# Towards Understanding the Fairness of Differentially Private Margin Classifiers

Wenqiang Ruan[1], Mingxin Xu[1], Yinan Jing[1] and Weili Han[1*]

[1*]Laboratory for Data Analytics and Security, Fudan University, 2005, Songhu Road, Shanghai, 200438, China.

*Corresponding author(s). E-mail(s): wlhan@fudan.edu.cn;
Contributing authors: 20110240031@fudan.edu.cn;
20212010078@fudan.edu.cn; jingyn@fudan.edu.cn;

**Abstract**

Margin classifiers, such as Support Vector Machine, are usually critical in the high-stakes decision domains. In recent years, differential privacy has been widely employed in margin classifiers to protect user privacy. However, incorporating differential privacy into margin classifiers might adversely cause the fairness issue in the sense that differentially private margin classifiers have significantly different true positive rates on different groups that are determined by sensitive attributes (e.g., race). In order to address this issue, we are motivated to identify the factor that dominates the fairness of differentially private margin classifiers based on well-designed experiments and further analysis. We first conduct an empirical study on three classical margin classifiers learned via three representative differentially private empirical risk minimization algorithms, respectively. The empirical result shows that the fairness of differentially private margin classifiers strongly depends on the fairness of their non-private versions. We then analyze how differential privacy impacts the fairness of margin classifiers and confirm the empirical study results. In a general sense, our study shows that when non-private margin classifiers are fair, the fairness of their differentially private counterparts can be ensured.

**Keywords:** Margin Classifiers, Differential Privacy, Fairness, Empirical Risk Minimization

# 1 Introduction

Margin classifiers are playing an important role in the high-stakes decision domains (e.g., credit assessment) [1, 2]. Recently, to protect user privacy when training margin classifiers on sensitive data, a number of differentially private empirical risk minimization (ERM) algorithms have been proposed [3–8]. Meanwhile, as an important social concern about machine learning, algorithmic fairness is receiving increasing attention from both public and academia [9, 10]. Among various machine learning models, the fairness of margin classifiers receives significant attention [11–15] for their wide application in high-stakes domains protected by anti-discrimination regulations.

However, previous studies [16, 17] showed that the differentially private ERM algorithms could make machine learning models unfairly treat different groups, such as recognizing black faces and white faces with different accuracy. Here, their studies are mainly empirical and lack an analysis of how differential privacy impacts the fairness of their studied models. As a result, we still did not know the dominant factor on the fairness of the differentially private machine learning models. Identifying the dominant factor would help find a correct way to ensure the fairness of differentially private margin classifiers.

In this paper, we show that the fairness of non-private margin classifiers dominates the fairness of corresponding differentially private margin classifiers based on well-designed experiments and further analysis. We first empirically evaluate the impact of three representative differentially private ERM algorithms [3–5] on the fairness of three classical margin classifiers: Linear support vector machine (SVM), Kernel SVM, and logistic regression (LR). Because in most high-stakes domains, the accuracy of the 'positive' label is more important than that of the 'negative' label [12, 18], we use *equal opportunity* [12], which requires that different groups should have the same true positive rate (TPR), as the fairness notion. By testing three datasets widely used in the algorithmic fairness field, we find that the fairness of differentially private margin classifiers strongly depends on the fairness of their non-private versions. In that sense, when a non-private margin classifier has almost the same TPR on different groups, its differentially private version also has almost the same TPR on these groups. Furthermore, when a non-private margin classifier has a significant TPR gap between two groups, differential privacy will amplify this TPR gap.

We confirm the empirical results through a theoretical analysis of how differential privacy impacts the fairness of margin classifiers. Concretely, we reveal that the main reason for significant TPR gaps in differentially private margin classifiers is that 'positive' data samples of different groups have significantly different margin distributions in their non-private versions, which is implied by TPR gaps. By contrast, when a non-private margin classifier has similar TPR on different groups, the 'positive' data samples from different groups will have similar margin distributions. Consequently, the negative impact brought by differential privacy can be largely ignored, even eliminated. We also show that

our analysis results can be extended to other accuracy-based group fairness notions (e.g., *equal odds* [12]).

In summary, we show that if non-private margin classifiers are fair with negligible TPR gaps, the fairness of their differentially private counterparts can be ensured. As is shown in Section 5.3, when we improve the fairness of non-private margin classifiers with a pre-processing method [11], the TPR gaps of differentially private margin classifiers are close to and even lower than those of non-private margin classifiers.

## 2 Related Work

**Algorithmic Fairness.** Chouldechova et al. [9] presented an overview of current studies on algorithmic fairness. Dwork et al. [19] proposed the notion of *individual fairness*. However, because the similarity of individuals is hard to measure, a series of group fairness notions [12, 19, 20] have been proposed. Based on these fairness notions, several studies proposed the related algorithms to train fair classifiers [11–14]. All of these studies took margin classifiers as typical cases to verify the effectiveness of their algorithms.

**Differential Privacy.** Differential privacy has become a *de facto* standard to protect user privacy of machine learning models. Since Chaudhuri et al. [21] created a novel sensitivity analysis method for convex and continuous loss functions, many differentially private ERM algorithms have been developed to achieve a better privacy-utility trade-off [6, 7, 22, 23], to make differentially private ERM algorithms more usable [5, 24] or to make a non-convex optimization process differentially private [3, 25]. In addition, Jagielski et al. [26] applied differential privacy to protect the sensitive attribute (e.g., gender) of data samples when training a fair classifier.

**Differential Privacy and Algorithmic Fairness.** Cummings et al. [27] showed that perfect fairness and differential privacy are incompatible under non-trivial accuracy. Bagdasaryan et al. [16] empirically revealed that a differentially private stochastic gradient descent algorithm has a disparate impact on the accuracy of different groups. Motivated by the above findings, some related algorithms [28–32] have been proposed to balance privacy protection and fairness on the classification problem, the selection problem, etc. However, there still lacks a comprehensive study on how differential privacy impacts the fairness of margin classifiers, which is critical to design differentially private and fair margin classifiers. Compared with previous studies, our study covers a wider spectrum of differentially private ERM algorithms. What is more, beyond the empirical study, we conduct a theoretical analysis of how differential privacy impacts the fairness of margin classifiers.

## 3 Preliminaries

To present the study results clearly, we list the symbols involved in this paper in Table 1.

**Table 1**: Notations involved in this paper

| Symbol | Description |
|---|---|
| $D$ | Dataset |
| $\ell$ | Loss Function |
| $L$ | Lipschitz constant |
| $\eta$ | Learning rate |
| $k$ | Batch size |
| $T$ | Iteration number |
| $\Lambda$ | Coefficient of $L_2$-regularization |
| $n$ | Size of training dataset |
| $\epsilon, \delta$ | Privacy parameters |
| $\theta$ | Model parameters |
| $p$ | Feature dimension |
| $\lambda, \alpha$ | Deviation parameters |
| $\gamma$ | The upper bound of gradients |

## 3.1 Margin Classifier

**Definition 1** ***Geometric margin*** *[33]. The geometric margin $\rho_h(\vec{x})$ of a linear classifier $h : \vec{x} \to \theta^{\mathsf{T}} \cdot \vec{x}$ at a data sample $\vec{x}$ is its Euclidean distance to the hyperplane whose normal vector is $\theta$:*

$$\rho_h(\vec{x}) = \frac{|\theta^{\mathsf{T}} \cdot \vec{x}|}{\|\theta\|_2}$$

**Margin classifier** [34]. Margin classifiers learn a model by optimizing a loss function that takes margins as inputs (e.g., maximizing the minimum margin). That is, the loss function of any margin classifier can be represented as a composite function of the margin function and a margin loss function $\phi(\rho_h(\vec{x})) : \mathbb{R}^p \to \mathbb{R}^+$, where $p$ is the dimension of input data.

