

# Disparate Impact in Differential Privacy from Gradient Misalignment

**Maria S. Esipova**  
Layer 6 AI  
maria@layer6.ai

**Atiyeh Ashari Ghomi**  
Layer 6 AI  
atiyeh@layer6.ai

**Yaqiao Luo**  
Layer 6 AI  
emily@layer6.ai

**Jesse C. Cresswell**  
Layer 6 AI  
jesse@layer6.ai

## Abstract

As machine learning becomes more widespread throughout society, aspects including data privacy and fairness must be carefully considered, and are crucial for deployment in highly regulated industries. Unfortunately, the application of privacy enhancing technologies can worsen unfair tendencies in models. In particular, one of the most widely used techniques for private model training, differentially private stochastic gradient descent (DPSGD), frequently intensifies disparate impact on groups within data. In this work we study the fine-grained causes of unfairness in DPSGD and identify gradient misalignment due to inequitable gradient clipping as the most significant source. This observation leads us to a new method for reducing unfairness by preventing gradient misalignment in DPSGD.

## 1 Introduction

The increasingly widespread use of machine learning throughout society has brought into focus social, ethical, and legal considerations surrounding its use. In highly regulated industries, such as healthcare and banking, regional laws and regulations require data collection and analysis to respect the privacy of individuals.<sup>1</sup> Other regulations focus on the fairness of how models are developed and used.<sup>2</sup> As machine learning is progressively adopted in highly regulated industries, the privacy and fairness aspects of models must be considered at all stages of the modelling lifecycle.

There are many privacy enhancing technologies including differential privacy [16], federated learning [26], secure multiparty computation [38], and homomorphic encryption [19] that are used separately or jointly to protect the privacy of individuals whose data is used for machine learning [12, 3, 24]. The latter three technologies find usage in data sharing schemes and can allow data to be analysed while preventing its exposure to the wrong parties. However, in these cases the procedures usually return a trained model which itself can leak private information [8]. On the other hand, differential privacy (DP) focuses on quantifying the privacy cost of disclosing aggregated information about a dataset, and can guarantee that nothing is learned about individuals that could not be inferred from population-level correlations [21]. Hence, DP is often used when the results of data analysis will be made publicly available, for instance when exposing the outputs of a model, or the results of the most recent US census [2].

Not only must data privacy be protected for applications in regulated industries, fairness of models must be ensured. While there is no single definition that captures what it means to be “fair”, with regards to model-based decision making, fairness may preclude disparate treatment or disparate impact [27]. Disparate treatment is usually concerned with how models are applied across populations,

<sup>1</sup>Examples of laws governing data privacy include the General Data Protection Regulation in Europe, Health Insurance Portability and Accountability Act in the USA, and Personal Information Protection and Electronic Documents Act in Canada.

<sup>2</sup>In the USA, fair lending laws including the Fair Housing Act, and Equal Credit Opportunity Act prohibit discrimination based on protected characteristics such as race, age, and sex.

whereas disparate impact can arise from biases in datasets that are amplified by the greedy nature of loss minimization algorithms [7]. Differences in model performance across protected groups can result in a significant negative monetary, health, or societal impact for individuals who are discriminated against [13].

Unfortunately, it has been observed that disparate impact can be exacerbated by applying DP in machine learning [5]. Applications of DP always come with a privacy-utility tradeoff, where stronger guarantees of privacy negatively impact the usefulness of results - model performance in this context [15]. Underrepresented groups within the population can experience disparity in the cost of adding privacy, hence, fairness concerns are a major obstacle to deploying models trained with DP.

The causes of unfairness in DP depend on the techniques used, but are not fully understood. For the most widely used technique, differentially private stochastic gradient descent (DPSGD), two sources of error are introduced that impact model utility. First, per-sample gradients are clipped to a fixed upper bound on their norm, and second, noise is added to the averaged gradient. Disparate impact from DPSGD was initially hypothesized to be rooted in unbalanced datasets [5], though counterexamples were found by [37]. More recent research claims disparate impact to be caused by incommensurate clipping errors across groups, in turn effected by a large difference in average group gradient norms [37, 34].

In this work we highlight the disparate impact of gradient misalignment. In particular, we claim that the most significant cause of disparate impact is the difference in the direction of the unclipped and clipped gradients, which in turn can be caused by aggressive clipping and imbalances of gradient norms between groups. Our analysis of direction errors leads to a variant of DPSGD with properly aligned gradients. We explore this alternate clipping method in relation to disparate impact and show that it not only significantly reduces the cost of privacy across all protected groups, it also reduces the *difference* in cost of privacy for all groups. Hence, it removes disparate impact and is more effective than previous proposals in doing so.

In summary our contributions are:

- A more fine-grained analysis of the causes of disparate impact in DPSGD identifying the role of gradient misalignment.
- Evidence that gradient misalignment is the most significant cause of disparate impact.
- A new algorithm, DPSGD-Global-Adapt, which properly aligns gradients and reduces disparate impact without requiring access to protected group labels.
- Experimental verification that aligning gradients is more successful at mitigating disparate impact than previous approaches.

## 2 Related Work

**Privacy and Fairness:** While privacy and fairness have been extensively studied separately, only recently have their interactions come into focus. Ekstrand et al. [17] considered the intersection of privacy and fairness for several definitions of privacy. This line of research gained new urgency when Bagdasaryan et al. [5] empirically observed that DPSGD exacerbated existing disparity in model accuracy when datasets were imbalanced due to underrepresented groups. Disparate impact due to DP was further observed in [33] and [18] for varying levels of group imbalance. Using an adversarial definition of privacy rather than DP, Jaiswal and Mower Provost [23] found that it is not necessarily underrepresented groups that incur higher privacy costs. Similar examples were shown in [37] for DPSGD, and disparate impact was linked to groups having larger gradient norms.

Other fairness-aware learning research has evaluated the fairness of a private model’s outcomes on protected groups of the population. In this context fairness might refer to a statistical condition of non-discrimination with respect to groups [30, 35], for example, equalized odds [21], equality of opportunity [14], or demographic parity [36, 18]. Chang and Shokri [9] empirically found that imposing fairness constraints on private models could lead to higher privacy loss for certain groups. In contrast, we consider cross-model fairness where the *cost of adding privacy* to a non-private model must be fairly distributed between groups.

**Adaptive Clipping:** Many variations on the clipping procedure in DPSGD have been proposed to improve properties other than fairness, but we find they can also benefit fairness. Adaptive clipping

comes in many forms, but usually tunes the clipping threshold as training proceeds which can allow for better privacy-utility tradeoffs [4, 32] and convergence [6]. The convergence of DPSGD connects to the symmetry properties of the distribution of gradients [11] which are affected by clipping.

### 3 Background

#### 3.1 Setting and Definitions

We begin by laying out the problem setting and review the relevant definitions for discussing fairness in privacy. For concreteness we consider a binary classification problem on a dataset  $D$  which consists of  $n$  points of the form  $(x_i, a_i, y_i)$ , where  $x_i \in \mathbb{R}^d$  is a feature vector,  $y_i \in \{0, 1\}$  is a binary label, and  $a_i \in [K]$  refers to a protected group attribute which partitions the data. The group label  $a_i$  can optionally be an attribute in  $x_i$ , the label value  $y_i$ , or some distinct auxiliary value.

The goal is to train a model  $f_\theta : \mathbb{R}^d \rightarrow [0, 1]$  with parameter vector  $\theta$  that is simultaneously useful and private, and in which the application of privacy is fair. Utility in the empirical risk minimization (ERM) problem is governed by the per-sample loss function  $\ell : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$ , with the optimal model minimizing the empirical risk objective  $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i \in D} \ell(f_\theta(x_i), y_i)$ , which happens for an optimal set of parameters  $\theta^* = \arg \min_\theta \mathcal{L}(\theta; D)$ . The requirement of privacy is applied to the model through its parameters; private parameters  $\tilde{\theta}$  must be obtained such that a minimal amount of private information in  $D$  is exposed. For this we apply the framework of differential privacy, recounted in the next section.

Fairness of the privacy methodology can be measured in terms of the disparate impact that applying privacy has on the protected groups. As in [5], we use a version of *accuracy parity*, the difference in classification accuracy across protected groups after adding privacy. For any subset of the data  $Z \subseteq D$ , denote  $Z_k = \{(x_i, a_i, y_i) \in Z \mid a_i = k\}$  as the set of points belonging to group  $k$ . A private model has accuracy parity for subset  $Z_k$  if it minimizes the *privacy cost*

$$\pi(\theta, Z_k) = \text{acc}(\theta^*; Z_k) - \mathbb{E}_{\tilde{\theta}}[\text{acc}(\tilde{\theta}; Z_k)], \quad (1)$$

where the expectation is over the randomness involved in acquiring private model parameters. Of course, metrics other than classification accuracy could be used as required by the problem setting. Alternatively, fairness for privacy can be measured at the level of the loss function as in [34], which is more amenable to analyzing the causes of unfairness. The *excessive risk* experienced by a group is

$$R(\theta, Z_k) = \mathbb{E}_{\tilde{\theta}}[\mathcal{L}(\tilde{\theta}; Z_k)] - \mathcal{L}(\theta^*; Z_k). \quad (2)$$

For convenience when the whole dataset is used we denote  $R(\theta; D_k)$  as  $R_k$ , and  $R(\theta; D)$  as  $R$  (and similar for privacy cost). For both accuracy and loss we consider the gap between disparate impact values across groups. The *privacy cost gap* is  $\pi_{a,b} = |\pi_a - \pi_b|$  for groups  $a, b \in [K]$ , and the *excessive risk gap* refers to  $\xi_{a,b} = |R_a - R_b|$  which is a slight modification from [34]. The goal of a fair private classifier is to minimize the privacy cost and/or excessive risk for all values of the protected group attribute, while maintaining small fairness gaps.

