Removing Disparate Impact on Model Accuracy in Differentially Private Stochastic Gradient Descent

Depeng Xu
University of Arkansas
Fayetteville, AR, USA
depengxu@uark.edu

Wei Du
University of Arkansas
Fayetteville, AR, USA
wd005@uark.edu

Xintao Wu
University of Arkansas
Fayetteville, AR, USA
xintaowu@uark.edu

ABSTRACT

In differentially private stochastic gradient descent (DPSGD), gradient clipping and random noise addition disproportionately affect underrepresented and complex classes and subgroups. As a consequence, DPSGD has disparate impact: the accuracy of a model trained using DPSGD tends to decrease more on these classes and subgroups vs. the original, non-private model. If the original model is unfair in the sense that its accuracy is not the same across all subgroups, DPSGD exacerbates this unfairness. In this work, we study the inequality in utility loss due to differential privacy, which compares the changes in prediction accuracy w.r.t. each group between the private model and the non-private model. We analyze the cost of privacy w.r.t. each group and explain how the group sample size along with other factors is related to the privacy impact on group accuracy. Furthermore, we propose a modified DPSGD algorithm, called DPSGD-F, to achieve differential privacy, equal costs of differential privacy, and good utility. DPSGD-F adaptively adjusts the contribution of samples in a group depending on the group clipping bias such that differential privacy has no disparate impact on group accuracy. Our experimental evaluation shows the effectiveness of our removal algorithm on achieving equal costs of differential privacy with satisfactory utility.

CCS CONCEPTS

• **Security and privacy**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Machine learning algorithms**;

KEYWORDS

differential privacy, fairness, stochastic gradient descent

ACM Reference Format:
Depeng Xu, Wei Du, and Xintao Wu. 2021. Removing Disparate Impact on Model Accuracy in Differentially Private Stochastic Gradient Descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467268

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467268

1 INTRODUCTION

Most researches on fairness-aware machine learning study whether the predictive decision made by machine learning model is discriminatory against the protected group [16, 24, 30, 31]. For example, demographic parity requires that a prediction is independent of the protected attribute. Equality of odds [16] requires that a prediction is independent of the protected attribute conditional on the original outcome. These fairness notions focus on achieving non-discrimination within one single model. In addition to the within-model fairness, cross-model fairness also arises in differential privacy preserving machine learning models when we compare the accuracy loss incurred by private model between the majority group and the protected group. Recently, research in [3] shows that the reduction in accuracy incurred by deep private models disproportionately impacts underrepresented subgroups. The unfairness in this cross-model scenario is that the reduction in accuracy due to privacy protection is discriminatory against the protected group.

In this paper, we study the inequality in utility loss due to differential privacy, which compares the changes in prediction accuracy w.r.t. each group between the private model and the non-private model. Differential privacy guarantees that the query results or the released model cannot be exploited by attackers to derive whether one particular record is present or absent in the underlying dataset [11]. When we enforce differential privacy onto a regular non-private model, the model trades some utility off for privacy. On one hand, with the impact of differential privacy, the within-model unfairness in the private model may be different from the one in the non-private model [7, 17, 29]. On the other hand, differential privacy may introduce additional discriminative effect towards the protected group when we compare the private model with the non-private model. The utility loss between the private and non-private models w.r.t. each group, such as reduction in group accuracy, may be uneven. The intention of differential privacy should not be to introduce more accuracy loss on the protected group regardless of the level of within-model unfairness in the non-private model.

There are several empirical studies on the relationship between the utility loss due to differential privacy and groups with different represented sample sizes. Research in [3] shows that the accuracy of private models tends to decrease more on classes that already have lower accuracy in the original, non-private model. In their case, the direction of inequality in utility loss due to differential privacy is the same as the existing within-model discrimination against the underrepresented group in the non-private model, i.e. "the poor become poorer". Research in [8] shows the similar observation that the contribution of rare training examples is hidden by random noise in differentially private stochastic gradient descent,

and that random noise slows down the convergence of the learning algorithm. Research in [18] shows different observations when they analyze if the performance on emotion recognition is affected in an imbalanced way for the models trained to enhance privacy. They find that while the performance is affected differently for the subgroups, the effect is not consistent across multiple setups and datasets. In their case, there is no consistent direction of inequality in utility loss by differential privacy against the underrepresented group. Hence, the impact of differential privacy on group accuracy is more complicated than the observation in [3] (details in Section 4.1). It needs to be cautionary to conclude that differential privacy introduces more utility loss on the underrepresented group. The bottom line is that the objective of differential privacy is to protect individual's privacy instead of introducing unfairness in the form of inequality in utility loss w.r.t. groups. Though the privacy metric increases when a model is adversarially trained to enhance privacy, we need to ensure that the performance of the model on that dataset does not harm one subgroup more than the other.

In this work, we first analyze the inequality in utility loss by differential privacy. We use "cost of privacy" to refer to the utility loss between the private and non-private models as the result of the utility-privacy trade-off. We study the cost of privacy w.r.t. each group in comparison with the whole population and explain how the group sample size is related to the privacy impact on group accuracy along with other factors (Section 4.2). The difference in group sample sizes leads to the difference in average group gradient norms, which results in different group clipping biases under the uniform clipping bound. It costs less utility trade-off to achieve the same level of differential privacy for the group with larger group sample size and/or smaller group clipping bias. In other words, the group with smaller group sample size and/or larger group clipping bias incurs more utility loss when the algorithm achieves the same level of differential privacy w.r.t. each group. Furthermore, we propose a modified DPSGD algorithm, called DPSGD-F, to remove the potential inequality in utility loss among groups (Section 5.2). DPSGD-F adjusts the contribution of samples in a group depending on the group clipping bias. For the group with smaller cost of privacy, their contribution is decreased and the achieved privacy w.r.t. their group is stronger; and vise versa. As a result, the final utility loss is the same for each group, i.e. differential privacy has no disparate impact on group accuracy in DPSGD-F. Our evaluation shows the effectiveness of our removal algorithm on achieving equal costs of privacy with satisfactory utility (Section 6).