## 3.2 Differentially Private Empirical Risk Minimization Algorithms

We first introduce the definition of neighboring datasets: $D$ and $D^{'} \in D^n$ are neighboring datasets if $D^{'}$ and $D$ differs in one data sample. We then introduce the definition of $(\epsilon, \delta)$-differential privacy as follows.

**Definition 2** $(\epsilon, \delta)$-***differentially privacy*** *[35]. For a random mechanism $M$ whose input is $D \in D^n$ and output is $r \in R$, we say $M$ is $(\epsilon, \delta)$-differentially private if for any subset $S \subseteq R$, $Pr(M(D) \in S) \leq e^{\epsilon} \cdot Pr(M(D^{'}) \in S) + \delta$, where $\epsilon$ is the privacy budget, a tunable parameter on the privacy-utility trade-off.*

The main idea of differential privacy is to bound the influence of each data sample on the output to prevent attackers from inferring any information about one single data sample from the output. A typical way to satisfy the definition of differential privacy is by adding random noise sampled from a predefined distribution to the computing process. If $\delta$ is 0, we say $M$ is $\epsilon$-differentially private.

We can design differentially private ERM algorithms according to the following three paradigms: (1) Objective perturbation (adding random noise to loss functions); (2) Gradient perturbation (adding random noise to gradients); (3) Output perturbation (adding random noise to the final model parameters). To comprehensively study the relationship between differential privacy and fairness of margin classifiers, we test three differentially private ERM algorithms, each of which follows one or two of the above three paradigms.

**Approximate Minimal Perturbation algorithm (AMP)** [4] combines the objective perturbation and the output perturbation paradigms. It thus divides the total privacy budget into two parts (i.e., the noise of objective perturbation and the noise of output perturbation). Note that even though AMP is a hybrid method, more than 99% of the privacy budget should be allocated to the objective perturbation phase as they recommend.

**Differentially Private Stochastic Gradient Descent algorithm (DPSGD)** [3] follows the gradient perturbation paradigm. It adds noise to the clipped gradients. DPSGD can be applied to train non-convex models because it has no assumption on the loss functions.

**Private convex permutation-based Stochastic Gradient Descent algorithm (PSGD)** [5] follows the output perturbation paradigm. The goal of PSGD is to help incorporate differential privacy into existed machine learning systems without modifying the original system. It adds noise to the final model parameters based on the sensitivity analysis on convex and continuous loss functions and the stochastic gradient descent process.

Despite adding the noise at different positions, all of the above differentially private ERM algorithms provide utility guarantees for convex models, which bound the difference between the losses of private and non-private models. They guarantee the utility by bounding the Euclidean distance between the private model parameters $\theta_{priv}$ and non-private model parameters $\theta^*$. As a result, we define $(\lambda, \alpha)$-deviation to quantify the deviation of model parameters led by differential privacy noise.

**Definition 3** $(\lambda, \alpha)$***-deviation.*** *We say a differentially private ERM algorithm is* $(\lambda, \alpha)$*-deviate if it can guarantee that when trained from the same dataset, with the probability at least 1-*$\alpha$*, the* $L_2$ *distance between private model parameters* $\theta_{priv}$ *and non-private model parameters* $\theta^*$ *is less than a given value* $\lambda$*. That is:*

$$Pr(\left\|\theta_{priv} - \theta^*\right\|_2 < \lambda) \geq 1 - \alpha$$

In Definition 3, $\alpha$ bounds the probability that the $L_2$ distance between the private model and the original model is higher than or equal to $\lambda$. We show the the deviation properties of the above three differentially private ERM algorithms in Lemma 1, Lemma 2 and Lemma 3.

**Lemma 1** *AMP follows* $(\frac{n\gamma}{\Lambda} + (\sqrt{2p\log\frac{2}{\alpha}})(\frac{4L}{\Lambda\epsilon_3}(1 + \sqrt{2\log\frac{1}{\delta_1}}) + \frac{n\gamma}{\Lambda\epsilon_2}(1 + \sqrt{2\log\frac{1}{\delta_2}})),\alpha)$*-deviation.*

**Lemma 2** *PSGD follows* $(\frac{2p\ln(p/\alpha)kTL\eta}{n\epsilon},\alpha)$*-deviation.*

**Lemma 3** *When applying DPSGD to optimize a* $\Delta$*-strongly convex and* $L_2$*-Lipchitz continuous loss function, if we set learning rate as* $\frac{1}{\Delta t}$*, DPSGD follows* $(\frac{4(L^2+p\sigma^2)}{\Delta^2 T\alpha},\alpha)$*-deviation.*

The proofs of the above lemmas are shown in Appendix A with the pseudocodes of three differentially private learning algorithms.

### 3.3 Equal Opportunity

Let $D = \{(\vec{x}_1, a_1, y_1), \cdots, (\vec{x}_n, a_n, y_n)\}$ be a dataset that consists of $n$ data samples from an unknown distribution over $(X, A) \times Y$, where $Y = \{+1, -1\}$ is the set of labels, $A$ is the set of sensitive attributes (e.g., gender, race) and $X$ is the set of other features in an input space. In this paper, we use *equal opportunity* [12], which requires that different groups should have the same true positive rate (TPR), as the fairness notion in our study.

Cummings et al. [27] has shown that perfect fairness and differential privacy are incompatible under non-trivial accuracy. We thus use *$\rho$-True Positive Rate Disparity* to measure the degree of fairness of a classifier.

**Definition 4** ***$\rho$-True Positive Rate Disparity*** *[36]. For any* $a_i, a_j$ $(i \neq j) \in A$ *and a classifier* $h_\theta$*, we say* $h_\theta$ *satisfies* $\rho$*-True Positive Rate Disparity if and only if* $|Pr\{h_\theta(\vec{x}_i, a_i) = +1 | y_i = +1\} - Pr\{h_\theta(\vec{x}_j, a_j) = +1 | y_j = +1\}| \leq \rho$*. Here* $\rho$ *is the maximum TPR difference among all groups.*

## 4 Empirical Study

In this section, we evaluate the impact of differential privacy on the fairness of margin classifiers by applying AMP, DPSGD, PSGD to train three classical margin classifiers: Linear SVM, Kernel SVM and LR, respectively. We try to answer the following research questions: **Are differentially private ERM algorithms bound to aggravate the TPR gaps of margin classifiers? If not, which factor dominates the aggravation of the TPR gaps?** The answers to these questions would help find a correct way to ensure the fairness of differentially private margin classifiers.

### 4.1 Experiment Setup

**Datesets.** We transform all data samples to one-hot encoded form and shuffle them before the training process. Then we take the first 80%

**Table 2**: Overview of datasets.

| DataSet | #Sample | Sensitive Attribute | Positive Label |
|---|---|---|---|
| `Compas` | 5,915 | Race | No Recidivism in Two Years |
| `Adult` | 45,220 | Gender | Income Higher than 50k Dollars |
| `Default` | 30,000 | Gender | No Default Payment |

as the training dataset and the rest 20% as the testing dataset. There are six datasets (`Compas`[1], `Adult` [37], `Default` [37], `German` [37], `Student` [37], `Arrhythmia` [37]) that are widely used in the algorithmic fairness field. Considering the size of datasets (larger than 1,000), we employ three datasets (`Compas`, `Adult`, `Default`) in our empirical study. The overview of these three datasets is shown in Table 2. (1) `Compas` dataset contains 7,214 data samples. The binary label indicates whether an offender recidivates within two years after the screening. We set 'No Recidivism in Two Years' as the 'positive' label and Race as the sensitive attribute. After filtering the data samples with null attributes and selecting the data samples whose races are African-American (black) or Caucasian (white), we obtain 5,915 data samples. (2) `Adult` dataset contains 45,220 data samples. The binary label indicates whether the income of one citizen is higher than 50k dollars. We set 'Income Higher than 50k Dollars' as the 'positive' label and Gender as the sensitive attribute. (3) `Default` dataset contains 30,000 data samples. The binary label indicates whether one user has a default payment. We set 'No Default Payment' as the 'positive' label and Gender as the sensitive attribute. Note that even with large variances, the results of the rest three datasets give the same answer to the research questions with three employed datasets. We discuss them in Appendix B.