#### 3.2 Differential privacy

Differential privacy (DP) [16] is a widely used framework for quantifying how much privacy is consumed by a data analysis procedure. Formally, let  $D$  represent a set of data points, and  $M$  a probabilistic function, or *mechanism*, acting on datasets. We say that the mechanism is  $(\epsilon, \delta)$ -*differentially private* if for all subsets of possible outputs  $S \subseteq \text{Range}(M)$ , and for all pairs of databases  $D$  and  $D'$  that differ by the addition or removal of one element,

$$\Pr[M(D) \in S] \leq \exp(\epsilon) \Pr[M(D') \in S] + \delta. \quad (3)$$

---

**Algorithm 1** DPSGD

---

**Require:** Iterations  $T$ , Dataset  $D$ , sampling rate  $q$ , clipping bound  $C_0$ , noise multiplier  $\sigma$ , learning

rates  $\eta_t$   
Initialize  $\theta_0$  randomly  
**for**  $t$  in  $0, \dots, T - 1$  **do**  
     $B \leftarrow$  Poisson sample of  $D$  with rate  $q$   
    **for**  $(x_i, y_i)$  in  $B$  **do**  
         $g_i \leftarrow \nabla_{\theta} \ell(f_{\theta_t}(x_i), y_i)$   $\triangleright$  Compute per-sample gradients  
         $\bar{g}_i \leftarrow g_i \cdot \min\left(1, \frac{C_0}{\|g_i\|}\right)$   $\triangleright$  Clip if gradient norm is greater than  $C_0$   
     $\bar{g}_B \leftarrow \frac{1}{|B|} \left(\sum_{i \in B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C_0^2 \mathbb{I})\right)$   $\triangleright$  Aggregate and add noise  
     $\theta_{t+1} \leftarrow \theta_t - \eta_t \bar{g}_B$

---

For the ERM problem, there are several ways to train a differentially private model [10]. In this work we consider models that can be trained with stochastic gradient descent (SGD), such as neural networks, and focus on the most successful approach, DPSGD [1], in which the Gaussian mechanism [15] is applied to gradient updates as in Alg. 1. Since per-sample gradients  $g_i$  generally do not have finite sensitivity, defined as  $\Delta_h = \max_{D, D'} \|h(D) - h(D')\|$  for a function  $h$ , they are first clipped to have norm upper bounded by a fixed hyperparameter  $C_0$ . Clipped gradients  $\bar{g}_i$  in a batch are aggregated into  $\bar{g}_B$  and noise is added to produce  $\tilde{g}_B$  used in the parameter update.

### 3.3 Fairness concerns from clipping and noise in DPSGD

The two most significant steps in DPSGD, clipping and adding noise, can impact the learning process disproportionately across groups, but the exact conditions where disparate impact will occur have been debated [5, 18, 37, 34]. The most concrete connection so far appears in [34], where the expected loss  $\mathcal{L}(\theta; D_a)$  is decomposed into terms contributing to the excessive risk for group  $a$ ,  $R_a$ :

**Proposition 1** ([34]). *Consider the ERM problem with twice-differentiable loss  $\ell$  with respect to the model parameters. The expected loss  $\mathbb{E}[\mathcal{L}(\theta_{t+1}; D_a)]$  of group  $a \in [K]$  at iteration  $t$  is approximated up to second order in  $\|\theta_{t+1} - \theta_t\|$  as:*

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\theta_{t+1}; D_a)] &\approx \mathcal{L}(\theta_t; D_a) - \eta_t \langle g_{D_a}, g_D \rangle + \frac{\eta_t^2}{2} \mathbb{E}[g_B^T H_\ell^a g_B] && \text{(non-private term)} \\ &+ \eta_t \langle g_{D_a}, g_D - \bar{g}_B \rangle + \frac{\eta_t^2}{2} (\mathbb{E}[\bar{g}_B^T H_\ell^a \bar{g}_B] - \mathbb{E}[g_B^T H_\ell^a g_B]) && (R_a^{\text{clip}}) \\ &+ \frac{\eta_t^2}{2} \text{Tr}(H_\ell^a) C_0^2 \sigma^2. && (R_a^{\text{noise}}) \end{aligned}$$

The expectation is taken over the randomness of the DP mechanisms, and batches of data.

Terms in the first line appear for ordinary SGD, and do not contribute to the excessive risk Eq. (2). The terms in the second line,  $R_a^{\text{clip}}$ , are caused by clipping since they cancel when  $\bar{g}_B = g_B$  for every batch. They involve gradients  $g_{D_a}$  and Hessians  $H_\ell^a$ , averaged over datapoints belonging to group  $a$ . The final term,  $R_a^{\text{noise}}$ , depends on the scale of noise added in Alg. 1, as well as the trace of the group Hessian, also called the Laplacian, averaged over  $D_a$ . Based on Proposition 1, Tran et al. [34] conclude that clipping causes excessive risk to groups that have large gradient norms, which can result from large input norms  $\|x_i\|$ . Whether or not a group is underrepresented has less influence. In the next section we provide a new perspective on  $R_a^{\text{clip}}$  and the underlying causes of unfairness in DPSGD.

## 4 Disparate impact is caused by gradient misalignment

Clipping in DPSGD introduces two types of error to the clipped batch gradient  $\bar{g}_B$ . It will generally have different norm than  $\|g_B\|$ , and be misaligned compared to the SGD batch gradient,  $g_B$ . At a high level, gradient misalignment poses a more serious problem to the convergence of DPSGD than magnitude error, as illustrated in Fig. 1. Changing only the norm means gradient descent will still step towards the (local) minimum of the loss function, and the norm error could be completely compensated for by adapting the learning rate  $\eta_t$ . In contrast, a misaligned gradient could result

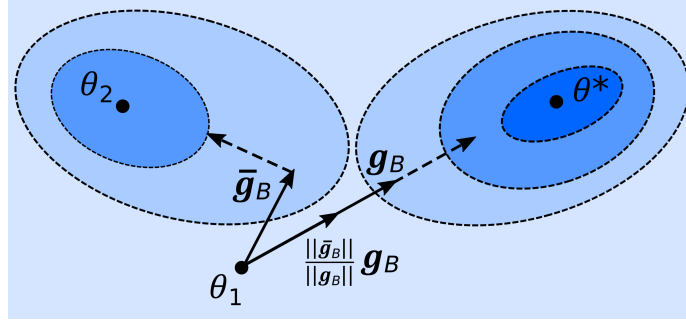


Figure 1: Direction errors from clipping are more severe than magnitude errors over the course of training and can lead to suboptimal convergence.

in a step towards significantly worse regions of the loss landscape causing catastrophic failures of convergence. We aim to quantify the relative impact of these effects and how they contribute to the excessive risk incurred due to clipping.

We can distinguish the effects by rewriting the clipped batch gradient as  $\bar{g}_B = M_B \left( \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right)$ <sup>3</sup> for an orthogonal matrix  $M_B$  such that  $\bar{g}_B$  and  $M_B g_B$  are colinear. As a proof of concept that gradient misalignment is the more severe error we compared models trained by taking steps  $\frac{\|\bar{g}_B\|}{\|g_B\|} g_B$  and  $M_B g_B$  with no noise. These represent magnitude errors and direction errors from clipping respectively. Details are provided in App. B. As seen in Table 1, direction error is more detrimental to performance than magnitude error. In particular, it disproportionately increases loss and decreases accuracy on the underrepresented class 8.

Table 1: Effect of direction vs. magnitude error on MNIST with class 8 underrepresented.

TYPE OF ERROR	ACC 2	ACC 8	LOSS 2	LOSS 8
MAGNITUDE	99.0	93.5	0.002	0.005
DIRECTION	96.8	84.1	0.076	0.518

Our first theoretical result quantifies the excessive risk from the two types of errors, and follows from a Taylor expansion of the expected loss using  $\bar{g}_B$  in the gradient descent update compared to  $g_B$ . The excessive risk from magnitude error comes from comparing  $g_B$  to  $\frac{\|\bar{g}_B\|}{\|g_B\|} g_B$ , while that of gradient misalignment is isolated by comparing  $\bar{g}_B = M_B \left( \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right)$  to  $\frac{\|\bar{g}_B\|}{\|g_B\|} g_B$  (see Fig. 1).

**Proposition 2.** *Consider the ERM problem with twice-differentiable loss  $\ell$  with respect to the model parameters. The excessive risk due to clipping experienced by group  $a \in [K]$  at iteration  $t$  is approximated up to second order in  $\|\theta_{t+1} - \theta_t\|$  as*

$$\begin{aligned}
R_a^{\text{clip}} \approx & \eta_t \left\langle g_{D_a}, \mathbb{E} \left[ \left( 1 - \frac{\|\bar{g}_B\|}{\|g_B\|} \right) g_B \right] \right\rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \left( \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} - 1 \right) g_B^T H_\ell^a g_B \right] & (R_a^{\text{mag}}) \\
& + \eta_t \left\langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (g_B - M_B g_B) \right] \right\rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right], & (R_a^{\text{dir}})
\end{aligned}$$

where  $g_Z$ ,  $\bar{g}_Z$  denote the average non-clipped and clipped gradients over  $Z \subseteq D$  at iteration  $t$ ,  $H_\ell^a$  refers to the Hessian over group  $a$ , and  $M_B$  is an orthogonal matrix such that  $\bar{g}_B$  and  $M_B g_B$  are colinear. The expectations are taken over batches of data.

We provide a derivation in App. A. Note that when the magnitude error is zero for all batches,  $\|g_B\| = \|\bar{g}_B\|$ , we have that  $R_a^{\text{mag}} = 0$  as expected. As well, when there is no gradient misalignment then  $M_B$  is the identity matrix for every batch, and so  $R_a^{\text{dir}} = 0$ .

<sup>3</sup>We point out that this is a relation, not the definition of  $\bar{g}_B$ ; both  $\bar{g}_B$  and  $g_B$  are defined in terms of the per-sample gradients that make up the batch  $B$ .