2 RELATED WORKS

2.1 Differential Privacy

Existing literature in differentially private machine learning targets both convex and non-convex optimization algorithms and can be divided into three main classes, input perturbation, output perturbation, and inner perturbation. Input perturbation approaches [9] add noise to the input data based on local differential privacy model. Output perturbation approaches [4] add noise to the model after the training procedure finishes, i.e. without modifying the training algorithm. Inner perturbation approaches modify the learning algorithm such that the noise is injected during learning. For example, research in [6] modifies the objective of the training procedure and research in [1] adds noise to the gradient output of each step of the training without modifying the objective. Research in [2] criticizes that limiting users to small contributions keeps the noise level low at the cost of introducing bias. They characterize the trade-off between bias and variance, and show that (1) a proper bound can be found depending on properties of the dataset and (2) a concrete cost of privacy cannot be avoided simply by collecting more data. Several works study how to adaptively bound the contributions of users and to clip the model parameters to improve learning accuracy and robustness. Research in [26] uses coordinate-wise adaptive clipping of the gradient to achieve the same privacy guarantee with much less added noise. In the federated learning setting, the proposed approach [27] adaptively clips to a value at a specified quantile of the distribution of update norms, where the value at the quantile is itself estimated online, with differential privacy. Other than adaptive clipping, research in [25] adaptively injects noise into features based on the contribution of each to the output so that the utility of deep neural networks under differential privacy is improved; research in [23] adaptively allocates per-iteration privacy budget to achieve zCDP on gradient descent.

2.2 Fairness-aware Machine Learning

Many methods have been proposed to modify the training data for mitigating biases and achieving fairness. These methods include: Massaging [19], Reweighting [5], Sampling [20], Disparate Impact Removal [14], Causation-based Removal [32] and Fair Representation Learning [12]. Some researches propose to mitigate discriminative bias in model predictions by adjusting the learning process [30] or changing the predicted labels [16]. Recent studies [24, 31] also use adversarial learning techniques to achieve fairness in classification and representation learning. Research in [22] uses adaptive sensitive reweighting to recognize sources of bias and to diminish their impact without affecting features or labels. Research in [21] uses agnostic learning to achieve good accuracy and fairness on all subgroups. However, these techniques cannot be directly combined with DPSGD because it is highly sensitive with unbounded sensitivity to find optimal balancing strategy.

2.3 Differential Privacy and Fairness

Recent works study the connection between achieving privacy protection and fairness. Research in [10] proposed a notion of fairness that is a generalization of differential privacy. Research in [15] developed a pattern sanitization method that achieves k-anonymity and fairness. Most recently, the position paper [13] argued for integrating recent research on fairness and non-discrimination to socio-technical systems that provide privacy protection. Later on, several works studied how to achieve within-model fairness (demographic parity [29], equality of odds [17], equality of opportunity [7]) in addition to enforcing differential privacy in the private model. Our work in this paper studies how to prevent disparate impact of the private model on model accuracy across different groups.

3 PRELIMINARY

Let D be a dataset with n tuples $x_1, x_2, \cdots, x_n$, where each tuple x_i includes the information of a user i on d unprotected attributes $A_1, A_2, \cdots, A_d$, protected attribute S, and decision Y. Let D^k denote

a subset in D with tuples with $S = k$. Given a set of examples D, the non-private model outputs a classifier $\eta(a; w)$ with parameter w which minimizes the loss function $\mathcal{L}_D(w) = \frac{1}{n}\sum_{i=1}^{n} L_i(w)$. The optimal model parameter w^* is defined as: $w^* = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} L_i(w)$. A differentially private algorithm outputs a classifier $\tilde{\eta}(a; \tilde{w})$ by selecting $\tilde{w}$ in a manner that satisfies differential privacy while keeping it close to the actual optimal w^*.

3.1 Differential Privacy

Differential privacy guarantees output of a query q be insensitive to the presence or absence of one record in a dataset.

Differential privacy [11]. A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ϵ, δ)-differentially private if, for any pair of datasets $D, D' \in \mathcal{D}$ that differ in exactly one record, and for any subset of outputs $\mathcal{O} \in \mathcal{R}$, we have $\Pr(\mathcal{M}(D) \in \mathcal{O}) \leq \exp(\epsilon) \cdot \Pr(\mathcal{M}(D') \in \mathcal{O}) + \delta$.

The parameter ϵ denotes the privacy budget, which controls the amount by which the distributions induced by D and D' may differ. The parameter δ is a broken probability. Smaller values of ϵ and δ indicate stronger privacy guarantee.

Global sensitivity [11]. Given a query $q : \mathcal{D} \to \mathbb{R}$, the global sensitivity Δ_f is defined as $\Delta_f = \max_{D,D'} |q(D) - q(D')|$.

The global sensitivity measures the maximum possible change in $q(D)$ when one record in the dataset changes.

The Gaussian mechanism with parameter σ adds Gaussian noise $N(0, \sigma^2)$ to each component of the model output.

Gaussian mechanism [11]. Let $\epsilon \in [0, 1]$ be arbitrary. For $c^2 > 2\log(1.25/\delta)$, the Gaussian mechanism with parameter $\sigma > c\Delta_f/\epsilon$ satisfies (ϵ, δ)-differential privacy.

3.2 Differentially Private SGD

The procedure of deep learning model training is to minimize the output of a loss function through numerous stochastic gradient descent (SGD) steps. Research in [1] proposed a differentially private SGD algorithm (DPSGD). DPSGD uses a clipping bound on l_2 norm of individual updates, aggregates the clipped updates, and then adds Gaussian noise to the aggregate. This ensures that the iterates do not overfit to any individual user's update.

The privacy leakage of DPSGD is measured by (ϵ, δ), i.e. computing a bound for the privacy loss ϵ that holds with certain probability δ. Each iteration t of DPSGD can be considered as a privacy mechanism $\mathcal{M}_t$ that has the same pattern in terms of sensitive data access. [1] further proposed a moment accounting mechanism which calculates the aggregate privacy bound when performing SGD for multiple steps. The moments accountant computes tighter bounds for the privacy loss compared to the standard composition theorems. The moments accountant is tailored to the Gaussian mechanism and employs the log moment of each $\mathcal{M}_t$ to derive the bound of the total privacy loss.

To reduce noise in private training of neural networks, DPSGD [1] truncates the gradient of a neural network to control the sensitivity of the sum of gradients. This is because the sensitivity of gradients and the scale of the noise would otherwise be unbounded. To fix this, a cap C on the maximum size of a user's contribution is adopted (Line 7 in Algorithm 1). This will bias our estimated sum but also reduce the amount of added noise, as the sensitivity of the sum is now C. One question is how to choose the truncation level for the gradient norm. If set too high, the noise level may be so great that any utility in the result is lost. If set too low, a large amount of gradients will be forced to clip. DPSGD simply suggests using the median of observed gradients. [2] investigated this bias-variance trade-off and showed that the limit we should choose is the $(1 - 1/b\epsilon)$-quantile of the gradients themselves. It does not matter how large or small the gradients are above or below the cutoff, only that a fixed number of values are clipped.