**Algorithm Implementation and Hyperparameter Configuration.** We implement AMP, DPSGD, PSGD based on the open-source code[2] released by Iyengar et al. [4]. All of three algorithms have at least four hyperparameters. To comprehensively study the relationship between differential privacy and fairness of margin classifiers, we conduct a grid search procedure to find the best hyperparameter configuration, which means that under the hyperparameter configuration, private models acquire the highest average test accuracy given a privacy budget. In addition, we independently train ten models for each hyperparameter configuration and average the TPR gaps between groups of these ten models as the final result. We also plot the error bars of the test results to show the statistical significance of our results. We list all potential values of hyperparameters in Table 3.

**Privacy Parameters.** To comprehensively study the impact of differential privacy on the fairness of margin classifiers, we test eight $\epsilon$ values (from 1 to 8), which covers most privacy budget values used in practice. In addition, following the settings of previous works [4, 5], we set another privacy parameter

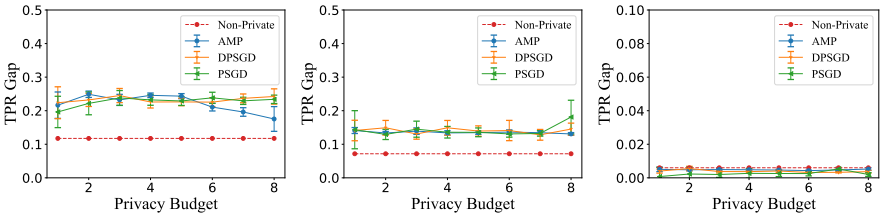
[1] https://github.com/propublica/Compas-analysis
[2] https://github.com/sunblaze-ucb/dpml-benchmark

**Table 3**: Potential hyperparameter values for the grid search procedure.

| Hyperparameter | Potential Values |
|---|---|
| $\Lambda$ (regularization factor) | 0, 0.001, 0.01, 0.05 |
| $\eta$ (learning rate) | 0.001, 0.01, 0.1, 1, 10 |
| $T$(iteration number) | 5, 10, 100, 500, 1000 |
| $f$ (output budget fraction of AMP) | 0.001, 0.01, 0.1, 0.5 |
| $f_1$ (privacy budget fraction of AMP) | 0.9, 0.95, 0.98, 0.99 |
| $L$ (clipping threshold) | 0, 0.05, 0.1, 1, 10 |

**Table 4**: Potential Privacy parameters.

| Privacy Parameter | Potential Values |
|---|---|
| $\epsilon$ | 1, 2, 3, 4, 5, 6, 7, 8 |
| $\delta$ | $\frac{1}{n^2}$ |




(a) `Compas` (b) `Adult` (c) `Default`

**Fig. 1**: TPR gaps of non-private and differentially private Linear SVM models.

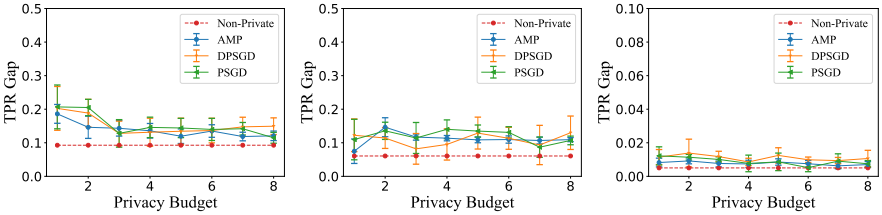
$\delta$ as $\frac{1}{n^2}$, where $n$ is the size of the training dataset. The potential values of privacy parameters are shown in Table 4.

**Sample Clipping.** All three differentially private ERM algorithms require that the loss functions should be $L_2$-Lipschitz continuous [4]. We achieve it by bounding the $L_2$ norm of each data sample. Before the training process, we clip the feature vector of each data sample $(\vec{x}_i, a_i)$ to $(\vec{x}_i, a_i) \cdot min(1, \frac{L}{\|(\vec{x}_i, a_i)\|_2})$.

## 4.2 Experimental Results

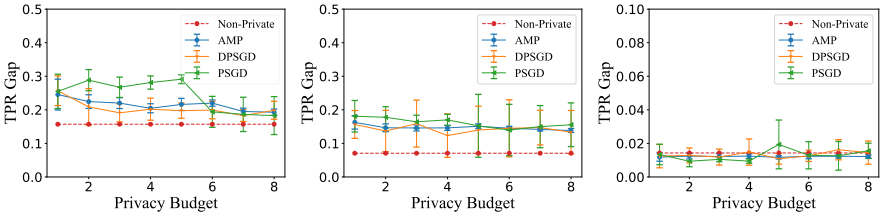
**Linear Support Vector Machine.** We obtain the non-private baselines by training $L_2$ regularized Linear Huber SVM models [38]. Then we train differentially private Linear SVM models via AMP, DPSGD and PSGD on same training datasets.

As is shown in Figure 1, the average TPR gaps of all private models trained on `Compas` and `Adult` datasets are larger than those of the non-private models. In contrast, the average TPR gaps of all private models trained on `Default` dataset are close to that of the non-private model. The TPR gap between the white samples and black samples of the non-private model trained on `Compas`




(a) $\texttt{Compas}_{dim=220,std=0.9}$ (b) $\texttt{Adult}_{dim=245,std=0.1}$ (c) $\texttt{Default}_{dim=120,std=0.3}$

**Fig. 2**: TPR gaps of non-private and differentially private Kernel SVM models, where dim and std are the parameters of kernel function approximation method.




(a) `Compas` (b) `Adult` (c) `Default`

**Fig. 3**: TPR gaps of non-private and differentially private LR models.

dataset is about 0.117 (more than 19 times of `Default`); the TPR gaps between the male samples and female samples of the non-private models trained on `Adult` and `Default` datasets are about 0.072 (12 times of `Default`), 0.006, respectively.

**Kernel Support Vector Machine.** We implement the non-private Kernel SVM and its differentially private versions through a Fourier transform-based function approximation method proposed by Rahimi et al. [39]. This method uses random cosine functions to approximate the kernel functions that project the original features to a high-dimension target space. Therefore, two additional parameters are involved in the Kernel SVM implementation: the dimension number of the target space (dim), the standard variance of random cosine functions (std). We approximate the Gaussian kernel function [33] here and use a grid search procedure to determine the values of these two parameters. Then we train Linear SVM models on the projected high-dimension features.

As is shown in Figure 2, the private models trained on `Compas` and `Adult` datasets all have larger average TPR gaps than the non-private models. Note that the TPR gaps of non-private models trained on `Compas` and `Adult` datasets are about 19 and 12 times more than that of `Default` dataset. While

**Table 5**: Overview of imbalanced datasets.

| DataSet | #Sample | Size ratio of the majority group to the minority group |
|---|---|---|
| `Compas` | 4,245 | 5:1 (Black: White) |
| `Adult` | 33,579 | 10:1 (Male: Female) |
| `Default` | 19,924 | 10:1 (Female: Male) |

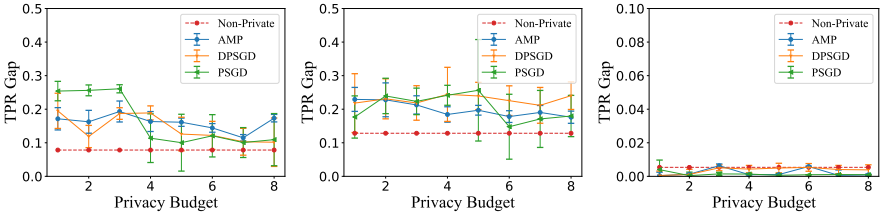
in `Default` dataset, the average TPR gaps of private models are similar to that of the non-private models. Meanwhile, as the privacy budget changes, the size of TPR gaps fluctuates up and down, which shows that the TPR gap changes are accidental errors introduced by the randomness of noise sampling.