To determine the characteristics of groups that will have unfair outcomes from clipping in DPSGD we can distill a simpler condition for when  $R_a^{\text{dir}} > R_b^{\text{dir}}$ . [34] already provides such a condition for clipping overall, however it does not effectively account for the danger of gradient misalignment. Their condition is sufficient, but not necessary, and some of its looseness stems from the inequality  $x^T y \geq -\|x\|\|y\|$  used to convert all terms in  $R_a^{\text{clip}}$  into expressions involving group gradient norms. This approach loses information about gradient direction. We instead propose a tighter analysis of  $R_a^{\text{dir}} - R_b^{\text{dir}}$  using  $x^T y = \|x\|\|y\| \cos \theta$ , where  $\theta = \angle(x, y)$ .

**Proposition 3.** *Assume the loss  $\ell$  is twice continuously differentiable and convex with respect to the model parameters. As well, assume that  $\eta_t \leq (\max_{z \in [K]} \lambda_z)^{-1}$  where  $\lambda_z$  is the maximum eigenvalue of the Hessian  $H_\ell^z$ . For groups  $a, b \in [K]$ ,  $R_a^{\text{dir}} > R_b^{\text{dir}}$  if*

$$\mathbb{E} [\|\bar{g}_B\|(\cos \theta_B^a - \cos \bar{\theta}_B^a)] > \frac{\|g_{D_b}\|}{\|g_{D_a}\|} \mathbb{E} [\|\bar{g}_B\|(\cos \theta_B^b - \cos \bar{\theta}_B^b)] + \frac{\mathbb{E}[\|\bar{g}_B\|^2]}{\|g_{D_a}\|}, \quad (4)$$

where  $\theta_B^z = \angle(g_{D_z}, g_B)$  and  $\bar{\theta}_B^z = \angle(g_{D_z}, \bar{g}_B)$  for a group  $z \in [K]$ .

Thus, if the clipping operation disproportionately and sufficiently increases the direction error for group  $a$  relative to group  $b$  such that the condition above is satisfied, then group  $a$  incurs larger excessive risk due to gradient misalignment.

In our experiments we will empirically show this condition is a tight lower bound for  $R_a^{\text{dir}} - R_b^{\text{dir}}$ . Hence, when the direction error for groups  $a, b$  is small (i.e. we expect that  $\theta_B^i \approx \bar{\theta}_B^i$  for  $i = a, b$ ), we have that  $R_a^{\text{dir}} - R_b^{\text{dir}} \approx 0$  regardless of size of  $\|g_{D_a}\|$  relative to  $\|g_{D_b}\|$  or aggressive clipping. It follows that clipping does not negatively impact excessive risk if gradients are aligned. On the other hand if direction error is not close to zero, large group gradient norms do exacerbate the error in direction, as the dominant term of  $R_a^{\text{dir}}$  scales with  $\|g_{D_a}\|$ .

We note that ultimately the excessive risk Eq. (2) is evaluated for trained models, whereas Prop. 2 estimates it for a single iteration. Fig. 1 demonstrates that the full impact of clipping errors may not be felt in a single iteration, but only at convergence. We postulate that gradient misalignment is the main cause of disparate impact in private models and seek a method to prevent it.

## 5 Preventing gradient misalignment in DPSGD

Our results so far show that gradient misalignment due to clipping is a significant cause of unfairness in DPSGD, and is potentially more sinister than magnitude error. Logically,  $R_a^{\text{dir}}$  would be minimized if the clipping operation left the direction of  $g_B$  unchanged. There are several ways this could be guaranteed.

Using a single gradient per batch is the simplest way to ensure that  $g_B$  and  $\bar{g}_B$  are colinear. Per-sample clipping only modifies the magnitude of  $g_i$ , so if no aggregation is done over  $\bar{g}_i$ , then no direction error occurs. However, in DPSGD, Poisson sampling should be used [1, 39] which means the batch size is not fixed, and we cannot guarantee single datapoint batches. As a second possibility, gradient misalignment could be avoided if the distribution of per-sample gradients had certain symmetry properties. For example, the dominant term in  $R_a^{\text{dir}}$ , namely  $\langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (g_B - M_B g_B) \right] \rangle$ , would be zero if the direction of the batch gradient  $g_B$  does not change after per-sample clipping in expectation. Similar symmetry requirements approximately hold in some real-world settings [11], though their relevance to the sources of unfairness in DPSGD have not been noted before.

We will explore a third avenue for reducing gradient misalignment. The problem is clearly caused by gradients in a batch being scaled down by different amounts, and some not at all. A simple workaround is to *scale down all per-sample gradients by the same amount*, i.e.  $\bar{g}_i = \gamma g_i$ ,  $0 < \gamma < 1$ , such that their average  $\bar{g}_B = \gamma g_B$  is scaled down but still aligned. Scaling alone is insufficient to ensure per-sample gradients have bounded sensitivity. However, supposing that there were a strict upper bound  $Z \geq \|g_i\| \forall i \in D$  then scaling all gradients by  $\gamma = C_0/Z$  would guarantee bounded sensitivity of  $C_0$  for each  $\bar{g}_i$ . Given sufficient smoothness of the loss function, for any empirical sample of data there will be such an upper bound  $\max_{i \in D} \|g_i\|$ , but determining it exactly cannot be done in a differentially private manner.

One option is to choose  $Z$  as a hyperparameter without looking at the data, in the same way  $C_0$  is chosen in DPSGD. Then, if  $Z$  fails to be a strict upper bound, any gradients with  $\|g_i\| > Z$  can be

---

**Algorithm 2** DPSGD-Global-Adapt

---

**Require:** Iterations  $T$ , Dataset  $D$ , sampling rate  $q$ , clipping bound  $C_0$ , strict clipping bound  $Z \geq C_0$ , noise multipliers  $\sigma_1, \sigma_2$ , learning rates  $\eta_t$ , clipping learning rate  $\eta_Z$ , threshold  $\gamma \geq 0$

Initialize  $\theta_0$  randomly

**for**  $t$  in  $0, \dots, T - 1$  **do**

$B \leftarrow$  Poisson sample of  $D$  with rate  $q$

**for**  $(x_i, y_i)$  in  $B$  **do**

$g_i \leftarrow \nabla_{\theta_t} \ell(f_{\theta_t}(x_i), y_i)$

$\triangleright$  Compute per-sample gradients

$\tilde{g}_i \leftarrow C_0 \cdot g_i / \max(\|g_i\|, Z)$

$\triangleright$  Scale down all gradients to have norm less than  $C_0$

$\tilde{g}_B \leftarrow \frac{1}{|B|} (\sum_{i \in B} \tilde{g}_i + \mathcal{N}(0, \sigma_1^2 C_0^2 \mathbb{I}))$

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}_B$

$b_t \leftarrow |\{i : \|g_i\| > \gamma \cdot Z\}|$

$\tilde{b}_t \leftarrow \frac{1}{|B|} (b_t + \mathcal{N}(0, \sigma_2^2))$

$\triangleright$  Privately estimate count of how many gradients are “large”

$Z \leftarrow Z \cdot \exp(-\eta_Z + \tilde{b}_t)$

$\triangleright$  Adaptively update  $Z$

---

clipped to  $C_0$  to guarantee a bound on sensitivity. If  $Z$  is chosen sufficiently large, then no gradients are clipped to  $C_0$  and gradient misalignment will be avoided. The drawback of choosing a large  $Z$  is that the scaled gradients  $\tilde{g}_i$  will become small and convergence of gradient descent may be hindered.

We find a more robust solution is to adaptively update  $Z$  such that it approximates or upper-bounds  $\max_{i \in B} \|g_i\|$  for all batches  $B$ . When  $Z$  is larger than all gradients we would like to reduce  $Z$  so that gradients are scaled down less, but if  $Z$  is too small and gradients are being clipped  $Z$  should be increased.  $Z$  can be updated each iteration by privately estimating  $b_t$ , the number of gradients in the batch that are larger than  $Z$  times a tolerance threshold  $\gamma \geq 0$ . Since  $b_t$  is a unit sensitivity quantity we can estimate it well with the sampled Gaussian mechanism,  $\tilde{b}_t = \frac{1}{|B|} (b_t + \mathcal{N}(0, \sigma_2^2))$ . We use the geometric update rule  $Z \leftarrow Z \cdot \exp(-\eta_Z + \tilde{b}_t)$  with a learning rate  $\eta_Z$  (c.f. [4]). When  $b_t$  is near zero, in other words when most samples have gradient norm less than or equal to  $\gamma \cdot Z$ , then in expectation  $\tilde{b}_t = 0$  and  $Z$  is decreased by a factor of  $\exp(-\eta_Z)$ . However, when  $b_t$  is sufficiently large, then in expectation  $\tilde{b}_t \geq \eta_Z$  and so  $Z$  is increased.

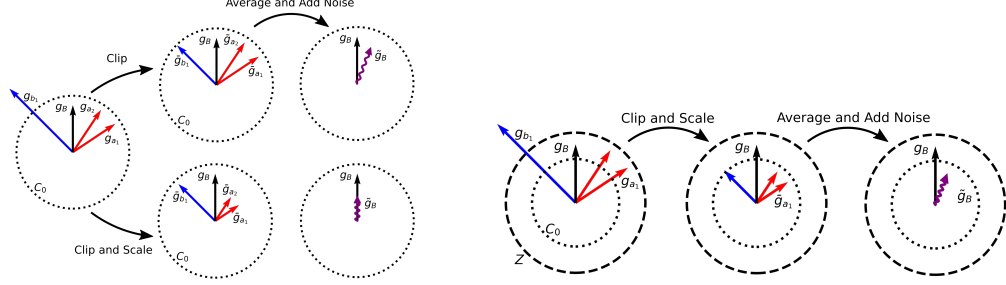
The additional step of privately estimating the number of gradients with norm larger than  $\gamma \cdot Z$  must be accounted for in the overall DP guarantee. However, our adaptive method is empirically not sensitive to the exact count  $b_t$ , so a relatively large amount of noise can be used, see [4] for comparison. In practice we used  $\sigma_2 \approx 10\sigma_1$  which produces a negligible impact on the overall DP guarantee.