Algorithm 1 DPSGD (Dataset D, loss function $\mathcal{L}_D(w)$, learning rate r, batch size b, noise scale σ, clipping bound C)

1: **for** $t \in [T]$ **do**
2: Randomly sample a batch B_t with $|B_t| = b$ from D
3: **for** each sample $x_i \in B_t$ **do**
4: $g_i = \nabla L_i(w_t)$
5: **for** each sample $x_i \in B_t$ **do**
6: $\bar{g}_i = g_i \times \min\left(1, \frac{C}{|g_i|}\right)$
7: $\tilde{G}_B = \frac{1}{b}\left(\sum_i \bar{g}_i + N(0, \sigma^2 C^2 \mathbf{I})\right)$
8: $\tilde{w}_{t+1} = \tilde{w}_t - r\tilde{G}_B$
9: Return $\tilde{w}_T$ and accumulated (ϵ, δ)

4 DISPARATE IMPACT ON MODEL ACCURACY

4.1 Preliminary Observations

To explain why DPSGD has disparate impact on model accuracy w.r.t. each group, [3] constructs an unbalanced MNIST dataset to study the effects of gradient clipping, noise addition, the size of the underrepresented group, etc. Training on the data of the underrepresented subgroups produces larger gradients, thus clipping reduces their learning rate and the influence of their data on the model. They also show random noise addition has the biggest impact on the underrepresented inputs. However, [18] reports inconsistent observations on whether differential privacy has negative discrimination towards the underrepresented group in terms of reduction in accuracy. To complement their observations, we use the unbalanced MNIST dataset used in [3] to reproduce their result, and we also use two benchmark census datasets (Adult and Dutch) in fair machine learning to study the inequality of utility loss due to differential privacy (Setup details in Section 6.1). Table 1 (row 4 and 5) shows the model accuracy w.r.t. the total population, the majority group and the minority group for SGD and DPSGD. Table 2 (row 4 and 5) shows the average loss and average gradient norm w.r.t. groups at the last training epoch for SGD and DPSGD. We summarize our key findings below.

On MNIST, the minority group (class 8) has significantly larger utility loss than the other groups in private model. In Table 1, DPSGD only results in −0.0707 decrease in accuracy on the well-represented classes but accuracy on the underrepresented class drops −0.6807, exhibiting a disparate impact on the underrepresented class. Table 2 shows that the small sample size reduces both the convergence rate and the optimal utility of class 8 in DPSGD in comparison with the non-private SGD. The model is far from converging, yet clipping and noise addition do not let it move closer

to the minimum of the loss function. Furthermore, the addition of noise, whose magnitude is similar to the updating vector, prevents the clipped gradients of the underrepresented class from sufficiently updating the relevant parts of the model. Training with more epochs does not reduce this gap while exhausting the privacy budget. Differential privacy also slows down the convergence and degrades the utility for each group. Hence, DPSGD introduces negative discrimination against the minority group (which already has lower accuracy in the non-private SGD model) on MNIST. This matches the observation in [3].

On Adult and Dutch, we have different observations from MNIST. The Adult dataset is an unbalanced dataset, where the female group is underrepresented. Even though the male group is the majority group, it has lower accuracy in the SGD and more utility loss in DPSGD than the female group (Table 1). The Dutch dataset is a balanced dataset, where the group sample sizes are similar for male and female. However, DPSGD introduces more negative discrimination against the male group (Table 1) and its direction (male group loses more accuracy due to DP) is even opposite to the direction of within-model discrimination (female group has less accuracy in SGD). Table 2 shows that the average gradient norm is much higher for the male group in DPSGD on the two datasets. It is not simply against the group with smaller sample size or lower accuracy in the SGD. Hence, Differential privacy does not always introduce more accuracy loss to the minority group on Adult and Dutch. This matches the observation in [18].

From the preliminary observations, we learn that the disproportionate effect from differential privacy is not guaranteed towards the underrepresented group or the group with "poor" accuracy. Why does differential privacy cause inequality in utility loss w.r.t. each group? It may depend on more than just the represented sample size of each group: the classification model, the mechanism to achieve differential privacy, and the relative complexity of data distribution of each group subject to the model. One common observation across all settings is that the group with more utility loss has larger gradients and worse convergence. The underrepresented class 8 has average gradient norm of over 100 and bad utility in DPSGD. The male group has much larger average gradient norm than the female group in DPSGD on both Adult and Dutch datasets. It is important to address the larger gradients and worse convergence directly in order to mitigate inequality in utility loss.

4.2 Cost of Privacy w.r.t. Each Group

In this section, we conduct analysis on the cost of privacy from the viewpoint of a single batch, where the utility loss is measured by the expected error of the estimated private gradient w.r.t. each group. For ease of discussion, our analysis follows [2] that investigates the bias-variance trade-off due to clipping in DPSGD with Laplace noise. Suppose that B_t is a collection of b samples, $x_1, \cdots, x_b$. Each x_i corresponds to a sample and generates the gradient g_i. We would like to estimate the average gradient G_B from B_t in a private way while minimizing the objective function.

We denote the gradient before clipping $G_B = \frac{1}{b}\sum_{i=1}^{b} g_i$, the gradient after clipping but before adding noise $\bar{G}_B = \frac{1}{b}\sum_{i=1}^{b} \bar{g}_i$, and the gradient after clipping and adding noise $\tilde{G}_B = \frac{1}{b}(\sum_{i=1}^{b} \bar{g}_i + Lap(\frac{C}{\epsilon}))$. The expected error of the estimate $\tilde{G}_B$ consists of a variance term (due to the noise) and a bias term (due to the contribution limit):

$$\mathbb{E}|\tilde{G}_B - G_B| \le \mathbb{E}|\tilde{G}_B - \bar{G}_B| + |\bar{G}_B - G_B| \le \frac{1}{b}\frac{C}{\epsilon} + \frac{1}{b}\sum_{i=1}^{b}\max(0, |g_i| - C).$$

In the above derivation, we base the fact that the mean absolute deviation of a Laplace variable is equal to its scale parameter. We can find the optimal C by noting that the bound is convex with subderivative $\frac{1}{\epsilon} - |\{i : g_i > C\}|$, thus the minimum is achieved when C is equal to the $\lceil 1/\epsilon \rceil$th largest value in gradients.

The expected error is tight as we have

$$\mathbb{E}|\tilde{G}_B - G_B| \ge \frac{1}{2}\left[\frac{1}{b}\frac{C}{\epsilon} + \frac{1}{b}\sum_{i=1}^{b} max(0, |g_i| - C)\right].$$

In other words, the limit we should choose is just the $(1 - 1/b\epsilon)$-quantile of the gradients themselves.

