SGD-P. The findings show that it outperforms the baseline in terms of performance. However, either of them ensures the alleviation of accuracy disparity across sensitive groups.

In summary, existing methodologies typically address these two research challenges in isolation, failing to integrate them in a manner that achieves both objectives, which are generally conflicting. This observation motivates the development of a novel approach that concurrently attains an optimal balance between mitigating accuracy disparity and achieving outcome fairness.

3 Group-Specific DP-SGD with Threshold Optimization

In this section, we propose the group-specific DP-SGD with a threshold optimization approach, which can achieve a fair and private ML model. The proposed method first uses group-specific DP-SGD to train a private model. Then, the threshold optimization strategy, i.e., threshold optimization classification, is adopted on the private model to improve the outcome fairness of the model.

3.1 DP-SGD using Group-Specific Training

In this work, we introduce GS-DP-SGD, a group-specific training strategy for DP-SGD, aimed at alleviating the accuracy discrepancy exacerbated by DP-SGD. As highlighted in [3], groups with larger gradient norms experience more significant accuracy declines due to clipping operations. Moreover, the optimization directions generated by SGD may be unfavorable to these groups, leading to ineffective or even deteriorated accuracy improvements. Through our GS-DP-SGD, we aim to tailor the optimization trajectory for specific groups, reducing the impact of gradient clipping and enhancing performance. Furthermore, by training DP-SGD separately with group-specific datasets, distinct privacy parameters can be set for each group, ensuring an equitable distribution of the privacy budget across all groups. This addresses concerns raised in the literature [32] regarding varying levels of privacy protection among different groups.

The overview workflow of GS-DP-SGD is depicted in Fig. 1 and detailed in Algorithm 2. Firstly, The dataset D is divided into D_a and D_b based on sensitive groups. Two neural networks are then trained independently on these group-specific datasets. At the end of each epoch, each neural network generates updated weights. These weights are then aggregated and averaged to produce the final weight $w^{(t+1)}$, which is subsequently propagated back to each neural network as the initial weight for the next training iteration. The workflow of GS-DP-SGD is similar to the training framework of FairDP [27], except for the weight clipping operation between line 2 and line 3 in Algorithm 2. FairDP clips the last layer of weights to a bounded norm to ensure outcome fairness. According to the proof in [27], the GS-DP-SGD satisfies (ϵ, δ)-DP.

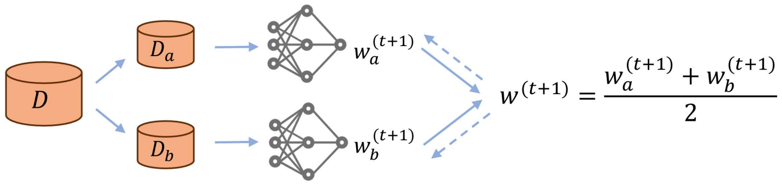

Fig. 1. The workflow of group-specific training mechanism.

Algorithm 2 GS-DP-SGD

Input: Training dataset $D = \{D_a, D_b\}$, the parameterized model $f_w(\cdot)$, loss function $\ell(\hat{y}, y)$ for prediction $\hat{y}$ and label y, iterations T, batch size b, learning rate η, noise scale σ, gradient norm bound C.

```
1:  Initialize $w^{(0)}$ randomly.
2:  for $t = 0, 1, ..., T-1$ do
3:      $w_a^{(t)} = w_b^{(t)} = w^{(t)}$
4:      Sample a batch $B_a^{(t)}$ from $D_a$ with sampling probability $b/|D_a|$.
5:      for $i \in B_a^{(t)}$ do
6:          $g_i \leftarrow \nabla \ell(f_{w_a^{(t)}}(\mathbf{x}_i), y_i)$
7:          $\bar{g}_i \leftarrow g_i \cdot min(1, \frac{C}{\|g_i\|_2})$
8:      end for
9:      $\tilde{g}_a \leftarrow \frac{1}{b}\left(\sum_{i \in B_a^{(t)}} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
10:     $w_a^{(t+1)} \leftarrow w_a^{(t)} - \eta \tilde{g}_a$
11:     Sample a batch $B_b^{(t)}$ from $D_b$ with sampling probability $b/|D_b|$.
12:     for $i \in B_b^{(t)}$ do
13:         $g_i \leftarrow \nabla \ell(f_{w_b^{(t)}}(\mathbf{x}_i), y_i)$
14:         $\bar{g}_i \leftarrow g_i \cdot min(1, \frac{C}{\|g_i\|_2})$
15:     end for
16:     $\tilde{g}_b \leftarrow \frac{1}{b}\left(\sum_{i \in B_b^{(t)}} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
17:     $w_b^{(t+1)} \leftarrow w_b^{(t)} - \eta \tilde{g}_b$
18:     $w^{(t+1)} \leftarrow \frac{w_a^{(t+1)} + w_b^{(t+1)}}{2}$
19: end for
```