The procedure described above and presented in Algorithm 2, which we call DPSGD-Global-Adapt, is similar to the recent method DPSGD-Global (Alg. 3 in App. B) [6], where scaling is applied globally to all per-sample gradients rather than locally to those with  $\|g_i\| > C_0$  as illustrated in Fig. 2. Compared to our approach, DPSGD-Global relies on  $Z$  being set correctly as a hyperparameter, and clips to zero rather than  $C_0$  which is unnecessarily aggressive and causes additional unfairness. DPSGD-Global was proposed for improving the convergence properties of DPSGD, but we have logically arrived at a similar algorithm as a method to mitigate unfairness.

Compared to other approaches for removing unfairness, such as DPSGD-F [37], DPSGD-Global-Adapt has the additional advantage of mitigating unfairness without accessing protected group data, which may be unavailable in practice.

## 6 Experiments

In our experiments we provide evidence that gradient misalignment is a significant cause of unfairness, and demonstrate that DPSGD-Global-Adapt can effectively reduce unfairness by aligning gradients. For comparison with prior work, we replicate as closely as possible the setting in [37]. Our code for reproducing the experiments will be provided at [github.com/layer6ai-labs/fair-dp](https://github.com/layer6ai-labs/fair-dp) upon publication.



(a) **Left:** Gradients are computed per-sample, and colored based on group membership. **Top:** DPSGD **Bottom:** DPSGD-Global-Adapt. (b) In DPSGD-Global-Adapt scaling alone does not guarantee finite sensitivity, so gradients with norm above  $Z$  are clipped to  $C_0$ .

Figure 2: DPSGD Variants

## 6.1 Experiment settings

For all experiments, full details are provided in App. B. We use an artificially unbalanced MNIST training dataset where class 8 only constitutes about 1% of the dataset on average, and protected group values are the class labels. We also use two census datasets popular in the ML fairness literature, Adult and Dutch, preprocessed as in [25]. For both datasets, by default, “sex” is the protected group attribute. Adult is balanced, while Dutch is unbalanced with a 3:1 male to female ratio.

We compare our method DPSGD-Global-Adapt with two methods designed to reduce unfairness, DPSGD-F [37] (Alg. 4) and the Fairness-Lens method [34] (Alg. 5) shown in App. B, as well as the cognate method DPSGD-Global [6]. Each method’s effectiveness in removing disparate impact is measured using privacy cost (Eq. 1), and excessive risk (Eq. 2) per group, as well as the privacy cost gap (PCG), and excessive risk gap (ERG) between groups. For MNIST, the underrepresented group 8 is compared to group 2 [37]. All experiments were run for 5 random seeds, and results are given as means  $\pm$  standard errors.

For MNIST, all methods train a convolutional neural network with two layers of 32 and 16 channels and tanh activations. For tabular datasets, an MLP model with two hidden layers of 256 units is used instead. For all private methods, we use an RDP accountant [28, 29] with  $\delta = 10^{-6}$ . As a baseline, for DPSGD we set  $\sigma = 1$ ,  $C_0 = 0.5$  for tabular datasets, and  $\sigma = 0.8$ ,  $C_0 = 1$  for MNIST. With this, training for 20 and 60 epochs respectively gives  $\epsilon = 3.41$  for Adult,  $\epsilon = 2.84$  for Dutch, and  $\epsilon = 5.90$  for MNIST. DPSGD-F has negligibly higher  $\epsilon$ , while our method achieves the same  $\epsilon$  guarantees to two significant digits. Complete hyperparameters are given in App. B.

## 6.2 Results

[37] finds that the male group experiences disparate impact from DPSGD in Adult, and Female for Dutch, while the underrepresented class 8 in MNIST experiences disparate impact.

Table 2: Performance and Fairness metrics for MNIST

METHOD	ACC 2	ACC 8	$\pi(2)$	$\pi(8)$	PCG	Loss 2	Loss 8	ER 2	ER 8	ERG
NON PRIVATE	98.0 $\pm$ 0.2	84.3 $\pm$ 2.6	-	-	-	0.06 $\pm$ 0.00	0.32 $\pm$ 0.03	-	-	-
DPSGD	89.0 $\pm$ 0.2	26.3 $\pm$ 0.8	8.9 $\pm$ 0.2	57.9 $\pm$ 2.9	48.9 $\pm$ 3.0	0.67 $\pm$ 0.03	2.56 $\pm$ 0.09	0.61 $\pm$ 0.03	2.24 $\pm$ 0.07	1.63 $\pm$ 0.06
DPSGD-F	89.5 $\pm$ 0.3	59.3 $\pm$ 1.0	8.5 $\pm$ 0.3	24.9 $\pm$ 2.9	16.4 $\pm$ 2.9	0.65 $\pm$ 0.02	1.47 $\pm$ 0.09	0.59 $\pm$ 0.02	1.16 $\pm$ 0.07	0.56 $\pm$ 0.09
DPSGD-G.	90.6 $\pm$ 0.5	62.0 $\pm$ 5.8	7.4 $\pm$ 0.4	22.2 $\pm$ 5.8	14.8 $\pm$ 6.0	0.34 $\pm$ 0.02	1.31 $\pm$ 0.08	0.28 $\pm$ 0.02	0.99 $\pm$ 0.08	0.71 $\pm$ 0.09
DPSGD-G.-A.	92.0 $\pm$ 0.5	65.5 $\pm$ 2.7	6.0 $\pm$ 0.5	18.8 $\pm$ 2.1	12.8 $\pm$ 2.1	0.35 $\pm$ 0.01	1.20 $\pm$ 0.08	0.29 $\pm$ 0.02	0.89 $\pm$ 0.08	0.60 $\pm$ 0.06

Table 2 displays the accuracy and loss, along with privacy cost and excessive risk metrics for MNIST<sup>4</sup> (cf. Tables 3 and 4 in App. B for Adult and Dutch). Recall that higher is better for accuracy, but for all other metrics lower is better. We see that our method for reducing gradient misalignment outperforms both DPSGD-F and DPSGD in model performance as well as privacy cost and excessive risk on all metrics.

<sup>4</sup>The Fairness Lens method [34] is not compared for MNIST because the author-provided code only handles binary classification problems.



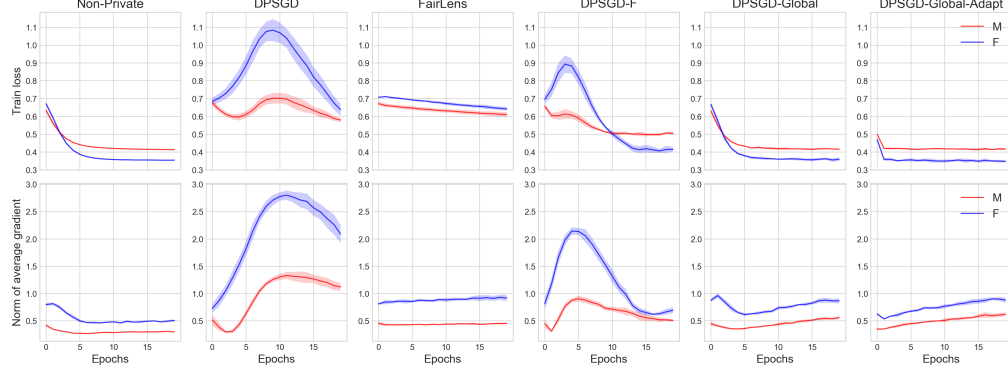


Figure 3: Dutch dataset: (a) Train loss per epoch (b)  $\|g_B\|$ , averaged over batches per epoch

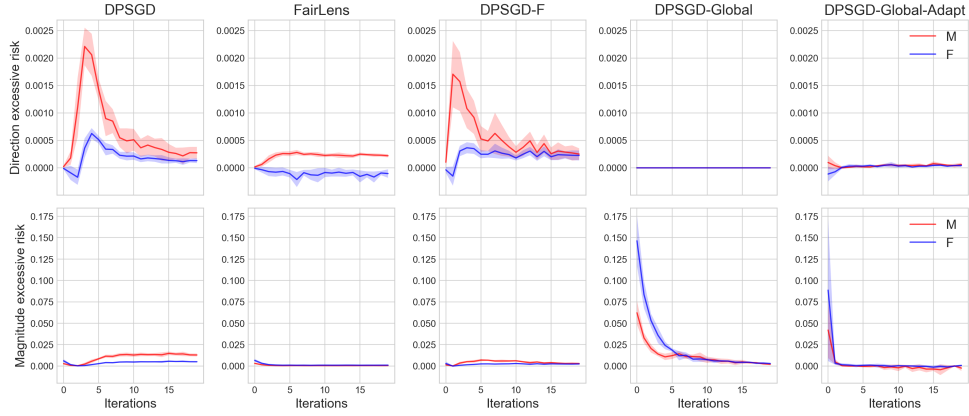


Figure 4: Adult dataset: (a)  $R_a^{\text{dir}}$ , excessive risk due to gradient misalignment per group. (b)  $R_a^{\text{mag}}$ , excessive risk due to magnitude error per group. See Proposition 2 for definitions.

Not only are final model metrics improved, we see more stable training with DPSGD-Global-Adapt in Fig. 3 for Dutch (cf. Figs. 7 and 8 in App. B for Adult and MNIST). This shows the average train loss per iteration, and average norm of the batched gradient. While the improvement in convergence from DPSGD to DPSGD-F is clear, the convergence for both groups in DPSGD-Global-Adapt resembles that of the non-private method much more closely. Consider Fig. 3(b), where the Male group average norms do not converge to 0 in DPSGD, which is somewhat improved in DPSGD-F. In DPSGD-Global-Adapt the norms for both groups converge to zero quickly, and the gap between the two is quickly reduced.

Fig. 4 shows the excessive risk terms due to gradient misalignment  $R_a^{\text{dir}}$ , and magnitude error  $R_a^{\text{mag}}$  for Adult at each iteration (see Figs. 9 and 10 in App. B for Dutch and MNIST). We see that global clipping almost completely removes direction errors as intended, but as a tradeoff increases magnitude error. However, we have argued that direction error is a more severe cause of disparate impact (and poor model convergence), which is borne out by the results in Tables 1, 2, 3, and 4.