**Logistic Regression.** We obtain the non-private baseline by training a $L_2$ regularized LR model on the same training datasets with private models. As is shown in Figure 3, the private models trained on `Compas` and `Adult` datasets all have larger average TPR gaps than the non-private models. By contrast, when the TPR gap in the non-private model is small (0.014 in `Default` dataset, about 1/11 and 1/5 of `Compas` and `Adult` datasets), the TPR gaps in private models are almost the same as that of the non-private model.

**Insights.** By analyzing the experimental results of three classical margin classifiers learned via three differentially private ERM algorithms over three widely used datasets, we can conclude that differentially private ERM algorithms are not bound to have a disparate impact on the TPR of different groups. That is, when the TPR gaps of non-private models are small enough (such as 0.006 in `Default` dataset by Linear SVM), differential privacy will not aggravate the TPR gaps of margin classifiers. On the other hand, when non-private models have significant TPR gaps between groups (such as 0.117 in `Compas` dataset and 0.072 in `Adult` dataset by Linear SVM), all differentially private ERM algorithms amplify the TPR gaps. In addition, in `Compas` dataset, the number of black samples is about 1.5 times that of white samples, but the TPR of black samples drops much more than white samples in private models. The result shows that differential privacy only amplifies the bias in the dataset rather than discriminates the minority group of the dataset. We will further justify this claim in Section 4.3.

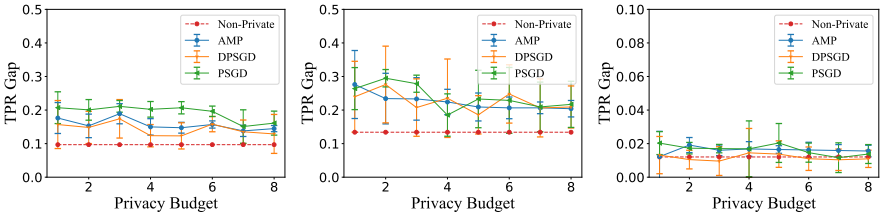
## 4.3 Impact of Data Imbalance

Bagdasaryan et al. [16] stated that differential privacy noise would cause less accuracy loss on majority groups and more accuracy loss on minority groups in differentially private neural network models. In order to test whether this claim is applicable in margin classifiers, we subsample the minority group of three datasets studied in Section 4 to construct imbalanced datasets. The details of constructed imbalanced datasets are shown in Table 5. Note that we set the size ratio of `Compas` dataset as 5:1 because it has much fewer samples than the other two datasets. Thus the testing results will have large variances if we set it as 10:1. We then train non-private and differentially private margin




(a) `Compas` (b) `Adult` (c) `Default`

**Fig. 4**: TPR gaps of non-private and differentially private SVM models trained on imbalanced datasets.




(a) `Compas` (b) `Adult` (c) `Default`

**Fig. 5**: TPR gaps of non-private and differentially private LR models trained on imbalanced datasets.

classifiers over these imbalanced datasets with the same grid search procedure used in Section 4. The testing results are shown in Figure 4 and Figure 5. In `Compas`, where the number of black samples is five times that of white samples, when non-private classifiers have significantly higher TPR on white samples, the differential privacy still enlarges the TPR gap between white samples and black samples. On the other hand, in `Default`, even though the number of female samples is ten times that of male samples, when non-private classifiers have similar TPR on different groups, differential privacy has a similar impact on these groups. The above results show that data imbalance has little impact on the accuracy loss of differentially private ERM algorithms caused on different groups.

## 5 Analysis of Impact Mechanism

In this section, we analyze how differentially private ERM algorithms impact the TPR gaps of margin classifiers. We synthesize a two-dimensional dataset to show the intuition behind our analysis in Figure 6. For clarity purposes, we only illustrate the 'positive' samples. As is shown in Figure 6, in the non-private model, Group1 has a higher TPR than Group2 (i.e., Group1 has more true positive (TP) samples and fewer false negative (FN) samples than Group2).

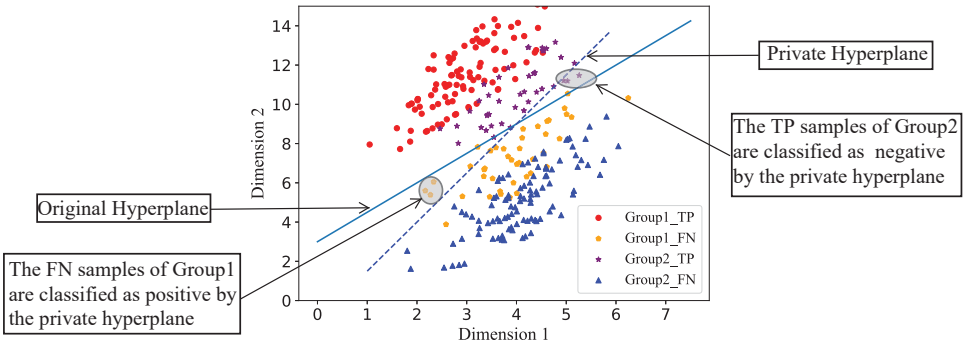



**Fig. 6**: An overview of the analysis of Section 5. Each point represents a data sample. The color and shape of one point indicate the group and type the data sample belongs to.

The TPR gap between Group1 and Group2 implies different margin distributions of their TP and FN data samples, i.e., the margins of TP data samples of Group2 mainly distribute on lower values (closer to the original hyperplane), while the margins of FN data samples of Group1 mainly distribute on lower values (Section 5.1). When the private hyperplane deviates from the original hyperplane, more TP samples of Group2 are misclassified as negative and more FN samples of Group1 are correctly classified as positive (Section 5.2). As a result, the TPR gap between these two groups is aggravated (Section 5.3).

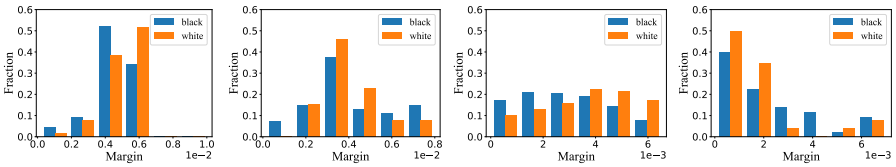
## 5.1 Bridging TPR Gap and Margin Gap

In this section, we show that if one group has a significantly higher TPR than another group in a non-private margin classifier, the margins of the group's TP data samples will distribute on higher values, while the margins of the group's FN data samples will distribute on lower values. We first analyze the correlation between the margin and the loss of one data sample. The loss functions of standard linear SVM [33] and LR [33] are:

$$loss_{SVM}(\theta, \vec{x}_i, y_i) = \begin{cases} max(0, 1 - \theta^T \vec{x}_i) & y_i = +1 \\ max(0, 1 + \theta^T \vec{x}_i) & y_i = -1 \end{cases}$$

$$loss_{LR}(\theta, \vec{x}_i, y_i) = \begin{cases} \log(1 + e^{-(\theta^T \vec{x}_i)}) & y_i = +1 \\ \log(1 + \frac{1}{e^{-(\theta^T \vec{x}_i)}}) & y_i = -1 \end{cases}$$

where $|\theta^T \vec{x}_i| = margin_{\vec{x}_i} * \|\theta\|_2$ according to Definition 1. Without loss of generality, we discuss the situation where $y_i = +1$ here. By the definitions of the above loss functions, when a data sample $\vec{x}_i$ is correctly classified (i.e., $\theta^T \vec{x}_i > 0$ ), a larger margin implies a smaller value of the loss function. Conversely, when $\vec{x}_i$ is wrongly classified (i.e., $\theta^T \vec{x}_i < 0$), a smaller margin implies




(a) SVM_TP (b) SVM_FN (c) LR_TP (d) LR_FN

**Fig. 7**: Margin distribution of TP samples and FN samples of `Compas` dataset on non-private Linear SVM and LR models.