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

3.2 Threshold Optimization for GS-DP-SGD

To mitigate the outcome unfairness of the private model trained by GS-DP-SGD, we incorporate a post-processing method called reject option based classification (ROC) [17]. Traditionally, a learned model assigns a data instance to the class with the highest predicted probability, whereas ROC leverages these probabilities to identify instances for labeling in a manner that neutralizes the effect of discrimination for the deprived group [17]. The decision mechanism of ROC is delineated in Algorithm 3. According to Algorithm 3, the threshold γ determines the decision behavior of the model and hence affects the outcome fairness. To adhere to a specified fairness constraint θ, we optimize the threshold for ROC on the validation dataset, termed Threshold Optimization Classification (TOC), as detailed in Algorithm 4.

In Algorithm 4, we use a simple grid search strategy to find the proper threshold value. Firstly, 100 threshold values are generated evenly spaced within [0.5, 1]. Then, for each threshold, we apply the ROC method to the validation set. We then calculate the fairness metric (in our work, we choose the DemParity) to identify the optimal threshold that meets the fairness constraint θ. If no threshold satisfies the fairness constraint, we select the threshold with minimal fairness metric.

Algorithm 3 Reject Option based Classification (ROC)

```
Input: Model $f_{w^{(T)}}(\cdot)$, threshold $\gamma$, feature Vector $\mathbf{x}$ (with sensitive attribute $s$).
$p = f_{w^{(T)}}(\mathbf{x})$
if $max(p, 1-p) < \gamma$ then
    if $s = s_a$ then $\hat{y} = 1$
    if $s = s_b$ then $\hat{y} = 0$
else
    if $p \geq 0.5$ then $\hat{y} = 1$
    else $\hat{y} = 0$
end if
Output: The predicted label $\hat{y}$ for the feature vector $\mathbf{x}$.
```

Algorithm 4 Threshold Optimization based Classification (TOC)

```
Input: Validation dataset $D_{valid} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$, fairness constraint $\theta$, model $f_{w^{(T)}}$.
for $\gamma \in$ linspace(0.5, 1, 100) do
    for $\mathbf{x}_i \in D_{valid}$ do
        $\hat{y_i}$ = ROC($f_{w^{(T)}}(\cdot)$, $\gamma$, $\mathbf{x}_i$)
    end for
    $m_\gamma = |P(\hat{y} = 1|s = s_a) - P(\hat{y} = 1|s = s_b)|$
    if $m_\gamma \leq \theta$ then Return $\gamma$
end for
Return the $\gamma$ that has minimal $m_\gamma$
Output: The threshold $\gamma$.
```