### 6.3 Tightness of lower bounds

We compare the usefulness of the lower bound of  $R_a^{\text{clip}} - R_b^{\text{clip}}$  given in the proof of Theorem 3 in [34], to the lower bound we give in Proposition 3 for  $R_a^{\text{dir}} - R_b^{\text{dir}}$ . We see that while group 0 experiences disparate impact due to clipping, the lower bound in 5 from [34] is negative for each iteration, failing to capture that  $R_0^{\text{clip}} > R_1^{\text{clip}}$ . On the other hand, the true values of  $R_0^{\text{dir}} - R_1^{\text{dir}}$  are closely lower-bounded in our version, such that disparate impact due to direction is accurately predicted.

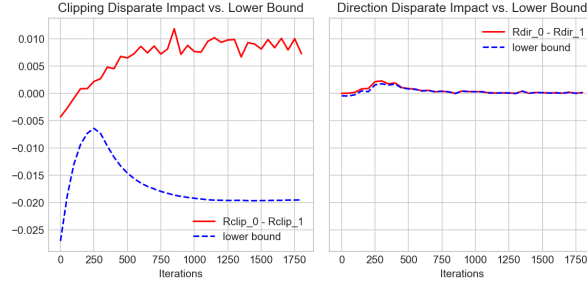


Figure 5: Comparison of exact excessive risk gaps (ERG) to lower bounds on Adult. **Left:** ERG due to clipping error, and lower bound from [34]. **Right:** ERG due to direction error, and lower bound from our Proposition 3.

## 7 Discussion

In this paper we identified a core cause of disparate impact in DPSGD, gradient misalignment, and proposed a mitigating solution, DPSGD-Global-Adapt. We empirically verified that DPSGD-Global-Adapt is successful in improving fairness in terms of accuracy and loss over DPSGD and other fair baselines on several datasets. Our method has additional advantages over other fair baselines in that it does not require the collection of protected group data, and it removes disparate impact for all groups simultaneously.

It is important to note that while DPSGD-Global-Adapt is effective at reducing disparate impact by aligning gradients, it does not resolve the privacy-utility trade-off, which exists in any private mechanism fundamentally. Nor does it ensure that the model is non-discriminatory towards subgroups, only that adding privacy does not exacerbate unfairness. For example, biases in data collection or discriminatory modelling assumptions can cause disparate impact within the non-private model, which overlaying DPSGD-Global-Adapt will not cure. Any models trained with DPSGD-Global-Adapt should still be validated for fairness independently; failure to do so could unknowingly cause additional unfairness.

## References

- [1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
- [2] J. M. Abowd. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2867, 2018. ISBN 9781450355520.
- [3] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1–10, 2022.
- [4] G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy. Differentially Private Learning with Adaptive Clipping. In *Advances in Neural Information Processing Systems*, volume 34, pages 17455–17466, 2021.
- [5] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32:15479–15488, 2019.
- [6] Z. Bu, H. Wang, Q. Long, and W. J. Su. On the Convergence of Deep Learning with Differential Privacy. *CoRR*, abs/2106.07830, 2021. URL <https://arxiv.org/abs/2106.07830>.
- [7] J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91, 2018.

- [8] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.
- [9] H. Chang and R. Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 292–303, 2021. doi: 10.1109/EuroSP51992.2021.00028.
- [10] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.
- [11] X. Chen, S. Z. Wu, and M. Hong. Understanding Gradient Clipping in Private SGD: A Geometric Perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 13773–13782, 2020. URL <https://proceedings.neurips.cc/paper/2020/file/9ecff5455677b38d19f49ce658ef0608-Paper.pdf>.
- [12] C. A. Choquette-Choo, N. Dullerud, A. Dziedzic, Y. Zhang, S. Jha, N. Papernot, and X. Wang. CaPC Learning: Confidential and Private Collaborative Learning. In *International Conference on Learning Representations*, 2020.
- [13] A. Chouldechova and A. Roth. A Snapshot of the Frontiers of Fairness in Machine Learning. *Commun. ACM*, 63(5):82–89, 2020. ISSN 0001-0782.
- [14] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. On the Compatibility of Privacy and Fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, UMAP’19 Adjunct, page 309–315, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367110. doi: 10.1145/3314183.3323847. URL <https://doi.org/10.1145/3314183.3323847>.
- [15] C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, Aug. 2014. ISSN 1551-305X. doi: 10.1561/04000000042. URL <https://doi.org/10.1561/04000000042>.
- [16] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.
- [17] M. D. Ekstrand, R. Joshaghani, and H. Mehrpouyan. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 35–47, 2018.
- [18] T. Farrand, F. Mireshghallah, S. Singh, and A. Trask. 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*, page 15–19, 2020.
- [19] C. Gentry. Fully Homomorphic Encryption Using Ideal Lattices. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, page 169–178, 2009.
- [20] M. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2): 433–450, 1990. doi: 10.1080/03610919008812866. URL <https://doi.org/10.1080/03610919008812866>.
- [21] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. S. Malvajerdi, and J. Ullman. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3000–3008, 2019.
- [22] M. Jagielski, J. Ullman, and A. Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

- [23] M. Jaiswal and E. Mower Provost. Privacy Enhanced Multimodal Neural Representations for Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7985–7993, 2020. doi: 10.1609/aaai.v34i05.6307. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6307>.
- [24] S. Kalra, J. Wen, J. C. Cresswell, M. Volkovs, and H. R. Tizhoosh. ProxyFL: Decentralized Federated Learning through Proxy Model Sharing. *arXiv preprint arXiv:2111.11343*, 2021.
- [25] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, page e1452, 2022.
- [26] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017. URL <http://arxiv.org/abs/1602.05629>.
- [27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6), 2021.
- [28] I. Mironov. Rényi differential privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, Aug 2017. doi: 10.1109/csf.2017.11. URL <http://dx.doi.org/10.1109/CSF.2017.11>.
- [29] I. Mironov, K. Talwar, and L. Zhang. Rényi Differential Privacy of the Sampled Gaussian Mechanism. *arXiv preprint arXiv:1908.10530*, 2019.
- [30] H. Mozannar, M. Ohannessian, and N. Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR, 2020.
- [31] M. Nasr, S. Songi, A. Thakurta, N. Papemoti, and N. Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 866–882. IEEE, 2021.
- [32] V. Pichapati, A. T. Suresh, F. X. Yu, S. J. Reddi, and S. Kumar. AdaClip: Adaptive Clipping for Private SGD. *CoRR*, abs/1908.07643, 2019. URL <http://arxiv.org/abs/1908.07643>.
- [33] D. Pujol, R. McKenna, S. Kuppam, M. Hay, A. Machanavajjhala, and G. Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 189–199, 2020.
- [34] C. Tran, M. Dinh, and F. Fioretto. Differentially Private Empirical Risk Minimization under the Fairness Lens. In *Advances in Neural Information Processing Systems*, volume 34, pages 27555–27565, 2021. URL <https://proceedings.neurips.cc/paper/2021/file/e7e8f8e5982b3298c8addedf6811d500-Paper.pdf>.
- [35] C. Tran, F. Fioretto, and P. Van Hentenryck. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9932–9939, May 2021. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17193>.
- [36] D. Xu, S. Yuan, and X. Wu. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference, WWW ’19*, page 594–599, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317584. URL <https://doi.org/10.1145/3308560.3317584>.
- [37] D. Xu, W. Du, and X. Wu. Removing Disparate Impact on Model Accuracy in Differentially Private Stochastic Gradient Descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, page 1924–1932, 2021.
- [38] A. C.-C. Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science*, pages 162–167, 1986.
- [39] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex. Differentially private model publishing for deep learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 332–349, 2019.

## A Theoretical Results

### A.1 Proofs of main results

In this section we provide complete proofs for our theoretical contributions.

**Proposition 2.** *Consider the ERM problem with twice-differentiable loss  $\ell$  with respect to the model parameters. The excessive risk due to clipping experienced by group  $a \in [K]$  at iteration  $t$  is approximated up to second order in  $\|\theta_{t+1} - \theta_t\|$  as*

$$\begin{aligned} R_a^{\text{clip}} &\approx \eta_t \left\langle g_{D_a}, \mathbb{E} \left[ \left( 1 - \frac{\|\bar{g}_B\|}{\|g_B\|} \right) g_B \right] \right\rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \left( \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} - 1 \right) g_B^T H_\ell^a g_B \right] \\ &+ \eta_t \left\langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (g_B - M_B g_B) \right] \right\rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right], \end{aligned} \quad \begin{aligned} (R_a^{\text{mag}}) \\ (R_a^{\text{dir}}) \end{aligned}$$

where  $g_Z, \bar{g}_Z$  denote the average non-clipped and clipped gradients over  $Z \subseteq D$  at iteration  $t$ ,  $H_\ell^a$  refers to the Hessian over group  $a$ , and  $M_B$  is an orthogonal matrix such that  $\bar{g}_B$  and  $M_B g_B$  are colinear. The expectations are taken over batches of data.

*Proof.*

The proof is based on a Taylor expansion of the excessive risk, as in [34].