For the same batch of samples, we derive the cost of privacy w.r.t. each group. Suppose the batch of samples B_t are from K groups and group k has sample size b^k. We have $G_B^k = \frac{1}{b^k}\sum_{i=1}^{b^k} g_i^k$ and $G_B = \frac{1}{b}\sum_{k=1}^{K} b^k G_B^k$. DPSGD bounds the sensitivity of gradient by clipping each sample's gradient with a clipping bound C. $\bar{G}_B^k = \frac{1}{b^k}\sum_{i=1}^{b^k} \bar{g}_i^k = \frac{1}{b^k}\sum_{i=1}^{b^k} g_i^k \times \min(1, \frac{C}{|g_i^k|})$. Then, DPSGD adds Laplace noise on the sum of clipped gradients. $\tilde{G}_B^k = \frac{1}{b^k}(b^k \tilde{G}_B^k + Lap(\frac{C}{\epsilon}))$. The expected error of the estimate $\tilde{G}_B^k$ consists of a variance term (due to the noise) and a bias term (due to the contribution limit):

$$\begin{aligned}\mathbb{E}|\tilde{G}_B^k - G_B^k| &\le \mathbb{E}|\tilde{G}_B^k - \bar{G}_B^k| + |\bar{G}_B^k - G_B^k| \\ &\le \frac{1}{b^k}\frac{C}{\epsilon} + \frac{1}{b^k}\sum_{i}^{b^k}\max(0, |g_i^k| - C) = \frac{1}{b^k}\frac{C}{\epsilon} + \frac{1}{b^k}\sum_{i}^{m^k}(|g_i^k| - C),\end{aligned} \quad (1)$$

where $m^k = |\{i : |g_i^k| > C\}|$ is the number of examples that get clipped in group k. Similarly, we can get the tight bound w.r.t. each group k is $\mathbb{E}|\tilde{G}_B^k - G_B^k| \ge \frac{1}{2}\left[\frac{1}{b^k}\frac{C}{\epsilon} + \frac{1}{b^k}\sum_{i}^{b^k}\max(0, |g_i^k| - C)\right]$.

From Equation 1, we know the utility loss of group k, measured by the expected error of the estimated private gradient, is bounded by two terms, the bias $\frac{1}{b^k}\sum_{i}^{b^k}\max(0, |g_i^k| - C)$ due to contribution limit (depending on the size of gradients and the size of clipping bound) and the variance of the noise $\frac{1}{b^k}\frac{C}{\epsilon}$ (depending on the scale of the noise). Next, we discuss their separate impacts in DPSGD.

Given the clipping bound C, the bias due to clipping w.r.t. the group with large gradients is larger than the one w.r.t. the group with small gradients. Before clipping, the group with large gradients has large contribution in the total gradient G_B in SGD, but it is not the case in DPSGD. The direction of the total gradient after clipping $\bar{G}_B$ is closer to the direction of the gradient of the group with small bias (small gradients) in comparison with the direction of the total gradient before clipping G_B. Due to clipping, the contribution and convergence of the group with large gradients are reduced.

The added noise increases the variance of the model gradient, as it tries to hide the influence of a single record on the model. It slows down the convergence rate of the model. Because the noise scales $\frac{C}{\epsilon}$ and the sensitivity of clipped gradients C are the same for all groups, the noisy gradients of all groups achieve the same level

of differential privacy ϵ. The direction of the noise is random, i.e. it does not favor a particular group in expectation.

Overall in DPSGD, the group with large gradients has larger cost of privacy, i.e. they have more utility loss to achieve ϵ level of differential privacy under the same clipping bound C.

We can also consider the optimal choice of C which is $(1-\frac{1}{b\epsilon})$-quantile for the whole batch. For each group, the optimal choice of C^k is $(1-\frac{1}{b^k\epsilon})$-quantile for group k. The distance between C and C^k is not the same for all groups, and C is closer to the choice of C^k for the group with small bias (small gradients).

Now we look back on the observations in Section 4.1. On MNIST, the group sample size affects the convergence rate for each group. The group with large sample size (the majority group, class 2) has larger contribution in the total gradient than the group with small sample size (the minority group, class 8), and therefore it leads to a relatively faster and better convergence. As the result, the gradients of the minority group are larger than the gradients of the majority group later on. In their case, the small sample size is the main cause of large gradient norm and large utility loss in class 8. On Adult and Dutch, the average bias due to clipping for each group is different because the distributions of gradients are quite different. The average gradient norm of the male group is larger than the average gradient norm of the female group, even though the male group is not underrepresented. As the result, the male group's contribution is limited due to clipping and it has larger utility loss in DPSGD. In there case, the group sample size is not the only reason to cause difference in the average gradient norm, and the other factors (e.g. the relative complexity of data distribution of each group subject to the model) out-weighs sample size, so the well-represented male group has larger utility loss.

This gives us an insight on the relation between differential privacy and the inequality in utility loss w.r.t. each group. The direct cause of the inequality is the large cost of privacy due to large average gradient norm (which can be caused by small group sample size along with other factors). In DPSGD, the clipping bound is selected uniformly for each group without consideration of the difference in clipping biases. As a result, the noise addition to achieve (ϵ, δ)-differential privacy on the learning model results in different utility-privacy trade-off for each group, where the underrepresented or the more complex group incurs a larger utility loss. After all, DPSGD is designed to protect individual's privacy with nice properties without consideration of its different impact towards each group. In order to avoid disparate utility loss among groups, we need to modify DPSGD such that each group needs to achieve different level of privacy to counter their difference in costs of privacy.

5 REMOVING DISPARATE IMPACT ON MODEL ACCURACY IN DPSGD

Our objective is to build a learning algorithm that outputs a classifier $\tilde{\eta}(a;\tilde{w})$ with parameter $\tilde{w}$ that achieves differential privacy, equality of utility loss w.r.t each group, and good accuracy. Based on our preliminary observation and analysis on cost of privacy, we propose a heuristic removal algorithm to achieve equal utility loss w.r.t. each group, called DPSGD-F.

Algorithm 2 DPSGD-F (Dataset D, loss function $\mathcal{L}_D(w)$, learning rate r, batch size b, noise scales σ_1, σ_2, base clipping bound C_0)

for $t \in [T]$ **do**
  Randomly sample a batch B_t with $|B_t| = b$ from D
  for each sample $x_i \in B_t$ **do**
    $g_i = \nabla L_i(w_t)$
  for each group $k \in [K]$ **do**
    $m^k = \left|\left\{i : |g_i^k| > C_0\right\}\right|$
    $o^k = \left|\left\{i : |g_i^k| \le C_0\right\}\right|$
    $\left\{\tilde{m}^k, \tilde{o}^k\right\}_{k\in[K]} = \left\{m^k, o^k\right\}_{k\in[K]} + N(0, \sigma_1^2\mathbf{I})$
  $\tilde{m} = \sum_{k\in[K]} \tilde{m}^k$
  for each group $k \in [K]$ **do**
    $\tilde{b}^k = \tilde{m}^k + \tilde{o}^k$
    $C^k = C_0 \times \left(1 + \frac{\tilde{m}^k/\tilde{b}^k}{\tilde{m}/b}\right)$
  for each sample $x_i \in B_t$ **do**
    $\bar{g}_i = g_i \times \min\left(1, \frac{C^k}{|g_i|}\right)$
  $C = \max_k C^k$
  $\tilde{G}_B = \frac{1}{b}\left(\sum_i \bar{g}_i + N(0, \sigma_2^2 C^2 \mathbf{I})\right)$
  $\tilde{w}_{t+1} = \tilde{w}_t - r\tilde{G}_B$
Return $\tilde{w}_T$ and accumulated (ϵ, δ)