Overall, in our work, we implement GS-DP-SGD with Threshold Optimization (referred to as GS-DP-SGD-TO) to address both the exacerbated accuracy disparity (i.e., AccParity) among sensitive groups and the outcome fairness (i.e., DemParity) in differentially private models. The pseudocode for GS-DP-SGD-TO is shown in Algorithm 5.

```
Algorithm 5 GS-DP-SGD-TO
Input: Training dataset $D_{train}$, validation dataset $D_{valid}$, the parameterized model
    $f_w(\cdot)$, loss function $\ell(\cdot,\cdot)$, iterations $T$, batch size $b$, learning rate $\eta$, noise scale $\sigma$,
    gradient norm bound $C$, fairness constraint $\theta$.
1: $f_{w^{(T)}}$, $(\epsilon, \delta)$ = GD-DP-SGD($D_{train}$, $f(\cdot)$, $\ell(\cdot,\cdot)$, $T$, $b$, $\eta$, $\sigma$, $C$)
2: $\gamma$ = TOC($D_{valid}$, $\theta$, $f_{w^{(T)}}$)
Output: The ROC model ROC($f_{w^{(T)}}, \gamma, \cdot$) and the accumulated $(\epsilon, \delta)$.
```

4 Experimental Study

In our experiments, we provide evidence that both the unfairness introduced by DP (AccParity) and the outcome unfairness (DemParity) in differentially private models are mitigated by our proposed approach, while also demonstrating reasonable performance.

4.1 Experimental Setting

Datasets : Our experiments will be conducted on six commonly used binary classification datasets [18]: Adult, Dutch, Bank, Credit, Compas, and Law. Their sensitive attributes are sex, sex, race, race, martial, and sex, respectively. Data preprocessing is performed following [10].

Compared Methods : We benchmark our method against three approaches designed to mitigate accuracy disparity among sensitive groups in differentially private models: DP-SGD-F [30], DP-SGD-A [10], and GS-DP-SGD(mentioned in 3.1). Then, we evaluate our method against two methods aimed at achieving outcome fairness in differentially private models: FairDP[27] and DP-SGD-P[22].

Models : Experiments are performed on Multi-layer Perceptron (MLP) with two hidden layers of 256 units and a maximum number of iterations of 20. The privacy hyperparameters are set with a default clipping bound of 0.5 and δ set to 1e-6. Additionally, the batch size is set to 256.

Evaluation Metrics : We utilize two commonly used group fairness metrics to evaluate model fairness: AccParity and DemParity, as described in Section 2.1. Each dataset is randomly partitioned into training, validation, and test sets with a ratio of 7:1:2. To ensure the reliability of the experiment results, the reported metrics represent the average performance across 15 independent repeated runs.

4.2 Experimental Results

Compare with methods of alleviating the accuracy unfairness intensified by DP. We conduct a comparative analysis between our method, GS-DP-SGD-TO, and three other methods designed to reduce accuracy disparity between sensitive groups. Additionally, we include SGD and DP-SGD in the experiments to observe the optimization effects of other methods on the AccParity metric. We use the Opacus open-source library to preset the same privacy protection value for all methods. Although the computed ϵ show slight differences in Tab. 1, these minor variations are negligible, allowing us to assume that all methods provide consistent privacy protection for each specific dataset. In addition, the GS-DP-SGD-TO method applies a fairness constraint on the DemParity metric which is set to 0.05. While a DemParity value of 0 is deemed optimal for outcome fairness, a model is considered to have acceptable outcome fairness if its DemParity measurement falls within the range of 0.05 to 0.1 [22, 29]. Next, we will analyze the results of experiments from two aspects.

Focus on the AccParity metric. From Tab. 1, we observe that DP-SGD yields significantly higher AccParity values compared to SGD. This indicates that incorporating DP into the gradient descent process exacerbates accuracy disparity between sensitive groups, which is consistent with findings in the literature [3, 30, 26]. Compared to DP-SGD, GS-DP-SGD effectively reduces the AccParity value across all datasets, demonstrating the effectiveness of using a group-specific training mechanism to mitigate accuracy unfairness issues. Overall, DP-SGD-A achieves the best performance in terms of the AccParity metric. While GS-DP-SGD-TO does not outperform DP-SGD-A on the AccParity metric, it achieves comparable results and consistently maintains a lower AccParity value than DP-SGD across all datasets. This indicates that GS-DP-SGD-TO effectively mitigates the unfairness introduced by DP. Additionally, DP-SGD-F fails to eliminate the impact of privacy on the Law dataset, resulting in an AccParity value identical to that of DP-SGD.