Let  $M_B$  be an orthogonal matrix such that  $\bar{g}_B = M_B \left( \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right)$ . In this way,  $\|\bar{g}_B\| = \left\| \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right\|$  and  $g_B$  and  $\frac{\|\bar{g}_B\|}{\|g_B\|} g_B$  are colinear, and so the former characterizes direction error, and the latter error in magnitude. The excessive risk due to error in magnitude for group  $a$  at iteration  $t$  is then given by

$$\mathbb{E}[\mathcal{L}(\theta_t - \eta_t \frac{\|\bar{g}_B\|}{\|g_B\|} g_B; D_a) - \mathcal{L}(\theta_t - \eta_t g_B; D_a)],$$

the cost in loss of using the update vector  $\frac{\|\bar{g}_B\|}{\|g_B\|} g_B$  rather than  $g_B$ , where the expectation is over randomness of batch sampling. We perform second-order Taylor expansion of  $\mathbb{E}[\mathcal{L}(\theta_t - \eta_t \frac{\|\bar{g}_B\|}{\|g_B\|} g_B; D_a)]$  and take the expectation to get that

$$\mathbb{E}[\mathcal{L}(\theta_t - \eta_t \frac{\|\bar{g}_B\|}{\|g_B\|} g_B; D_a)] \approx \mathcal{L}(\theta_t; D_a) - \eta_t \langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right] \rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} g_B^T H_\ell^a g_B \right].$$

Hence,

$$\begin{aligned} R_a^{\text{clip}} &= \eta_t \langle g_{D_a}, g_D - \bar{g}_D \rangle + \frac{\eta_t^2}{2} (\mathbb{E}[\bar{g}_B^T H_\ell^a \bar{g}_B] - \mathbb{E}[g_B^T H_\ell^a g_B]) \\ &= \eta_t \langle g_{D_a}, g_D - \bar{g}_D \rangle + \frac{\eta_t^2}{2} (\mathbb{E}[\bar{g}_B^T H_\ell^a \bar{g}_B] - \mathbb{E}[g_B^T H_\ell^a g_B]) \\ &\quad - \eta_t \langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right] \rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} g_B^T H_\ell^a g_B \right] \\ &\quad + \eta_t \langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right] \rangle - \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} g_B^T H_\ell^a g_B \right] \\ &= \eta_t \langle g_{D_a}, g_D - \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right] \rangle + \frac{\eta_t^2}{2} \left( \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} g_B^T H_\ell^a g_B \right] - \mathbb{E} [g_B^T H_\ell^a g_B] \right) \\ &\quad + \eta_t \langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right] - \bar{g}_D \rangle + \frac{\eta_t^2}{2} \left( \mathbb{E} [\bar{g}_B^T H_\ell^a \bar{g}_B] - \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} g_B^T H_\ell^a g_B \right] \right) \\ &= \eta_t \langle g_{D_a}, g_D - \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right] \rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \left( \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} - 1 \right) g_B^T H_\ell^a g_B \right] \\ &\quad + \eta_t \langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right] - \bar{g}_D \rangle + \frac{\eta_t^2}{2} \left( \mathbb{E} [\bar{g}_B^T H_\ell^a \bar{g}_B] - \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} g_B^T H_\ell^a g_B \right] \right). \end{aligned} \quad \begin{aligned} (R_a^{\text{mag}}) \\ (R_a^{\text{dir}}) \end{aligned}$$

We can also further simplify  $R_a^{\text{dir}}$  by using that  $\bar{g}_D = \mathbb{E}[\bar{g}_B]$ ,  $\bar{g}_B = M_B \left( \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right)$  and that  $M_B$  is a linear transformation

$$\begin{aligned} R_a^{\text{dir}} &= \eta_t \langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} g_B - M_B \left( \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right) \right] \rangle + \frac{\eta_t^2}{2} \left( \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} (M_B g_B)^T H_\ell^a (M_B g_B) \right] \right. \\ &\quad \left. - \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} g_B^T H_\ell^a g_B \right] \right) \quad (5) \\ &= \eta_t \langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (g_B - M_B g_B) \right] \rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right]. \quad (6) \end{aligned}$$

□

**Proposition 3.** Assume the loss  $\ell$  is twice continuously differentiable and convex with respect to the model parameters. As well, assume that  $\eta_t \leq (\max_{z \in [K]} \lambda_z)^{-1}$  where  $\lambda_z$  is the maximum eigenvalue of the Hessian  $H_\ell^z$ . For groups  $a, b \in [K]$ ,  $R_a^{\text{dir}} > R_b^{\text{dir}}$  if

$$\mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^a - \cos \bar{\theta}_B^a)] > \frac{\|g_{D_b}\|}{\|g_{D_a}\|} \mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^b - \cos \bar{\theta}_B^b)] + \frac{\mathbb{E} [\|\bar{g}_B\|^2]}{\|g_{D_a}\|}, \quad (7)$$

where  $\theta_B^z = \angle(g_{D_z}, g_B)$  and  $\bar{\theta}_B^z = \angle(g_{D_z}, \bar{g}_B)$  for a group  $z \in [K]$ .

*Proof.*

This proof follows some steps presented in Lemma 2 of [34]. We seek a simplified condition for when the following is positive,

$$R_a^{\text{dir}} - R_b^{\text{dir}} = \eta_t \left\langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (g_B - M_B g_B) \right] \right\rangle - \eta_t \left\langle g_{D_b}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (g_B - M_B g_B) \right] \right\rangle \quad (8)$$

$$+ \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right] \quad (9)$$

$$- \frac{\eta_t^2}{2} \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^b (M_B g_B) - g_B^T H_\ell^b g_B) \right]. \quad (10)$$

Looking at one of the inner product terms, we use that  $\langle x, y \rangle = \|x\| \|y\| \cos(x, y)$  and linearity of expectation to obtain

$$\left\langle g_{D_a}, \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (g_B - M_B g_B) \right] \right\rangle = \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (\langle g_{D_a}, g_B \rangle - \langle g_{D_a}, M_B g_B \rangle) \right] \quad (11)$$

$$= \|g_{D_a}\| \mathbb{E} \left[ \frac{\|\bar{g}_B\|}{\|g_B\|} (\|g_B\| \cos(g_{D_a}, g_B) - \|M_B g_B\| \cos(g_{D_a}, M_B g_B)) \right] \quad (12)$$

$$= \|g_{D_a}\| \mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^a - \cos \bar{\theta}_B^a)], \quad (13)$$

where  $\theta_B^a = \angle(g_{D_a}, g_B)$  and  $\bar{\theta}_B^a = \angle(g_{D_a}, M_B g_B) = \angle(g_{D_a}, \bar{g}_B)$ . The last equality follows from the definition of  $M_B$  such that  $\bar{g}_B$  and  $M_B g_B$  are aligned and  $\|g_B\| = \|M_B g_B\|$ .

We can also get a bound on the difference in conjugates of the Hessian,  $\mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right]$ . Note that since we assume the loss  $\ell$  is convex, the Hessian  $H_\ell^a$  is positive semi-definite such that  $x^T H_\ell^a x \geq 0$  for all vectors  $x$ . It follows that  $\mathbb{E}[x^T H_\ell^a x] \geq 0$  and so using linearity of expectation,

$$\mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right] \leq \mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} (M_B g_B)^T H_\ell^a (M_B g_B) \right]. \quad (14)$$

Since  $\ell$  is twice continuously differentiable we have that  $H_\ell^a$  is symmetric and hence  $x^T H_\ell^a x \leq \lambda_a \|x\|^2$  where  $\lambda_a$  is the maximum eigenvalue of  $H_\ell^a$ . We then again use that  $\|M_B g_B\| = \|g_B\|$  and linearity of expectation to obtain

$$\mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right] \leq \lambda_a \mathbb{E} [\|\bar{g}_B\|^2]. \quad (15)$$

Similar analysis gives that  $\mathbb{E} \left[ \frac{\|\bar{g}_B\|^2}{\|g_B\|^2} ((M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B) \right] \geq -\lambda_a \mathbb{E} [\|\bar{g}_B\|^2]$ .

Combining the above, it follows that

$$R_a^{\text{dir}} - R_b^{\text{dir}} \geq \eta_t (\|g_{D_a}\| \mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^a - \cos \bar{\theta}_B^a)] - \|g_{D_b}\| \mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^b - \cos \bar{\theta}_B^b)]) \quad (16)$$

$$- \frac{\eta_t^2}{2} (\lambda_a + \lambda_b) \mathbb{E} [\|\bar{g}_B\|^2] \quad (17)$$

and since we assume  $\eta_t \leq \frac{1}{\max_{k \in [K]} \lambda_k}$

$$R_a^{\text{dir}} - R_b^{\text{dir}} \geq \eta_t (\|g_{D_a}\| \mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^a - \cos \bar{\theta}_B^a)] - \|g_{D_b}\| \mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^b - \cos \bar{\theta}_B^b)] - \mathbb{E} [\|\bar{g}_B\|^2]) \quad (18)$$

It follows that  $R_a^{\text{dir}} > R_b^{\text{dir}}$  when the following is satisfied:

$$\mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^a - \cos \bar{\theta}_B^a)] > \frac{\|g_{D_b}\|}{\|g_{D_a}\|} \mathbb{E} [\|\bar{g}_B\| (\cos \theta_B^b - \cos \bar{\theta}_B^b)] + \frac{\mathbb{E} [\|\bar{g}_B\|^2]}{\|g_{D_a}\|} \quad (19)$$

□

## A.2 Alternate decompositions of the clipping error

In Section A.2 we proposed a decomposition of the clipped batch gradient into parts representing magnitude and direction error,  $\bar{g}_B = M_B \left( \frac{\|\bar{g}_B\|}{\|g_B\|} g_B \right)$ . We presented a simple experiment in Table 1 to demonstrate that direction error causes the most severe problems for the final performance of models, and analysed the contributions of the two effects to the excessive risk in Proposition 2.

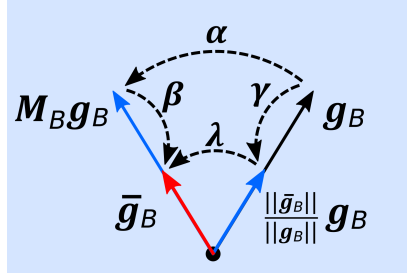


Figure 6: Decomposition of steps between  $g_B$  and  $\bar{g}_B$ .