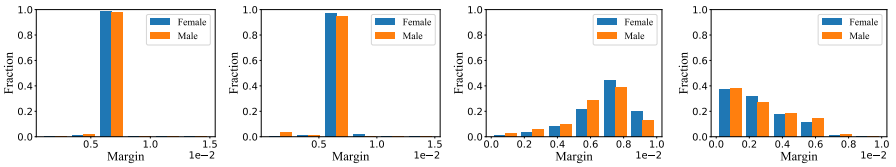
a smaller value of $|\theta^T \vec{x}_i|$ (i.e., $-\theta^T \vec{x}_i$), thus a smaller value of the loss function. Consequently, if the average loss of one group (refer to as $g_a$) is lower than another group (refer to as $g_b$), at least one of the following two situations will happen: (1) The correctly classified data samples of $g_a$ have a larger average margin than correctly classified data samples of $g_b$. (2) The wrongly classified data samples of $g_a$ have a smaller average margin than wrongly classified data samples of $g_b$. An concrete example of the above $g_a$ and $g_b$ is Group 1 and Group 2 in Figure 6.

If one group has a higher TPR than another one, its 'positive' data samples should have a lower average loss than the other one. Therefore, the TPR gap between groups inevitably implies the margin distribution difference between their TP data samples (situation (1)) or their FN data samples (situation (2)), even both simultaneously. On the other hand, if two groups have similar TPR, their 'positive' samples should have a similar loss and thus have a similar margin distribution.

To further verify the above analysis results, we plot the frequency histograms of data samples' margins to show the margin distributions of `Compas` and `Default` datasets in Figure 7 and 8. Because the only difference between Linear SVM and Kernel SVM is that the former is trained on original features and the latter is trained on projected high-dimension features, the results of Linear SVM can be generalized to Kernel SVM. In Linear SVM and LR models trained on `Compas` dataset, the TPR gaps between white samples and black samples are about 0.117 and 0.157, respectively. Consequently, the margins of TP black samples mainly distribute on lower values than the white ones, while the margins of FN white samples mainly distribute on the lower values than the black ones. By contrast, in `Default` dataset, where the TPR gaps of two non-private margin classifiers are both less than 0.015, the margin distributions of different groups' TP and FN samples are very similar.

## 5.2 Impact of Margin Gap

We then show that when the private hyperplane deviates from the original hyperplane, the TP samples with smaller margins are more likely to be wrongly classified as negative, and the FN samples with smaller margins are more likely to be correctly classified as positive.




(a) SVM_TP (b) SVM_FN (c) LR_TP (d) LR_FN

**Fig. 8**: Margin distribution of TP samples and FN samples of `Default` dataset on non-private Linear SVM and LR models.

**Theorem 1** *Let $m$ denote the margin of one data sample $\vec{x}$ to the original hyperplane whose normal vector is $\theta^*$. If $m$ is greater than $\frac{\lambda L}{\|\theta^*\|_2}$, then with the probability less than or equal to $\alpha$, the private model $\theta_{priv}$ trained by a differentially private ERM algorithm that is $(\lambda, \alpha)$-deviate makes a different prediction with the original model on $\vec{x}$, i.e.,*

$$Pr((\theta^{*\mathsf{T}} \cdot \vec{x})(\theta_{priv}^{\mathsf{T}} \cdot \vec{x}) < 0) < \alpha$$

*where $L$ is the upper bound of data samples' $L_2$ norm.*

*Proof*

$$\begin{aligned}(\theta^{*\mathsf{T}} \cdot \vec{x})(\theta_{priv}^{\mathsf{T}} \cdot \vec{x}) &= (\theta^{*\mathsf{T}} \cdot \vec{x})((\theta^* + \theta_{priv} - \theta^*)^{\mathsf{T}} \cdot \vec{x}) \\ &= (\theta^{*\mathsf{T}} \cdot \vec{x})(\theta^{*\mathsf{T}} \cdot \vec{x} + (\theta_{priv} - \theta^*)^{\mathsf{T}} \cdot \vec{x})\end{aligned}$$

According to Cauchy-Schwarz inequality,

$$|(\theta_{priv} - \theta^*)^{\mathsf{T}} \cdot \vec{x}| \leq \left\|\theta_{priv} - \theta^*\right\|_2 \cdot \|\vec{x}\|_2 < \lambda L$$

As we stated in Section 4.1, to ensure the loss functions are $L_2$-Lipchitz continuous, the $L_2$ norm of all data samples are not larger than $L$. Therefore,

$$\|\vec{x}\|_2 \leq L$$

Meanwhile, according to the deviation property of differentially private learning algorithms, with the probability at least 1-$\alpha$,

$$\left\|\theta_{priv} - \theta^*\right\|_2 < \lambda$$

According to the definition of margin, $\rho_h(\vec{x}) \geq \frac{\lambda L}{\|\theta^*\|_2}$ implies that $|\theta^{*\mathsf{T}} \cdot \vec{x}| \geq \lambda L$. Therefore, the sign of $(\theta^{*\mathsf{T}} \cdot \vec{x} + (\theta_{priv} - \theta^*)^{\mathsf{T}} \cdot \vec{x})$ would be consistent with the sign of $(\theta^{*\mathsf{T}} \cdot \vec{x})$ with probability at least 1-$\alpha$. Thus,

$$Pr((\theta^{*\mathsf{T}} \cdot \vec{x})(\theta_{priv}^{\mathsf{T}} \cdot \vec{x}) < 0) < \alpha$$

□

According to Definition 3 and deviation properties of three differentially private ERM algorithms identified in Section 3.2, a smaller deviation $\lambda$ implies a higher $\alpha$. Meanwhile, in Theorem 1, a smaller $m$ implies a smaller $\lambda$. Consequently, the bound of Theorem 1 shows that a differentially private margin

classifier $\theta_{priv}$ is more likely to make a different prediction with the non-private model on one data sample that has a lower $m$. As is shown in Figure 6, when the hyperplane deviates from its original position, the data samples that are closer to the original hyperplane are more likely to be classified as different classes. When a private model makes different predictions with the non-private model on them, TP samples suffer accuracy loss, while FN samples gain accuracy. Therefore, Theorem 1 shows that the hyperplane deviation led by the differential privacy noise causes more accuracy losses to the TP data samples that are closer to the original hyperplane, and more accuracy gains to the FN data samples that are closer to the original hyperplane.

### 5.3 Deep Analysis of Empirical Results

With the analysis results from Section 5.1 and Section 5.2, we analyze the empirical results from Section 4 as follows.

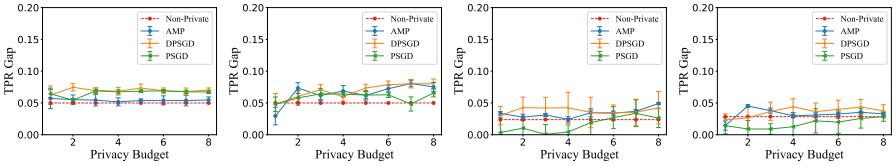
According to Section 5.1, the TPR gap between groups implies different margin distributions of these groups. Concretely, the group with a higher TPR would have more TP data samples whose margins distribute on high values and more FN data samples whose margins distribute on low values. Meanwhile, as the bound of Theorem 1 shows, when the original hyperplane is deviated by differential privacy noise, the group with a higher TPR will suffer less accuracy loss on TP data samples and gain more accuracy on FN data samples. Therefore, the significant TPR gaps of non-private margin classifiers trained on `Compas` and `Adult` datasets are amplified in their differentially private versions.