Focus on the DemParity metric. From Tab. 1, on Adult, Dutch, and Compas datasets, it is evident that except for GS-DP-SGD-TO, the DemParity values obtained by the other three methods are significantly high. On the law dataset, except for DP-SGD-F and GS-DP-SGD-TO, the DemParity values for the other two methods remain significantly high. On bank and credit datasets, we observe that any method including SGD and DP-SGD has acceptable DemParity values. This indicates that our approach excels at addressing outcome unfairness especially when models are trained on biased datasets. It is worth noting that since our threshold optimization is performed on the validation set, there might be slight deviations when applied to the test set. That means the fairness constraint may not be fully satisfied, revealing a limitation of our method.

In conclusion, DP-SGD-F lacks stability, failing to reduce the AccParity value or achieve acceptable outcome fairness across all datasets. While both DP-SGD-A and GS-DP-SGD fulfill the AccParity requirements, they fall short of meeting outcome fairness criteria. Conversely, the GS-DP-SGD-TO method effectively reduces AccParity between sensitive groups and consistently meets outcome fairness requirements compared to GS-DP-SGD. Notably, on the majority of datasets, including Adult, Dutch, Bank, and Credit, GS-DP-SGD-TO even achieves a lower AccParity value than GS-DP-SGD.

Table 1. The comparison results of methods to mitigate the unfairness caused by DP, with all evaluation metrics obtained from test datasets. The number in horizontal lines mean the AccParity values obtained by DP-SGD for the convenience of judging whether other methods effectively remove unfairness.

Dataset	**Method**	ϵ	**Acc** (↑)	**AccParity** (↓)	**DemParity** (↓)
Adult	SGD		0.848 ± 0.003	0.114 ± 0.008	0.191 ± 0.007
	DP-SGD	2.654	0.789 ± 0.005	0.154 ± 0.010	0.064 ± 0.005
	DP-SGD-F	2.667	0.829 ± 0.003	0.112 ± 0.008	0.210 ± 0.005