5.1 Equal Costs of Differential Privacy

In the within-model fairness, equal odds results in the equality of accuracy for different groups. Note that equal accuracy does not result in equal odds. As a trade-off for privacy, differential privacy results in accuracy loss on the model. However, different groups may incur different levels of accuracy loss. We use reduction in accuracy w.r.t. group k to measure utility loss between the private model $\tilde{\eta}$ and the non-privacte model η, denoted by Δ^k. We define a new fairness notion called *equal costs of differential privacy*, which requires that the utility loss due to differential privacy is the same for all groups.

Equal costs of differential privacy Given a labeled dataset D, a classifier η and a differentially private classifier $\tilde{\eta}$, a differentially private mechanism satisfies equal equal costs of privacy if $\Delta^i(\tilde{\eta} - \eta) = \Delta^j(\tilde{\eta} - \eta)$, where i, j are any two values of S.

5.2 Removal Algorithm

We propose a heuristic approach for differentially private SGD that removes disparate impact across different groups. The intuition of our heuristic approach is to balance the level of privacy w.r.t. each group based on their utility-privacy trade-off. Algorithm 2 shows the framework of our approach. Instead of uniformly clipping the gradients for all groups, we propose to do adaptive sensitive clipping where each group k gets its own clipping bound C^k. For the group with larger clipping bias (due to large gradients), we choose a larger clipping bound to balance their higher cost of privacy. The large gradients may be due to group sample size or other factors.

Based on our observation and analysis in the previous section, to balance the difference in costs of privacy for each group, we need to adjust the clipping bound C^k such that the contribution of each

group is proportional to the size of their average gradient (Line 12 in Algorithm 2). Ideally, we would like to adjust the clipping bound based on the private estimate of the average gradient norm. However, the original gradient before clipping has unbounded sensitivity. It would not be practical to get its private estimate. We need to construct a good approximate estimate of the relative size of the average gradient w.r.t. each group and it needs to have a small sensitivity for private estimation.

In our algorithm, we choose adaptive clipping bound C^k based on the m^k, where $m^k = |\{i : |g_i^k| > C_0\}|$. To avoid the influence of group sample size, we use the fraction of $\frac{m^k}{b^k}$ that represents the fraction of samples in the group with gradients larger than C_0. The relative ratio of $\frac{m^k}{b^k}$ and $\frac{m}{b}$ can approximately represent the relative size of the average gradient (Line 12). To choose the clipping bound C^k for group k in a private way, we get the private $\tilde{m}^k, \tilde{b}^k$ and $\tilde{m}$ from the collection $\{m^k, o^k\}_{k\in[K]}$ (Line 5-11). The collection $\{m^k, o^k\}_{k\in[K]}$ has sensitivity of 1, which is much smaller than the sensitivity of the actual gradients when we estimate the relative size of the average gradient.

After the adaptive clipping, the sensitivity of the clipped gradient of group k is $C^k = C_0 \times (1 + \frac{\tilde{m}^k/\tilde{b}^k}{\tilde{m}/b})$. The sensitivity of the clipped gradient of the total population would be $\max_k C^k$ as the worst case in the total population needs to be considered.

Note that in Algorithm 2 we have two steps of adding noise in each iteration t. We first use a relatively large noise scale σ_1 (small privacy budget) to get a private collection $\{\tilde{m}^k, \tilde{o}^k\}_{k\in[K]}$ (Line 8). Then we use a relatively small noise scale σ_2 to perturb the gradients (Line 16). The composition theorem (Theorem 2 in [1]) is applied when we compute the accumulated (ϵ, δ) from moments accountant (Line 18). Because $\sigma_1 > \sigma_2$, only a small fraction of privacy budget is spent on getting C^k.

For the total population, Algorithm 2 still satisfies (ϵ, δ)-differential privacy as it accounts for the worst clipping bound $\max_k C^k$. On the group level, each group achieves different levels of privacy depending on their utility-privacy trade-off. Algorithm 2 achieves $(\frac{C^k}{\max_{k'} C^{k'}}\epsilon, \delta)$-differential privacy w.r.t. group k.

With our modified DPSGD algorithm, we continue our discussion in Secion 4.2. On MNIST, the difference in gradient norms is primarily decided by the group sample size. Consider a majority group s^+ and a minority group s^-. In Algorithm 1, each group achieves the same level of privacy, but the underrepresented group s^- has higher cost of privacy. In Algorithm 2, we choose a higher clipping bound C^- for the underrepresented group. Because the noise scale is $\frac{C}{\epsilon} = \frac{C^-}{\epsilon}$ and the sensitivity of clipped gradients for the underrepresented group is C^-, the noisy gradient w.r.t. the underrepresented group achieves (ϵ, δ)-differential privacy. The well-represented group s^+ has a smaller cost of privacy, so we choose a lower clipping bound C^+. Because the noise scale is $\frac{C}{\epsilon} = \frac{C^-}{\epsilon}$ and the sensitivity of clipped gradients for the well-represented group is C^+, the noisy gradient w.r.t. the well-represented group then achieves $(\frac{C^+}{C^-}\epsilon, \delta)$-differential privacy. Two groups have different clipping bounds C^+, C^- and the same noise addition based on $C = \max(C^+, C^-)$ (same ϵ but different relative scales w.r.t. their group sensitivities). Hence, when we enforce the same level of utility loss for groups with different sample sizes, the well-represented group achieves stronger privacy (smaller than ϵ) than the underrepresented group. On Adult/Dutch, the male group has larger gradients regardless of the sample size. The group with smaller gradients based on model and data distribution has smaller cost of privacy. Algorithm 2 adjusts the clipping bound for each group. The group with smaller gradients achieves stronger level of privacy.