By contrast, if a non-private margin classifier has almost the same TPR on different groups, the 'positive' data samples of these groups will have similar margin distributions. Then the TP and FN data samples of different groups will obtain similar bounds in Theorem 1. As a result, the differentially private version of the margin classifier has almost the same TPR on these groups, too.

To further verify the effectiveness of the above results, we use a preprocessing method proposed by Donini et al. [11] to mitigate the biases that exist in `Compas` and `Adult` datasets. Then we train non-private and private Linear SVM and LR models on debiased datasets. The TPR gap testing results are shown in Figure 9. When we reduce the TPR gaps of the non-private models trained on `Compas` dataset from 0.117, 0.157 to 0.050, 0.050, the negative impact brought by differential privacy is largely mitigated, even eliminated. In `Adult` dataset, when we reduce the TPR gaps of non-private models from 0.072, 0.071 to 0.024, 0.028, the TPR gaps of private models are very similar to those of non-private models. These results further show that the fairness of differentially private margin classifiers strongly depends on the fairness of their non-private versions.

## 6 Discussion and Future Work

**Non-convex models.** Currently, domains that are protected by anti-discrimination laws are mainly high-stakes, such as credit assessment and




(a) `Compas` (SVM) (b) `Compas` (LR) (c) `Adult` (SVM) (d) `Adult` (LR)

**Fig. 9**: TPR gaps of non-private and private margin classifiers trained on `Compas` and `Adult` datasets that have been pre-processed by the method proposed by [11].

criminal justice. Deep learning models would still be far from being widely deployed in these domains due to their lack of interpretability and robustness [40–42]. Therefore, we focus on the fairness of differentially private margin classifiers in this paper. Besides, current differentially private ERM algorithms for non-convex models still lack rigorous utility guarantees. As a result, we put identifying the deviation properties of non-convex models as our future work.

**Extending our results to other accuracy-based fairness notions.** We have shown that the TPR gap of a non-private margin classifier implies the margin distribution difference between TP samples or FN samples of different groups. According to the qualitative analysis on the loss functions of SVM and LR, we can obtain the same result with the TPR gap when it comes to the true negative rate gap or the total accuracy gap. That is, a true negative rate gap or a total accuracy gap between two groups would also imply the different margin distributions of corresponding data samples. As Theorem 1 only assumes the margin of a data sample, the results of our paper can be extended to other accuracy-based fairness notions, including *equal odds* [12], which requires that the different groups should have the same true negative rate and true positive rate, and *accuracy parity* [18], which requires the different groups should have the same accuracy.

**Future work.** In the future, we will quantitatively analyze the correlation between the TPR gap and margin distribution difference among groups in the non-private margin classifier to understand the impact of differential privacy on the fairness of margin classifiers more deeply.

# 7 Conclusion

In this paper, we study the dominant factor on the fairness of differentially private margin classifiers. Through conducting a well-designed empirical study and analyzing how differential privacy impacts the fairness of margin classifiers, we show that the fairness of differentially private margin classifiers strongly depends on the fairness of their non-private counterparts. To summarize, we argue that if non-private margin classifiers are fair with negligible TPR gaps, the fairness of their differentially private versions can be ensured.

**Acknowledgements.** This paper is supported by the National Key R&D Program of China (2019YFE0103800) and Natural Science Foundation of China (U1836207).

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

[1] de Paula, D.A.V., Artes, R., Ayres, F., Minardi, A.: Estimating credit and profit scoring of a brazilian credit union with logistic regression and machine-learning techniques. RAUSP Management Journal **54**, 321–336 (2019)

[2] Zhang, L., Hu, H., Zhang, D.: A credit risk assessment model based on svm for small and medium enterprises in supply chain finance. Financial Innovation **1**(14), 1–21 (2015)

[3] 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)

[4] Iyengar, R., Near, J.P., Song, D., Thakkar, O., Thakurta, A., Wang, L.: Towards practical differentially private convex optimization. In: Proceedings of 2019 IEEE Symposium on Security and Privacy (SP), pp. 299–316 (2019). IEEE

[5] Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., Naughton, J.: Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In: Proceedings of the 2017 ACM International Conference on Management of Data, pp. 1307–1322 (2017)

[6] Yu, D., Zhang, H., Chen, W., Liu, T.-Y.: Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In: ICLR 2021 (2021)

[7] Zhou, Y., Wu, S., Banerjee, A.: Bypassing the ambient dimension: Private {sgd} with gradient subspace identification. In: International Conference on Learning Representations (2021)

[8] Huang, X., Ding, Y., Jiang, Z.L., Qi, S., Wang, X., Liao, Q.: Dp-fl: a novel differentially private federated learning framework for the unbalanced data. World Wide Web **23**(4), 2529–2545 (2020)

[9] Chouldechova, A., Roth, A.: A snapshot of the frontiers of fairness in machine learning. Communications of the ACM **63**(5), 82–89 (2020)

[10] Ranjbar Kermany, N., Zhao, W., Yang, J., Wu, J., Pizzato, L.: A fairness-aware multi-stakeholder recommender system. World Wide Web **24**(6), 1995–2018 (2021)

[11] Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., Pontil, M.: Empirical risk minimization under fairness constraints. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18, pp. 2796–2806. Curran Associates Inc., Red Hook, NY, USA (2018)

[12] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, pp. 3315–3323 (2016)

[13] Mandal, D., Deng, S., Jana, S., Wing, J., Hsu, D.J.: Ensuring fairness beyond the training data. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 18445–18456 (2020)

[14] Roh, Y., Lee, K., Whang, S.E., Suh, C.: Fairbatch: Batch selection for model fairness. In: International Conference on Learning Representations (2021)

[15] Hu, R., Zhu, X., Zhu, Y., Gan, J.: Robust svm with adaptive graph learning. World Wide Web **23**(3), 1945–1968 (2020)

[16] Bagdasaryan, E., Poursaeed, O., Shmatikov, V.: Differential privacy has disparate impact on model accuracy. In: Advances in Neural Information Processing Systems, pp. 15479–15488 (2019)

[17] Farrand, T., Mireshghallah, F., Singh, S., Trask, A.: Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In: Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice. PPMLP'20, pp. 15–19. Association for Computing Machinery, New York, NY, USA (2020)

[18] Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021)

[19] 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)

[20] Hebert-Johnson, U., Kim, M., Reingold, O., Rothblum, G.: Multicalibration: Calibration for the (Computationally-identifiable) masses. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1939–1948 (2018)

[21] Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research **12**(3) (2011)

[22] Bassily, R., Smith, A., Thakurta, A.: Private empirical risk minimization: Efficient algorithms and tight error bounds. In: 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, pp. 464–473 (2014). IEEE

[23] Su, D., Cao, J., Li, N., Bertino, E., Lyu, M., Jin, H.: Differentially private k-means clustering and a hybrid approach to private optimization. ACM Transactions on Privacy and Security (TOPS) **20**(4), 1–33 (2017)

[24] Jain, P., Kothari, P., Thakurta, A.: Differentially private online learning. In: Proceedings of Conference on Learning Theory, pp. 24–1 (2012)

[25] Bu, Z., Dong, J., Long, Q., Su, W.J.: Deep learning with gaussian differential privacy. Harvard data science review **2020**(23) (2020)

[26] 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 (2019). PMLR

[27] 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. UMAP'19 Adjunct, pp. 309–315. Association for Computing Machinery, New York, NY, USA (2019)

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

[29] Khalili, M.M., Zhang, X., Abroshan, M., Sojoudi, S.: Improving fairness and privacy in selection problems. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)

[30] Mozannar, H., Ohannessian, M.I., Srebro, N.: Fair learning with private demographic data. arXiv preprint arXiv:2002.11651 (2020)