	DP-SGD-A	2.661	0.848 ± 0.003	0.114 ± 0.007	0.191 ± 0.010
	GS-DP-SGD	2.657	0.832 ± 0.005	0.116 ± 0.010	0.199 ± 0.057
	GS-DP-SGD-TO	2.658	0.837 ± 0.005	0.108 ± 0.015	0.047 ± 0.015
Dutch	SGD		0.834 ± 0.003	0.070 ± 0.006	0.335 ± 0.016
	DP-SGD	2.269	0.793 ± 0.005	0.111 ± 0.011	0.231 ± 0.033
	DP-SGD-F	2.280	0.812 ± 0.005	0.092 ± 0.008	0.232 ± 0.024
	DP-SGD-A	2.275	0.834 ± 0.004	0.069 ± 0.006	0.332 ± 0.017
	GS-DP-SGD	2.267	0.805 ± 0.008	0.081 ± 0.015	0.208 ± 0.064
	GS-DP-SGD-TO	2.267	0.786 ± 0.006	0.070 ± 0.026	0.052 ± 0.020
Compas	SGD		0.673 ± 0.014	0.027 ± 0.015	0.291 ± 0.021
	DP-SGD	4.118	0.623 ± 0.025	0.041 ± 0.022	0.170 ± 0.057
	DP-SGD-F	4.204	0.632 ± 0.021	0.035 ± 0.019	0.184 ± 0.044
	DP-SGD-A	4.161	0.678 ± 0.013	0.024 ± 0.016	0.281 ± 0.020
	GS-DP-SGD	4.113	0.676 ± 0.011	0.024 ± 0.020	0.287 ± 0.037
	GS-DP-SGD-TO	4.113	0.668 ± 0.010	0.029 ± 0.017	0.073 ± 0.060
Law	SGD		0.898 ± 0.004	0.160 ± 0.014	0.190 ± 0.018
	DP-SGD	3.671	0.888 ± 0.004	0.202 ± 0.015	0.000 ± 0.000
	DP-SGD-F	3.694	0.888 ± 0.004	0.202 ± 0.015	0.001 ± 0.001
	DP-SGD-A	3.683	0.898 ± 0.004	0.158 ± 0.015	0.182 ± 0.017
	GS-DP-SGD	3.679	0.895 ± 0.004	0.163 ± 0.021	0.139 ± 0.053
	GS-DP-SGD-TO	3.679	0.894 ± 0.004	0.176 ± 0.017	0.049 ± 0.024
Bank	SGD		0.900 ± 0.003	0.036 ± 0.007	0.035 ± 0.004
	DP-SGD	2.831	0.884 ± 0.004	0.050 ± 0.006	0.002 ± 0.001
	DP-SGD-F	2.845	0.887 ± 0.004	0.047 ± 0.007	0.007 ± 0.003
	DP-SGD-A	2.838	0.902 ± 0.003	0.036 ± 0.006	0.037 ± 0.005
	GS-DP-SGD	2.837	0.888 ± 0.005	0.043 ± 0.013	0.048 ± 0.031
	GS-DP-SGD-TO	2.837	0.901 ± 0.004	0.037 ± 0.008	0.026 ± 0.013
Credit	SGD		0.812 ± 0.004	0.024 ± 0.010	0.030 ± 0.006
	DP-SGD	3.365	0.778 ± 0.005	0.031 ± 0.008	0.000 ± 0.000
	DP-SGD-F	3.381	0.779 ± 0.006	0.029 ± 0.008	0.005 ± 0.006
	DP-SGD-A	3.373	0.817 ± 0.005	0.023 ± 0.011	0.033 ± 0.007
	GS-DP-SGD	3.366	0.809 ± 0.004	0.031 ± 0.013	0.035 ± 0.023
	GS-DP-SGD-TO	3.366	0.821 ± 0.006	0.026 ± 0.009	0.020 ± 0.009