However, the decomposition we used is not unique, and furthermore it is not possible to completely isolate the two effects in the excessive risk analysis. For example, if we think of magnitude error as the difference in loss between using update vector  $g_B$  and  $\frac{\|\bar{g}_B\|}{\|g_B\|} g_B$  ( $\gamma$  in Fig. 6), then it follows that the remaining error is due to gradient misalignment, in other words, the difference in loss between using update vector  $\frac{\|\bar{g}_B\|}{\|g_B\|} g_B$  and  $\bar{g}_B$  ( $\lambda$  in Fig. 6). In this example, the error due to gradient misalignment includes both error in direction and error in magnitude, while magnitude error is “pure”,

$$R_a^{\text{mag}} = \mathbb{E}[\mathcal{L}(\theta_t - \eta_t \frac{\|\bar{g}_B\|}{\|g_B\|} g_B; D_a) - \mathcal{L}(\theta_t - \eta_t g_B; D_a)], \quad (20)$$

$$R_a^{\text{dir}} = \mathbb{E}[\mathcal{L}(\theta_t - \eta_t \bar{g}_B; D_a) - \mathcal{L}(\theta_t - \eta_t \frac{\|\bar{g}_B\|}{\|g_B\|} g_B; D_a)]. \quad (21)$$

A different way of decomposing the clipping error is considering the direction error as the difference in loss between using update vector  $g_B$  and  $M_B g_B$  ( $\alpha$  in Fig. 6). In this case, direction error is pure, i.e. does not include difference in magnitudes. It follows that the remaining error is magnitude error, so is the difference in loss between using update vector  $M_B g_B$  and  $\bar{g}_B$  ( $\beta$  in Fig. 6). Thus, the magnitude error in this case quantifies the difference in loss of scaling the already misaligned  $\bar{g}_B$ ,

$$R_a^{\text{dir}*} = \mathbb{E}[\mathcal{L}(\theta_t - \eta_t M_B g_B; D_a) - \mathcal{L}(\theta_t - \eta_t g_B; D_a)], \quad (22)$$

$$R_a^{\text{mag}*} = \mathbb{E}[\mathcal{L}(\theta_t - \eta_t \bar{g}_B; D_a) - \mathcal{L}(\theta_t - \eta_t M_B g_B; D_a)]. \quad (23)$$

In our analysis we used the first decomposition where magnitude error can be completely corrected by an adjustment of the learning rate, and direction error, what we hypothesized to be the largest cause of disparate impact, is the remaining part of the clipping error. For completeness, by using the second decomposition we can derive alternative versions of Propositions 2 and 3:

**Proposition 2\*.** *Consider the ERM problem with twice-differentiable loss  $\ell$  with respect to the model parameters. The excessive risk due to clipping experienced by group  $a \in [K]$  at iteration  $t$  is approximated up to second order in  $\|\theta_{t+1} - \theta_t\|$  as*

$$R_a^{\text{clip}} \approx \eta_t \left\langle g_{D_a}, \mathbb{E} \left[ \left( \frac{\|g_B\|}{\|\bar{g}_B\|} - 1 \right) \bar{g}_B \right] \right\rangle + \frac{\eta_t^2}{2} \mathbb{E} \left[ \left( 1 - \frac{\|g_B\|^2}{\|\bar{g}_B\|^2} \right) \bar{g}_B^T H_\ell^a \bar{g}_B \right], \quad (R_a^{\text{mag}*})$$

$$+ \mathbb{E}[\eta_t \langle g_{D_a}, g_D - M_B g_B \rangle] + \frac{\eta_t^2}{2} \mathbb{E}[(M_B g_B)^T H_\ell^a (M_B g_B) - g_B^T H_\ell^a g_B], \quad (R_a^{\text{dir}*})$$

where  $g_Z, \bar{g}_Z$  denote the average non-clipped and clipped gradients over  $Z \subseteq D$  at iteration  $t$ ,  $H_\ell^a$  refers to the Hessian over group  $a$ , and  $M_B$  is an orthogonal matrix such that  $\bar{g}_B$  and  $M_B g_B$  are colinear. The expectations are taken over batches of data.

**Proposition 3\*.** *Assume the loss  $\ell$  is twice continuously differentiable and convex with respect to the model parameters. As well, assume that  $\eta_t \leq (\max_{z \in [K]} \lambda_z)^{-1}$  where  $\lambda_z$  is the maximum eigenvalue of the Hessian  $H_\ell^z$ . For groups  $a, b \in [K]$ ,  $R_a^{\text{dir}} > R_b^{\text{dir}}$  if*

$$\mathbb{E}[\|g_B\|(\cos \theta_B^a - \cos \bar{\theta}_B^a)] > \frac{\|g_{D_b}\|}{\|g_{D_a}\|} \mathbb{E}[\|g_B\|(\cos \theta_B^b - \cos \bar{\theta}_B^b)] + \frac{\mathbb{E}[\|g_B\|]}{\|g_{D_a}\|} \quad (24)$$

where  $\theta_B^z = \angle(g_{D_z}, g_B)$  and  $\bar{\theta}_B^z = \angle(g_{D_z}, \bar{g}_B)$  for a group  $z \in [K]$ .

We omit the proofs since they are directly analogous to those in A.1.

## B Experimental Details

### B.1 Dataset preprocessing

**MNIST** We use the artificially unbalanced MNIST training dataset where class 8 is sampled with probability 9% such that class 8 only constitutes about 1% of the dataset on average. This gives about 6000 data samples for each class, other than class 8 with about 500. The protected group values are the class labels. As in [37], we compare models on how they treat the under-represented class 8 versus the well-represented class 2. The test set remains balanced, with approximately 1000 samples for each class. Data is normalized to be in the domain  $[0, 1]$ .

**Adult** The original Adult dataset<sup>5</sup> consists of 48,842 samples, reduced to 45,222 by removing all samples with missing values. The final weight feature is removed and the race attribute is discretized by {white, non-white}, giving 5 numerical, 3 binary and 6 categorical features. The numerical features are normalized and the categorical features are one-hot encoded. As is typical in the fairness literature, choices for the protected attribute are “sex”, “race” (binary) and possibly the discretized “age”. We use “sex” by default. The classification label is “income” (whether income exceeds \$50K). Prior to sampling, the Adult dataset is unbalanced with respect to sex with 30,527 males and 14,695 females. We sample a balanced dataset as in [37] with 14,000 females and 14,000 males on average.

<sup>5</sup>The Adult dataset is available at [archive.ics.uci.edu/ml/datasets/Adult](http://archive.ics.uci.edu/ml/datasets/Adult).



**Dutch** The Dutch dataset<sup>6</sup> is preprocessed by dropping underage samples (14 and under) and removing the weight feature. As well, all unemployed samples are removed, as well as those with missing or middle-level occupation, for a total of 60,420 samples. Specifically, occupation values 3,6,7,8 are considered middle-level. Occupation is then made binary by considering values 4,5,9 as low-level professions (0) and 1,2 as high-level professions (1). The binary classification task is to predict occupation, given the rest of the features. We consider “sex” as the protected group attribute. The original (processed) dataset is balanced with respect to sex with 30,147 male and 30,273 female samples. We follow Xu et al. [37] and sample an unbalanced dataset with approximately a 3:1 male to female ratio, for 30K male samples, and 10K female on average.

We use a 80/20 train/test split for both tabular datasets.

## B.2 Experiment settings

For tabular datasets, we set  $\sigma = 1$ ,  $C_0 = 0.5$ , while for MNIST, we set  $\sigma = 0.8$  and  $C_0 = 1$ . For DPSGD-F, the gradient noise is unchanged  $\sigma_2 = \sigma$ , and  $\sigma_1 = 10\sigma_2$ . For FairLens, we use regularization weights as in [34],  $\lambda_1 = \lambda_2 = 1$ . For non-global methods, the learning rate is  $\eta = 0.01$ . For DPSGD-Global we have  $\eta = 1$ ,  $Z = 50$  for Adult,  $\eta = 0.2$ ,  $Z = 50$  for Dutch, and  $\eta = 0.2$ ,  $Z = 100$  for MNIST. For DPSGD-Global-Adapt we have  $\sigma_2 = 10$ ,  $Z = 50$ ,  $\eta_Z = 0.1$  for all datasets and  $\eta = 0.2$ ,  $\gamma = 1$  for Adult,  $\eta = 1$ ,  $\gamma = 1$  for Dutch, and  $\eta = 0.1$ ,  $\gamma = 0.7$  for MNIST. All methods for all datasets use training and test batches of size 256.

Experiments were conducted on single TITAN V GPU machines. Approximately two GPU-days were used to train each method over five seeds for the three datasets.

## B.3 Implementation Details

The excessive risk terms for different groups ( $R_a^{\text{clip}}$  and  $R_a^{\text{noise}}$  in Proposition 1 and  $R_a^{\text{mag}}$  and  $R_a^{\text{dir}}$  in Proposition 2) all involve the Hessian of the loss function with respect to the model parameters. Calculating the Hessian as a matrix is computationally expensive, but more crucially requires memory that scales quadratically in the number of parameters. In the previous work studying  $R_a^{\text{clip}}$  and  $R_a^{\text{noise}}$ , Tran et al. [34] use the PyHessian library to compute the Hessian as a matrix, and then used it to compute the products and traces needed for  $R_a^{\text{clip}}$  and  $R_a^{\text{noise}}$ . Because this approach incurs a high memory burden, the models trained were limited to small MLPs with a single hidden layer of 20 hidden units.<sup>7</sup>

In our implementation, provided as supplemental material, we avoid computing the Hessian as a matrix altogether which allows us to scale our experiments to reproduce the setting of [37]. For the three datasets, our MLPs have parameter counts of  $N = 91650$  for Adult,  $N = 81666$  for Dutch, and  $N = 80522$  for MNIST, which would produce Hessian matrices with between 6 and 9 billion entries. Instead, we compute the terms involving Hessians like  $H_\ell^a g_B$  through Hessian-vector products (HVPs) using the functorch<sup>8</sup> library with PyTorch 1.11. Using HVPs requires memory comparable to that used when computing gradients for SGD.