[31] Tran, C., Fioretto, F., Hentenryck, P.V.: Differentially private and fair deep learning: A lagrangian dual approach. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 9932–9939 (2021)

[32] Xu, D., Du, W., Wu, X.: Removing disparate impact of differentially private stochastic gradient descent on model accuracy. arXiv preprint arXiv:2003.03699 (2020)

[33] Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning, (2012)

[34] Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Large margin classifiers: convex loss, low noise, and convergence rates. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1173–1180 (2004)

[35] Dwork, C.: Differential privacy. In: Proceedings of Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Part II, pp. 1–12 (2006)

[36] Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. WWW '17, pp. 1171–1180. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017)

[37] Dua, D., Graff, C.: UCI Machine Learning Repository (2017). http://archive.ics.uci.edu/ml

[38] Cherkassky, V., Ma, Y.: Practical selection of svm parameters and noise estimation for svm regression. Neural networks **17**(1), 113–126 (2004)

[39] Rahimi, A., Recht, B.: Uniform approximation of functions with random bases. In: 2008 46th Annual Allerton Conference on Communication, Control, and Computing, pp. 555–561 (2008)

[40] Heaven, D.: Why deep-learning ais are so easy to fool. Nature, 163–166 (2019)

[41] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 206–215 (2019)

[42] Xue, M., He, C., Wang, J., Liu, W.: One-to-n & n-to-one: Two advanced

backdoor attacks against deep learning models. IEEE Transactions on Dependable and Secure Computing (2020)

[43] Dasgupta, S., Schulman, L.: A probabilistic analysis of em for mixtures of separated, spherical gaussians. J. Mach. Learn. Res. **8**, 203–226 (2007)

[44] Rakhlin, A., Shamir, O., Sridharan, K.: Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Coference on International Conference on Machine Learning. ICML'12, pp. 1571–1578. Omnipress, Madison, WI, USA (2012)

# A Deviation Properties of AMP, PSGD, DPSGD

In this section, we first identify the deviation properties of AMP, PSGD and DPSGD. Then we show the detailed proof of Theorem 1.

**Algorithm 1** Approximate Minima Perturbation [4]

**Input:** Data set: $D = \{d_1, \cdots, d_n\}$; loss function: $\ell(\theta;\ d_i)$ with $L_2$-Lipschitz constant $L$, is convex in $\theta$, has a continous Hessian, and is $\beta$-smooth for all $\theta \in \mathcal{R}^p$ and all $d_i$; Hessian rank bound parameter: $r$; privacy parameters: $(\epsilon, \delta)$; gradient norm bound: $\gamma$.

1: Set $\epsilon_1, \epsilon_2, \epsilon_3, \delta_1, \delta_2 > 0$ such that $\epsilon = \epsilon_1 + \epsilon_2$, $\delta = \delta_1 + \delta_2$, and $0 < \epsilon_1 - \epsilon_3 < 1$

2: Set $\Lambda \geq \frac{r\beta}{\epsilon_1 - \epsilon_3}$

3: $b_1 \sim \mathcal{N}(0, \sigma_1^2 I_{p\times p})$, where $\sigma_1 = \frac{(\frac{2L}{n})(1+\sqrt{2\log\frac{1}{\delta_1}})}{\epsilon_3}$

4: Let $\mathcal{L}_{priv}(\theta;\ D) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta;\ d_i) + \frac{\Lambda}{2n}\|\theta\|_2^2 + b_1^T\theta$

5: $\theta_{approx} \leftarrow \theta$ such that $\|\nabla\mathcal{L}_{priv}(\theta, D)\|_2 \leq \gamma$

6: $b_2 \sim \mathcal{N}(0, \sigma_2^2 I_{p\times p})$, where $\sigma_2 = \frac{(\frac{n\gamma}{\Lambda})(1+\sqrt{2\log\frac{1}{\delta_2}})}{\epsilon_2}$

7: Output $\theta_{out} = \theta_{approx} + b_2$

The pseudocodes of AMP are shown in Algorithm 1. According to the design of AMP, we identify its deviation property as follows.

**Proof of Lemma 1.** AMP follows $(\frac{n\gamma}{\Lambda} + (\sqrt{2p\log\frac{2}{\alpha}})(\frac{4L}{\Lambda\epsilon_3}(1+\sqrt{2\log\frac{1}{\delta_1}}) + \frac{n\gamma}{\Lambda\epsilon_2}(1+\sqrt{2\log\frac{1}{\delta_2}})),\alpha)$-deviation.

*Proof* The utility guarantee of AMP contains two parts. First, it bounds the distance between optimal model parameters $\theta_{approx}$ under private loss function and optimal model parameters $\theta^*$ under non-private loss function. Second, it bounds the distance between private output $\theta_{out}$ and $\theta_{approx}$. The first bound is $\frac{2n\left\|\vec{b_1}\right\|_2}{\Lambda}$ (see inequality

10 of [4]). The second bound is $\frac{n\gamma}{\Lambda} + \left\|\vec{b_2}\right\|_2$ (see inequality 5 of [4]). Therefore, the total bound of the deviation of model parameters is $\frac{n(\gamma+2\left\|\vec{b_1}\right\|_2)}{\Lambda} + \left\|\vec{b_2}\right\|_2$, where $\vec{b_1}$ and $\vec{b_2}$ are distributed as $\mathcal{N}(0, \sigma_1^2 I_{p\times p}), \mathcal{N}(0, \sigma_2^2 I_{p\times p})$, here $\sigma_1 = \frac{\frac{2L}{n}(1+\sqrt{2\log\frac{1}{\delta_1}})}{\epsilon_3}$, $\sigma_2 = \frac{\frac{n\gamma}{\Lambda}(1+\sqrt{2\log\frac{1}{\delta_2}})}{\epsilon_2}$.

According to Lemma 2 in [43]: with probability $\geq 1 - \frac{\alpha}{2}$,

$$\left\|\vec{b_s}\right\|_2 \leq \sigma_s \sqrt{2p\log\frac{2}{\alpha}} \ ,$$

AMP thus follows $(\frac{n\gamma}{\Lambda} + (\sqrt{2p\log\frac{2}{\alpha}})(\frac{4L}{\Lambda\epsilon_3}(1+\sqrt{2\log\frac{1}{\delta_1}}) + \frac{n\gamma}{\Lambda\epsilon_2}(1+\sqrt{2\log\frac{1}{\delta_2}})),\alpha)$-deviation. □

**Algorithm 2** Differentially Private Permutation-based Stochastic Gradient Descent [5]

**Input:** Data set $D = \{d_1, \cdots, d_n\}$, loss function: $\ell(\theta;\ d_i)$ with $L_2$-Lipschitz constant $L$, privacy parameters: $(\epsilon, \delta)$, number of iterations: $T$, batch size: $k$, constant learning rate: $\eta$.

$\theta_1 = 0^p$
Let $\tau$ be a random permutation of $[n]$
**for** $t = 1$ *to* $T-1$ **do**
  **for** $b = 1$ *to* $\frac{n}{k}$ **do**
    Let $s_1 = d_{\tau(bk)}, \cdots, s_k = d_{\tau(b(k+1)-1)}$
    $\theta \leftarrow \theta - \eta(\frac{1}{k}\sum_{i=1}^{k} \nabla\ell(\theta;\ s_i))$
  **end for**
**end for**
$\sigma^2 \leftarrow \frac{8T^2L^2\eta^2\log(\frac{2}{\delta})}{k^2\epsilon^2}$
$b \sim \mathcal{N}(0, \sigma^2 I_{p\times p})$
Output $\theta_{priv} = \theta + b$

We show the pseudocodes of PSGD in Algorithm 2. We then identify the deviation property of PSGD.