Compare with methods of mitigating outcome unfairness for differentially private models. We compare our method with two methods that address the issue of outcome fairness in differentially private models. For FairDP, we obtain different DemParity values by setting various hyperparameters, i.e., the weights clipping bound M (ranging from 0.1 to 1, same as [27]). For DP-SGD-P and GS-DP-SGD-TO, we obtain different DemParity values by adjusting the DemParity constraint values. We apply the three methods on six different datasets and obtain the corresponding AccParity and accuracy(Acc) metrics for each DemParity value shown in Fig. 2 and Fig. 3.

Focus on the AccParity metric. From Fig. 2, we can see that both GS-DP-SGD-TO and DP-SGD-P can achieve a DemParity value of less than 0.1 across all datasets. However, FairDP fails to achieve outcome fairness requirements on the Dutch and Compas datasets as the minimum DemParity value exceeds 0.1, which is unacceptable in practical applications. Additionally, we observe that for the same DemParity value, the GS-DP-SGD-TO consistently achieves the lowest AccParity values and remains below the dashed line on the y-axis across all datasets. That indicates that among the three methods, only GS-DP-SGD-TO reliably mitigates the unfairness caused by DP.

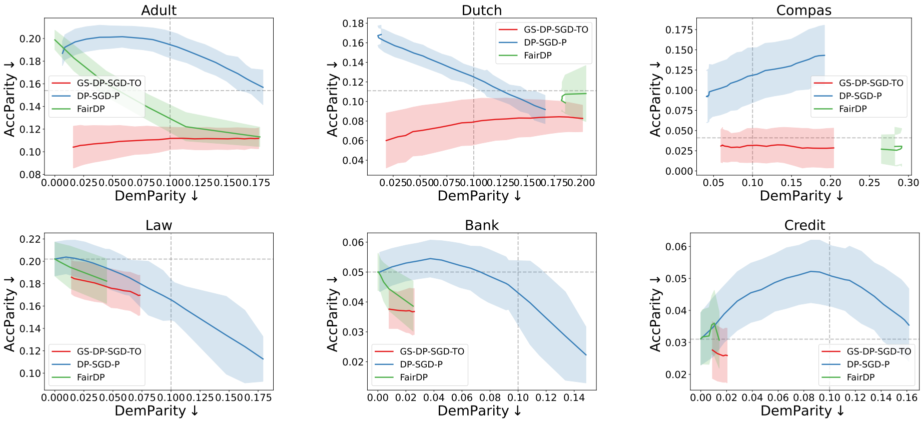

Fig. 2. The comparison results of three approaches on different datasets, with all evaluation metrics obtained from test datasets. The dashed lines on the y-axis represent the AccParity values obtained by DP-SGD, and the dashed lines on the x-axis represent the DemParity values of 0.1.

Focus on the Acc metric. From Fig. 3, it is evident that for almost all datasets (except for the Dutch and Compas datasets, due to the FairDP cannot be compared), GS-DP-SGD-TO achieves the highest accuracy for the same DemParity value. Therefore, we can conclude that compared to the other two methods, GS-DP-SGD-TO not only achieves a better trade-off between the AccParity and DemParity metrics but also demonstrates superior accuracy compared to the other two methods.

In summary, our proposed method GS-DP-SGD-TO exhibits excellent stability across all datasets. It not only effectively addresses the amplified accuracy disparity between sensitive groups due to DP integration but also ensures outcome fairness. Furthermore, it achieves a good trade-off between the group fairness metrics AccParity and DemParity without compromising model performance.

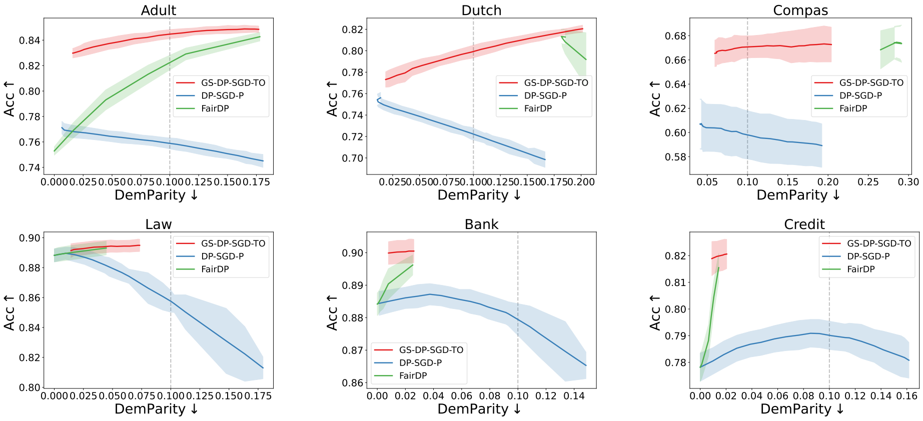

Fig. 3. The comparison results of three approaches on different datasets, with all evaluation metrics obtained from test datasets. The dashed lines on the x-axis represent the DemParity values of 0.1.

5 Conclusion

In our work, we propose a method aiming at simultaneously addressing the exacerbation of accuracy disparity among sensitive groups caused by DP and the issue of outcome unfairness in differentially private models. Our approach involves utilizing the group-specific DP-SGD algorithm during model training to alleviate the accuracy disparity issues associated with traditional DP-SGD. Subsequently, we employ a threshold optimization strategy on the trained model to enhance outcome fairness. Through this method, we successfully mitigate accuracy disparity and outcome unfairness, ultimately achieving fair and private ML models. Extensive experiments have confirmed that our method can effectively tackle both problems across all datasets. Moreover, we achieve a good trade-off between the group fairness metrics AccParity and DemParity with reasonable utility.