Algorithm 3 Naïve (Dataset D, loss function $\mathcal{L}_D(w)$, learning rate r, batch size b, noise scales σ_1, σ_2, base clipping bound C_0)

1: **for** $t \in [T]$ **do**
2: Randomly sample a batch B_t with $|B_t| = b$ from D
3: **for** each sample $x_i \in B_t$ **do**
4: $g_i = \nabla L_i(w_t)$
5: $\left\{\tilde{b}^k\right\}_{k\in[K]} = \left\{b^k\right\}_{k\in[K]} + N(0, \sigma_1^2\mathbf{I})$
6: **for** each group $k \in [K]$ **do**
7: $\theta^k = 1 \times \frac{b/K}{\tilde{b}^k}$
8: **for** each sample $x_i \in B_t$ **do**
9: $\bar{g}_i = \theta^k \times g_i \times \min\left(1, \frac{C_0}{|g_i|}\right)$
10: $C = C_0 \times \max_k \theta^k$
11: $\tilde{G}_B = \frac{1}{b}\left(\sum_i \bar{g}_i + N(0, \sigma_2^2 C^2 \mathbf{I})\right)$
12: $\tilde{w}_{t+1} = \tilde{w}_t - r\tilde{G}_B$
13: Return $\tilde{w}_T$ and accumulated (ϵ, δ)

5.3 Baseline

There is no previous work on how to achieve equal utility loss in DPSGD. For experimental evaluation, we present a naïve baseline algorithm based on reweighting (shown as Algorithm 3) in this section, as reweighting is a common way to mitigating biases. The naïve algorithm considers group sample size as the main cause of disproportional impact in DPSGD and adjusts sample contribution of each group to mitigate the impact of sample size. For the group with larger group sample size, we reweight the sample contribution with $\theta^k \propto \frac{1}{\tilde{b}^k}$ instead of using uniform weight of 1 for all groups, where $\tilde{b}^k$ is privately estimated (Line 5 in Algorithm 3). Note that G_B in Algorithm 1 is estimated based on uniform weight of each sample regardless of their group membership. The sensitivity for group k is $C^k = C_0 \times \theta^k$. The sensitivity for the total population would be $C^0 \times \max_k \theta^k$. As we add noise based on the sensitivity for the total population, Algorithm 3 satisfies (ϵ, δ)-differential privacy for the total population and $(\frac{\theta^k}{\max_{k'} \theta^{k'}}\epsilon)$-differential privacy w.r.t. group k. The result also matches the idea that we limit the sample contribution of the group with smaller cost of privacy to achieve stronger privacy level w.r.t. the group. However, Algorithm 3 only considers the group sample size. As we know from previous analysis, the factors that affect the gradient norm and bias due to clipping are more complex than just the group sample size. We will compare with this Naïve approach in experiments.

Table 1: Model accuracy w.r.t. the total population, the majority group and the minority group on the MNIST ($\epsilon = 6.55, \delta = 10^{-6}$), Adult ($\epsilon = 3.1, \delta = 10^{-6}$) and Dutch ($\epsilon = 2.66, \delta = 10^{-6}$) datasets

Dataset	MNIST			Adult			Dutch		
Group	Total	Class 2	Class 8	Total	M	F	Total	M	F
Sample size	54649	5958	500	45222	30527	14695	60420	30273	30147
SGD	0.9855	0.9903	0.9292	0.8099	0.7610	0.9117	0.7879	0.8013	0.7744
DPSGD vs. SGD	-0.1081	-0.0707	-0.6807	-0.0592	-0.0740	-0.0281	-0.1001	-0.1534	-0.0466
Naïve vs. SGD	-0.1500	-0.1512	-0.1510	-0.0593	-0.0742	-0.0281	-0.1004	-0.1549	-0.0458
DPSGD-F vs. SGD	-0.0293	-0.0281	-0.0432	-0.0254	-0.0298	-0.0161	-0.0130	-0.0160	-0.0099

Table 2: The average loss and the average gradient norm w.r.t. groups at the last training epoch on the MNIST ($\epsilon = 6.55, \delta = 10^{-6}$), Adult ($\epsilon = 3.1, \delta = 10^{-6}$) and Dutch ($\epsilon = 2.66, \delta = 10^{-6}$) datasets

	Average loss						Average gradient norm					
Dataset	MNIST		Adult		Dutch		MNIST		Adult		Dutch	
Group	Class 2	Class 8	M	F	M	F	Class 2	Class 8	M	F	M	F
SGD	0.04	0.04	0.48	0.27	0.52	0.53	0.68	4.76	0.08	0.11	0.19	0.19
DPSGD	0.41	2.16	0.68	0.31	0.59	0.53	13.53	100.46	0.41	0.12	0.26	0.12
Naïve	3.08	1.89	0.71	0.32	0.59	0.53	0.83	0.76	0.43	0.13	0.26	0.12
DPSGD-F	0.20	0.42	0.50	0.27	0.51	0.52	1.45	2.53	0.12	0.08	0.19	0.18

6 EXPERIMENTS

6.1 Experiment Setup

6.1.1 Datasets. We use MNIST dataset and replicate the setting in [3]. We constructed an unbalanced MNIST dataset with each class having about 6,000 samples expect for class 8. Class 8 has the most false negatives, hence we choose it as the artificially underrepresented group (reducing the number of training samples to 500). We compare the underrepresented class 8 with the well-represented class 2 that shares fewest false negatives with the class 8 and therefore can be considered independent. The testing dataset has 10,000 testing samples with about 1,000 for each class. We also use two census datasets, Adult and Dutch. For both datasets, we consider "Sex" as the protected attribute and "Income" as decision. For unprotected attributes, we convert categorical attributes to one-hot vectors and normalize numerical attributes to $[0, 1]$ range. After preprocessing, we have 40 unprotected attributes for Adult and 35 unprotected attributes for Dutch. In all settings, we split the census datasets into 80% training data and 20% testing data.

6.1.2 Model. For the MNIST dataset, we use a neural network with 2 convolutional layers and 2 linear layers with 431K parameters in total. We use learning rate $r = 0.01$, batch size $b = 256$, and the number of training epochs is 60. For the census datasets, we use a logistic regression model with regularization parameter 0.01. We use learning rate $r = 1/\sqrt{T}$, batch size $b = 256$, and the number of training epochs is 20. Our **source code** can be downloaded at https://tinyurl.com/dpsgdf.

6.1.3 Baselines. We compare our proposed method DPSGD-F (Algorithm 2) with the original DPSGD (Algorithm 1) and the Naïve approach (Algorithm 3). For each setting, the learning parameters are the same. We set C_0, σ_2 in DPSGD-F and Naïve equal to C, σ in DPSGD, respectively. We set $\sigma_1 = 10\sigma_2$. For the MNIST dataset, we set noise scale $\sigma = 0.8$, clipping bound $C = 1$, and $\delta = 10^{-6}$. For the census datasets, we set $\sigma = 1$, $C = 0.5$ and $\delta = 10^{-6}$. The accumulated privacy budget ϵ for each setting is computed using the privacy moments accounting method [1]. Because we set $\sigma_1 = 10\sigma_2$, most of ϵ is spent on gradients from σ_2. Only about 0.01 budget is from σ_1. To compare DPSGD-F and Naïve with DPSGD under the same privacy budget, the algorithm runs a few less iterations than DPSGD in the last epoch, where the total number of iterations $T = \text{epochs} \times n/b$ in SGD and DPSGD. For DPSGD-F and Naïve, T is 19, 11 and 17 less on the MNIST, Adult and Dutch datasets, respectively, which are very small differences in proportion to T. All DP models are compared with the non-private SGD when we measure the utility loss due to differential privacy.