**Proof of Lemma 2.** PSGD follows $(\frac{2p\ln(p/\alpha)kTL\eta}{n\epsilon},\alpha)$-deviation.

*Proof* The sensitivity of PSGD is $\frac{2kTL\eta}{n}$ (see Corollary 1 in [5]). As the noise is directly added on the final model, the Euclidean distance between private model and non-private model is the $L_2$ norm of the noise, which is distributed as Gamma distribution $\Gamma$(p, $\frac{2kTL\eta}{n\epsilon}$). According to Theorem 2 in [5]: for the noise vector $\kappa$, whose $L_2$ norm is distributed according to the Gamma distribution $\Gamma(p, \Delta)$, we have that with probability at least 1-$\alpha$, $\|\kappa\|_2 \leq p\Delta ln(\frac{p}{\alpha})$. Therefore, PSGD follows $(\frac{2p\ln(p/\alpha)kTL\eta}{n\epsilon},\alpha)$-deviation. □

We then identify the deviation property of DPSGD under a strong convexity and continuity assumption on loss functions. The pseudocodes of DPSGD are shown in Algorithm 3.

**Algorithm 3** Differentially Private Stochastic Gradient Descent [3, 22]

**Input:** Data set $D = \{d_1, \cdots, d_n\}$, loss function: $\ell(\theta;\ d_i)$ with $L_2$-Lipschitz constant $L$, privacy parameters: $(\epsilon, \delta)$, number of iterations: $T$, batch size: $k$, learning rate function: $\eta : [T] \to \mathbb{R}$.

$\sigma^2 \leftarrow \frac{16L^2 T log\frac{1}{\delta}}{n^2\epsilon^2}$
$\theta_1 = 0^p$
**for** $t = 1$ *to* $T - 1$ **do**
  $s_1, \cdots, s_k \leftarrow$ Sample $k$ samples uniformly with replacement from $D$
  $b_t \sim \mathcal{N}(0, \sigma^2 I_{p\times p})$
  $\theta_{t+1} = \theta_t - \eta(t)[\frac{1}{k}\sum_{i=1}^{k} \nabla\ell(\theta;\ s_i) + b_t]$
**end for**
Output $\theta_T$

**Proof of Lemma 3.** When applying DPSGD to optimize a $\Delta$-strongly convex and $L_2$-Lipchitz continuous loss function, if we set learning rate as $\frac{1}{\Delta t}$, DPSGD follows $(\frac{4(L^2+p\sigma^2)}{\Delta^2 T\alpha}, \alpha)$-deviation.

*Proof* Let $G_t$ as the gradient at iteration $t$, according to Theorem 2.4 of [22],

$$\mathbb{E}[\|G_t\|_2^2] \leq L^2 + p\sigma^2$$

Then according to Lemma 1 of [44],

$$\mathbb{E}[\|\theta_t - \theta^*\|_2] \leq \frac{4(L^2+p\sigma^2)}{\Delta^2 t}$$

Finally, according to Markov inequality,

$$Pr(\left\|\theta_{priv} - \theta^*\right\|_2 \leq \frac{4(L^2+p\sigma^2)}{\Delta^2 T\alpha}) \geq 1 - \alpha$$

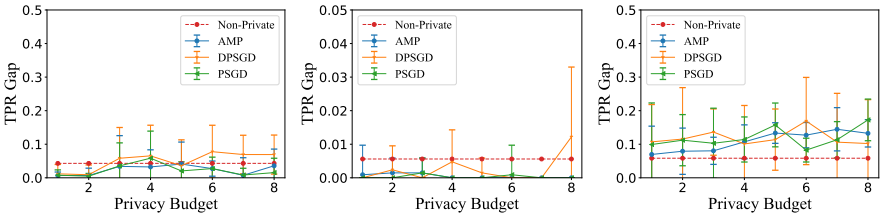
□

The deviation properties of AMP, PSGD and DPSGD show that $\lambda$ is inversely propertional to $\alpha$. Therefore, they deviate private hyperplane from the original hyperplane little with high probability.

For convenience, we discuss the situation where $y_i$= +1 here (the result can be generalized to the situation where $y_i$= -1). By the definition of the loss function of LR, when a data sample $\vec{x}_i$ is correctly classified (i.e., $\theta^T\vec{x}_i > 0$), a larger margin of $\vec{x}_i$ implies a smaller value of loss function. Conversely, when $\vec{x}_i$ is wrongly classified (i.e., $\theta^T\vec{x}_i < 0$), a larger margin of $\vec{x}_i$ implies a larger value of loss function. The following analysis is the same as that of Linear SVM.

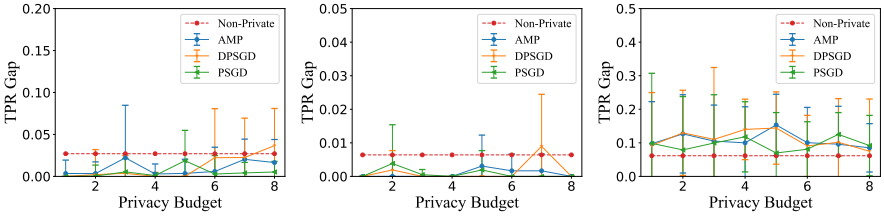
| DataSet | #Sample | Sensitive Attribute | Positive Label |
|---|---|---|---|
| `German` | 1,000 | Gender | Good Credit Risk |
| `Student` | 649 | Gender | Course Grade Higher than 10 |
| `Arrhythmia` | 452 | Gender | No Cardiac Arrhythmia |

**Table 6**: Overview of supplementary datasets.




(a) `German` (b) `Student` (c) `Arrhythmia`

**Fig. 10**: TPR gaps of non-private and differentially private SVM models.

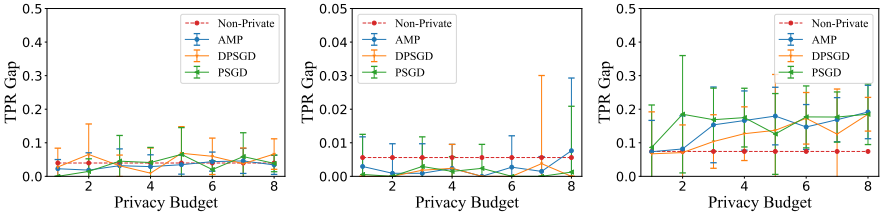



(a) $\texttt{German}_{dim=185,std=0.3}$ (b) $\texttt{Student}_{dim=225,std=0.1}$ (c) `Arrhythmia` $_{dim=390,std=0.3}$

**Fig. 11**: TPR gaps of non-private and differentially private Kernel SVM models, where dim and std are the parameters of kernel functions approximation method.

# B Empirical Results of the Rest Three Datasets

We train Linear SVM, Kernal SVM and LR models on `German`, `Student`, `Arrhythmia` datasets under the same setting with that of Section 4. The test results are shown in Figure 10, Figure 11 and Figure 12. Even though with large variances, from the average results, we can find that when a significant TPR gap exists in the non-private model, the private models will have larger TPR gaps. On the other hand, when the TPR gaps of non-private models are negligible, the private models will have similar, even smaller, TPR gaps with the non-private models. We then explain why the TPR gaps of margin classifiers trained on these three datasets have such large variances.




(a) `German` (b) `Student` (c) `Arrhythmia`

**Fig. 12**: TPR gaps of non-private and differentially private LR models.

The overview of `German, Student, Arrhythmia` datasets is shown in Table 6. The sizes of these three datasets are all less than 1,000. Consequently, the sizes of their testing datasets are less than or equal to 200. Even though labels are balanced distributed and different groups have the same number of data samples, the number of 'positive' samples of each group in testing datasets is less than or equal to 50. Therefore, the inversion of one data sample's prediction changes the TPR of the corresponding group by at least 2%. As a result, the test results of these datasets are greatly impacted by the randomness of noise sampling, and all have large variances.