Nevertheless, our method has some limitations. The achieved outcome fairness metric may surpass the predetermined constraint value because the threshold selection is based on the validation set. Additionally, our method is currently limited to binary classification tasks and single sensitive attributes. In the future, we will gradually refine our method to address these issues, aiming to ensure it consistently satisfies the constraints, and then extend it to multi-class classification and multifairness.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62250710682), the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001), the Program for Guangdong Introducing Innovative and

Entrepreneurial Teams (Grant No.2017ZT07X386), and the Research Institute of Trustworthy Autonomous Systems.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. 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 volumes] p. 129 (2018)
3. Bagdasaryan, E., Poursaeed, O., Shmatikov, V.: Differential privacy has disparate impact on model accuracy. Advances in neural information processing systems **32** (2019)
4. Chang, H., 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)
5. Cummings, R., Gupta, V., Kimpara, D., Morgenstern, J.: On the compatibility of privacy and fairness. In: Adjunct publication of the 27th conference on user modeling, adaptation and personalization. pp. 309–315 (2019)
6. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012)
7. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25. pp. 486–503. Springer (2006)
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings. vol. 3, pp. 265–284. Springer Berlin Heidelberg, Springer (2006)
9. Ekstrand, M.D., Joshaghani, R., Mehrpouyan, H.: Privacy for all: Ensuring fair and equitable privacy protections. In: Conference on fairness, accountability and transparency. pp. 35–47. PMLR (2018)
10. Esipova, M.S., Ghomi, A.A., Luo, Y., Cresswell, J.C.: Disparate impact in differential privacy from gradient misalignment. arXiv preprint arXiv:2206.07737 (2022)
11. Fioretto, F., Tran, C., Van Hentenryck, P., Zhu, K.: Differential privacy and fairness in decisions and learning tasks: A survey. arXiv preprint arXiv:2202.08187 (2022)
12. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Proceedings of the forty-first annual ACM symposium on Theory of computing. pp. 169–178 (2009)
13. Ghoukasian, H., Asoodeh, S.: Differentially private fair binary classifications. arXiv preprint arXiv:2402.15603 (2024)
14. Hansen, V.P.B., Neerkaje, A.T., Sawhney, R., Flek, L., Søgaard, A.: The impact of differential privacy on group disparity mitigation. arXiv preprint arXiv:2203.02745 (2022)

15. Huang, C., Zhang, Z., Mao, B., Yao, X.: An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence **4**(4), 799–819 (2023)
16. Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., Ullman, J.: Differentially private fair learning. In: International Conference on Machine Learning. pp. 3000–3008. PMLR (2019)
17. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th international conference on data mining. pp. 924–929. IEEE (2012)
18. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), e1452 (2022)
19. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
21. Mironov, I.: Rényi differential privacy. In: 2017 IEEE 30th computer security foundations symposium (CSF). pp. 263–275. IEEE (2017)
22. Pannekoek, M., Spigler, G.: Investigating trade-offs in utility, fairness and differential privacy in neural networks. arXiv preprint arXiv:2102.05975 (2021)
23. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Computing Surveys (CSUR) **55**(3), 1–44 (2022)
24. Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H.B., Vassilvitskii, S., Chien, S., 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)
25. Sanyal, A., Hu, Y., Yang, F.: How unfair is private learning? In: Uncertainty in Artificial Intelligence. pp. 1738–1748. PMLR (2022)
26. Tran, C., Dinh, M., Fioretto, F.: Differentially private empirical risk minimization under the fairness lens. Advances in Neural Information Processing Systems **34**, 27555–27565 (2021)
27. Tran, K., Fioretto, F., Khalil, I., Thai, M.T., Phan, N.: Fairdp: Certified fairness with differential privacy. arXiv preprint arXiv:2305.16474 (2023)
28. Voigt, P., 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)
29. Wang, Z., Huang, C., Yao, X.: Procedural fairness in machine learning. arXiv preprint arXiv:2404.01877 (2024)
30. Xu, D., Du, W., 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)
31. Xu, D., Yuan, S., Wu, X.: Achieving differential privacy and fairness in logistic regression. In: Companion proceedings of The 2019 world wide web conference. pp. 594–599 (2019)
32. Yu, D., Kamath, G., Kulkarni, J., Liu, T.Y., Yin, J., Zhang, H.: Individual privacy accounting for differentially private stochastic gradient descent. arXiv preprint arXiv:2206.02617 (2022)