6.1.4 Metric. We use the test data to measure the model utility and fairness. We use reduction in model accuracy for each group between the private SGD and the non-private SGD (Δ^k) as the metric to measure the cost of differential privacy w.r.t. each group. The difference between the costs on groups ($|\Delta^i - \Delta^j|$ for any i, j) measures the level of inequality in utility loss due to differential privacy. If the costs for all groups are independent of the protected attribute ($|\Delta^i - \Delta^j| \leq \tau$, $\tau = 0.05$ used in the paper), we consider the private SGD has equal reduction in model accuracy w.r.t. each group, i.e. the private SGD achieves equal cost of differential privacy. We also report the average loss and average gradient norm to show the convergence w.r.t. each group during training.

6.2 MNIST Dataset

Table 1 shows the model accuracy w.r.t. class 2 and 8 on the MNIST datasets. The non-private SGD model converges to 0.9292 accuracy on class 8 vs. 0.9903 accuracy on class 2. The DPSGD model causes −0.6807 accuracy loss on class 8 vs. −0.0707 on class 2, which exhibits a significant disparate impact on the underrepresented class. The Naïve approach achieves −0.1510 accuracy loss on class 8 vs. −0.1512 on class 2, which achieves equal costs of privacy. Our DPSGD-F algorithm has −0.0432 accuracy loss on class 8 vs.

DPSGD对于 underrepresented class 表现不同的影响

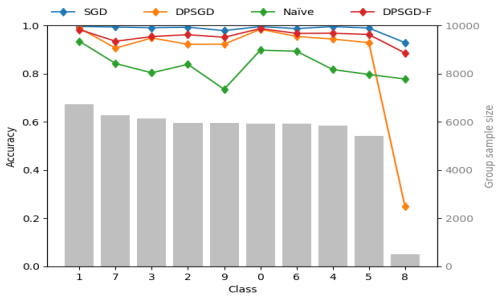


Figure 1: Model accuracy w.r.t. each class on MNIST

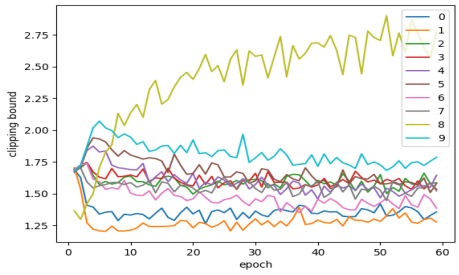


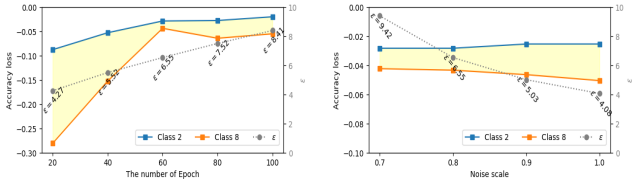
Figure 2: The clipping bound C^k w.r.t. each class over training epochs for DPSGD-F on the MNIST dataset

Table 3: Model accuracy for different uniform clipping bound ($C = 1, 2, 3, 4, 5$) in DPSGD vs. adaptive clipping bound ($C_0 = 1$) in DPSGD-F on the MNIST dataset

Group	Total	Class 2	Class 8
Sample size	54649	5958	500
SGD	0.9855	0.9903	0.9292
DPSGD ($C = 1$) vs. SGD	-0.1081	-0.0707	-0.6807
DPSGD ($C = 2$) vs. SGD	-0.0587	-0.0426	-0.3286
DPSGD ($C = 3$) vs. SGD	-0.0390	-0.0232	-0.2013
DPSGD ($C = 4$) vs. SGD	-0.0286	-0.0194	-0.1376
DPSGD ($C = 5$) vs. SGD	-0.0240	-0.0145	-0.1099
DPSGD-F ($C_0 = 1$) vs. SGD	-0.0293	-0.0281	-0.0432

-0.0281 on class 2, which also achieves equal costs of privacy. The total model accuracy also drops much less for DPSGD-F (-0.0293) than the original DPSGD (-0.1081). We further show in Figure 1 the model accuracy w.r.t. all classes on the MNIST dataset. In summary, the difference between DPSGD and DPSGD-F is small and consistent across all classes.

Table 2 shows the average loss and average gradient norm w.r.t. class 2 and 8 for SGD and different DP models at the last training epoch. In DPSGD, the average gradient norm for class 8 is over 100 and the average loss for class 8 is 2.16. Whereas, in DPSGD-F, the average gradient norm for class 8 is only 2.53 and the average loss for class 8 is only 0.42. The convergence loss and the gradient norm for class 8 are much closer to the ones for class 2 in DPSGD-F. The detailed convergence trends are included in the technical report [28]. The trend in DPSGD-F is the closest to the trend in SGD among all DP models. It shows our adjusted clipping bound helps to achieve the same group utility loss regardless of the group sample size.



(a) Varying the number of epochs **(b) Varying noise scale**

Figure 3: The accuracy loss on class 2 and 8 (DPSGD-F vs. SGD) with different ϵ on the MNIST dataset

Figure 2 shows how our adaptive clipping bound changes over epochs in DPSGD-F. Because class 8 has larger clipping bias due to its underrepresented group sample size, DPSGD-F gives class 8 a higher clipping bound to increase its sample contribution in the total gradient. The maximal C^k is close to 3. To show that the fair performance of DPSGD-F is not caused by increasing clipping bound uniformly, we run the original DPSGD with increasing clipping bound from $C = 1$ to $C = 5$. Table 3 shows the level of inequality in utility loss for different clipping bound in DPSGD vs. the adaptive clipping bound in DPSGD-F. Even though increasing clipping bound C in DPSGD can improve the accuracy on class 8, there is still significant difference between the accuracy loss on class 8 (-0.1099 when $C = 5$) and the accuracy loss on class 2 (-0.0145 when $C = 5$). This is because the utility-privacy trade-offs are different for the minority group and the majority group under the same clipping bound. So the inequality in utility loss cannot be removed by simply increasing the clipping bound in DPSGD. On the contrary, DPSGD-F achieves equal costs of privacy by adjusting the clipping bound for each group according to the utility-privacy trade-off. The group with smaller cost of privacy achieves a stronger level of privacy as a result of adaptive clipping.

We also evaluate the effectiveness of DPSGD-F over different ϵ (we keep δ the same). There are two factors, the number of epochs and the noise scale, affecting the accumulated ϵ. Figure 3a shows the group accuracy loss over different accumulated ϵ by altering the number of epochs while setting other parameters the same as default. With the number of epochs increasing, the accumulated ϵ increases (shown as dashed line in Figure 3a), and the difference between the accuracy losses of class 2 and 8 decreases. From 60 epochs on, the difference is below the threshold τ, i.e. DPSGD-F achieves equal costs of privacy. Figure 3b shows the group accuracy loss over different accumulated ϵ by altering the noise scale while setting others parameters the same as default. With the noise scale increasing, the accumulated ϵ decreases, and the difference between the accuracy losses of class 2 and 8 slightly increases, yet the difference is consistently below the threshold τ, i.e. DPSGD-F achieves equal costs of privacy. It suggests that it is better to enforce stronger privacy by increasing the noise scale than prematurely terminate training.

6.3 Adult and Dutch Datasets

Table 1 shows the model accuracy w.r.t. male and female on the Adult and Dutch datasets. The clipping biases for both census datasets are not primarily decided by group sample size. We observe disparate impact on DPSGD in comparison to SGD against the male group, even though the male group is not underrepresented. The Naïve approach does not work at all to achieve equal costs of privacy in this case, as the group sample size is not as important as in the MNIST dataset. There are still other factors that affect the gradient norm and the clipping bias w.r.t. each group. DPSGD-F can achieve similar accuracy loss for male and female groups.

Table 2 shows the average loss and average gradient norm w.r.t. male and female for SGD and different DP models at the last training epoch. On the Adult dataset, the average gradient norm in DPSGD for male is 5 times of the one in SGD and the average loss in DPSGD for male is 50% more than the one in SGD. Whereas, in DPSGD-F, the average gradient norm and the average loss for the male group are much closer to the ones in SGD. Similar to the Adult dataset, on the Dutch dataset, the average gradient norm and the average loss in DPSGD-F for the male group are much closer to the ones in SGD. On both datasets, the trends in DPSGD-F are the closest to the trends in SGD among all DP models. It shows that our adjusted clipping bound helps to achieve the same group utility loss.

7 CONCLUSION AND FUTURE WORK

Gradient clipping and random noise addition, which are the core techniques in differentially private SGD, disproportionately affect underrepresented and complex classes and subgroups. As a consequence, DPSGD has disparate impact: the accuracy of a model trained using DPSGD tends to decrease more on these classes and subgroups vs. the original, non-private model. If the original model is unfair in the sense that its accuracy is not the same across all subgroups, DPSGD exacerbates this unfairness. In this work, we have proposed DPSGD-F to achieve differential privacy, equal costs of differential privacy, and good utility. DPSGD-F adjusts the contribution of samples in a group depending on the group clipping bias such that differential privacy has no disparate impact on group accuracy. Our experimental evaluation shows how group sample size and group clipping bias affect the cost of differential privacy in DPSGD, and how adaptive clipping for each group helps to mitigate the disparate impact caused by differential privacy in DPSGD-F. Gradient clipping in the non-private context may improve the model robustness against outliers. However, examples in the minority group are not outliers. They should not be ignored by the (private) learning model. In future work, we plan to further improve our adaptive clipping method from group-wise adaptive clipping to element-wise (from perspectives of users or parameters) adaptive clipping, so the model can be fair even to the unseen minority class.

ACKNOWLEDGMENTS

This work was supported in part by NSF 1502273, 1920920, 1937010, and 1946391.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *SIGSAC*. 308–318.
[2] Kareem Amin, Alex Kulesza, Andres Muñoz Medina, and Sergei Vassilvitskii. 2019. Bounding User Contributions: A Bias-Variance Trade-off in Differential Privacy. In *ICML*. 263–271.
[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In *NeurIPS*. 15453–15462.
[4] Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. 2018. Model-Agnostic Private Learning. In *NeurIPS*. 7102–7112.
[5] T. Calders, F. Kamiran, and M. Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *ICDM Workshops*. 13–18.
[6] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. 2011. Differentially Private Empirical Risk Minimization. *J. Mach. Learn. Res.* 12 (2011), 1069–1109.
[7] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the Compatibility of Privacy and Fairness. In *UMAP*. 309–315.
[8] Min Du, Ruoxi Jia, and Dawn Song. 2020. Robust anomaly detection and backdoor attack detection via differential privacy. In *ICLR*.
[9] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2013. Local Privacy and Statistical Minimax Rates. In *FOCS*. 429–438.
[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science*. 214–226.
[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third*. 265–284.
[12] Harrison Edwards and Amos J. Storkey. 2016. Censoring Representations with an Adversary. In *ICLR*.
[13] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In *FAT**. 35–47.
[14] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *KDD*. 259–268.
[15] Sara Hajian, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2015. Discrimination- and privacy-aware patterns. *Data Min. Knowl. Discov.* 29, 6 (2015), 1733–1782.
[16] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*. 3315–3323.
[17] Matthew Jagielski, Michael J. Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. 2019. Differentially Private Fair Learning. In *ICML*. 3000–3008.
[18] Mimansa Jaiswal and Emily Mower Provost. 2020. Privacy Enhanced Multimodal Neural Representations for Emotion Recognition. In *AAAI*. 7985–7993.
[19] F. Kamiran and T. Calders. 2009. Classifying without discriminating. In *ICCCC*. 1–6.
[20] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33, 1 (2011), 1–33.
[21] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *ICML*. 2569–2577.
[22] Emmanouil Krasanakis, Eleftherios Spyromitros Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *WWW*. 853–862.
[23] Jaewoo Lee and Daniel Kifer. 2018. Concentrated Differentially Private Gradient Descent with Adaptive per-Iteration Privacy Budget. In *KDD*. 1656–1665.
[24] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*. 3381–3390.
[25] NhatHai Phan, Xintao Wu, Han Hu, and Dejing Dou. 2017. Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning. In *ICDM*. 385–394.
[26] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. 2019. AdaCliP: Adaptive Clipping for Private SGD. *CoRR* abs/1908.07643 (2019).
[27] Om Thakkar, Galen Andrew, and H. Brendan McMahan. 2019. Differentially Private Learning with Adaptive Clipping. *CoRR* abs/1905.03871 (2019).
[28] Depeng Xu, Wei Du, and Xintao Wu. 2020. Removing Disparate Impact of Differentially Private Stochastic Gradient Descent on Model Accuracy. *CoRR* abs/2003.03699 (2020).
[29] Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving Differential Privacy and Fairness in Logistic Regression. In *Companion of WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. 594–599.
[30] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*. 962–970.
[31] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AIES*. 335–340.
[32] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *IJCAI*. 3929–3935.