For the trace of the Hessian matrix, also called the Laplacian, one possible approach that does not require realizing the entire matrix in memory is to compute HVPs with unit vectors to isolate each diagonal element:  $\text{Tr}(H_\ell^a) = \sum_{i=1}^N I_i^T H_\ell^a I_i$  where  $I_i$  is the  $i$ th column of the identity matrix. While exact, this approach requires  $N$  HVPs for each group  $a \in K$ , of which there are at least two. Since this method is much too expensive for even the simple MLPs and CNNs we used, we instead employed Hutchinson’s trace estimator [20] to estimate  $\text{Tr}(H_\ell^a) = \mathbb{E}_z[z^T H_\ell^a z]$ . This estimator is unbiased when  $z$  is drawn from a Rademacher distribution which we used, and only requires  $n$  HVPs per group, where  $n$  can be chosen as large as required for convergence of the estimate. In practice we used  $n = 100$ .

<sup>6</sup>The Dutch dataset is available through the work [25] at [raw.githubusercontent.com/tailequy/fairness\\_dataset/main/Dutch\\_census/dutch\\_census\\_2001.arff](https://raw.githubusercontent.com/tailequy/fairness_dataset/main/Dutch_census/dutch_census_2001.arff).

<sup>7</sup>See implementation available at [openreview.net/forum?id=7EFdodSWee4](https://openreview.net/forum?id=7EFdodSWee4).

<sup>8</sup>See documentation at [pytorch.org/functorch/stable/](https://pytorch.org/functorch/stable/).

Additionally, whereas [34] replaces dataset gradients  $g_D$  and  $g_{D_a}$  with batch gradients when computing  $R_a^{\text{clip}}$  and  $R_a^{\text{noise}}$  in Proposition 1, we use the exact  $g_D$  and  $g_{D_a}$ . This eliminates an easily preventable source of noise in our results.

To further reduce computation time, we only evaluate excessive risk terms (Hessians) every 50, 100, 200 iterations for Adult, Dutch, and MNIST datasets respectively.

#### B.4 Direction error is more severe than magnitude error

As noted earlier, Proposition 2 only evaluates excessive risk for a single iteration, not necessarily capturing how each of  $R_a^{\text{dir}}$  and  $R_a^{\text{mag}}$  contribute to convergence and disparate impact over the course of training. In order to evaluate the full impact of magnitude error and error due to gradient misalignment, we consider the difference in final loss and accuracy between models which have zero magnitude error and zero direction error in Table 1. In these experiments, we consider zero magnitude error to be when  $\|\bar{g}_B\| = \|g_B\|$  for all batches, and zero direction error to be when  $g_B$  and  $\bar{g}_B$  are aligned for all batches. Note that these definitions correspond to comparing update vectors  $g_B$  and  $\frac{\|g_B\|}{\|\bar{g}_B\|}\bar{g}_B$  for the zero magnitude error experiment, and comparing update vectors  $g_B$  and  $\frac{\|\bar{g}_B\|}{\|g_B\|}g_B$  for the zero direction error experiment. These do not correspond to the definitions of  $R_a^{\text{dir}}$  and  $R_a^{\text{mag}}$  in Proposition 2, but capture the intuitive definitions of direction and magnitude error. As described in Section A.2, while  $R_a^{\text{clip}} = R_a^{\text{mag}} + R_a^{\text{dir}}$ , direction error and magnitude error cannot be purely separated with any definition of  $R_a^{\text{mag}}$ ,  $R_a^{\text{dir}}$ .

#### B.5 Baseline methods

We compared our approach DPSGD-Global-Adapt with its predecessor DPSGD-Global, which was designed to improve convergence not fairness, as well as two approaches specifically designed to improve fairness.

DPSGD-Global [6] is presented in Algorithm 3, and involves scaling almost all per-sample gradients by a global factor rather than only scaling large gradients with  $\|g_i\| > C_0$  by a norm-dependent factor. We say “almost all”, because scaling alone does not provide a strict upper bound on the sensitivity, as required for an application of the Gaussian mechanism, see Fig. 2b. The method additionally clips gradients to zero if their norm is above a strict upper bound  $Z$ . Otherwise, the global scaling factor is  $C_0/Z$ , which ensures that the sensitivity, namely  $C_0$ , is finite. The advantage of DPSGD-Global is that it can better preserve the direction of  $\bar{g}_B$ , especially when no gradients are clipped to zero. Hence, [6] advocates for setting  $Z$  larger than  $\|g_i\|$  for any sample in the batch. The drawback of a large  $Z$  is that all gradients are scaled down by a larger factor, so the convergence will be slowed unless the learning rate is increased to compensate. Setting  $Z$  is itself a challenge because we cannot inspect the batch to determine  $\max_i \|g_i\|$  without accounting for that expense in our privacy budget. In Section 5 we described how DPSGD-Global-Adapt resolves these concerns, first by clipping less aggressively, to  $C_0$  instead of 0, while maintaining the same sensitivity, and second by adaptively setting  $Z$  each round according to a private estimate of how many gradients in a batch exceeded  $\gamma Z$  (using the tolerance threshold  $\gamma$ ).

---

#### Algorithm 3 DPSGD-Global

---

**Require:** Iterations  $T$ , Dataset  $D$ , sampling rate  $q$ , clipping bound  $C_0$ , strict clipping bound  $Z \geq C_0$ , noise multiplier  $\sigma$ , learning rates  $\eta_t$

Initialize  $\theta_0$  randomly

**for**  $t$  in  $0, \dots, T - 1$  **do**

$B \leftarrow$  Poisson sample of  $D$  with rate  $q$

**for**  $(x_i, y_i)$  in  $B$  **do**

$g_i \leftarrow \nabla_{\theta_t} \ell(f_{\theta_t}(x_i), y_i)$

$\triangleright$  Compute per-sample gradients

$\gamma_i \leftarrow \begin{cases} \frac{C_0}{Z}, & \|g_i\| \leq Z \\ 0, & \|g_i\| > Z \end{cases}$

$\bar{g}_i \leftarrow \gamma_i g_i$

$\triangleright$  Clip to zero if too large, else scale down by a global factor.

$\bar{g}_B \leftarrow \frac{1}{|B|} (\sum_{i \in B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C_0^2 \mathbb{I}))$

$\theta_{t+1} \leftarrow \theta_t - \eta_t \bar{g}_B$

---

---

**Algorithm 4** DPSGD-F

---

**Require:** Iterations  $T$ , Dataset  $D$ , sampling rate  $q$ , clipping bound  $C_0$ , noise multipliers  $\sigma_1, \sigma_2$ , learning rates  $\eta_t$   
Initialize  $\theta_0$  randomly  
**for**  $t$  in  $0, \dots, T - 1$  **do**  
     $B \leftarrow$  Poisson sample of  $D$  with rate  $q$   
    **for**  $(x_i, a_i, y_i)$  in  $B$  **do**  
         $g_i \leftarrow \nabla_{\theta} \ell(f_{\theta_t}(x_i), y_i)$   $\triangleright$  Compute per-sample gradients  
    **for**  $k$  in  $[K]$  **do**  
         $m^k \leftarrow |\{i : \|g_i^k\| > C_0\}|$   $\triangleright$  Count samples per-group above/below clipping bound  
         $o^k \leftarrow |\{i : \|g_i^k\| \leq C_0\}|$   
         $\{\tilde{m}^k, \tilde{o}^k\}_{k \in [K]} \leftarrow \{m^k, o^k\}_{k \in [K]} + \mathcal{N}(0, \sigma_1^2 \mathbb{I})$   $\triangleright$  Privatize unit sensitivity count vectors  
         $\{\tilde{m}^k, \tilde{o}^k\}_{k \in [K]} \leftarrow \{\max(\lfloor \tilde{m}^k \rfloor, 0), \max(\lfloor \tilde{o}^k \rfloor, 0)\}_{k \in [K]}$   $\triangleright$  Postprocessing  
     $\tilde{m} = \sum_{k \in [K]} \tilde{m}^k$   
    **for**  $k$  in  $[K]$  **do**  
         $\tilde{b}^k = \tilde{m}^k + \tilde{o}^k$   
         $C_k = C_0 \cdot \left(1 + \frac{\tilde{m}^k / \tilde{b}^k}{\tilde{m} / |B|}\right)$   
    **for**  $(x_i, a_i, y_i)$  in  $B$  **do**  
         $\bar{g}_i \leftarrow g_i \cdot \min\left(1, \frac{C_k}{\|g_i\|}\right)$  where  $k = a_i$   $\triangleright$  Clip according to per-group clipping bounds  
     $\tilde{g}_B \leftarrow \frac{1}{|B|} \left(\sum_{i \in B} \bar{g}_i + \mathcal{N}(0, \sigma_2^2 C_0^2 \mathbb{I})\right)$   
     $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}_B$

---

Xu et al. [37] designed DPSGD-F as a method for removing disparate impact caused by DPSGD by adaptively setting the clipping threshold for different protected groups. The method was based on the observation that negatively impacted groups tended to have large gradient norms which were affected more by clipping. Hence, the clipping threshold is raised for groups with larger gradient norms, based on a private estimate of how many gradients per-group have  $\|g_i\| > C_0$ . Given large enough batch sizes, the private estimate can be done with much more noise as compared to the gradient update, so it does not meaningfully increase the privacy budget.

One drawback of this approach is that it requires group label information for every datapoint in the training set. In practice, especially in highly regulated industries, such information may not be permissible to use or even collect. Collecting additional private information from data subjects on protected attributes can itself be a negative process and creates unnecessary privacy risks. One major advantage of DPSGD-Global-Adapt is that it reduces unfairness without ever using group label information.

While each group is clipped using its own threshold, noise is added to the batched gradient based on the sensitivity, determined by the largest group threshold. While all groups receive the same theoretical privacy guarantee in terms of  $(\epsilon, \delta)$ , groups that are clipped to smaller thresholds may enjoy stronger empirical privacy guarantees, as determined for example by adversarial attacks [22, 31]. Hence, it appears likely that DPSGD-F can produce unfairness in the amount of privacy afforded to different groups.

DPSGD-F is shown in Alg. 4. Note that we present the algorithm as implemented in the author’s codebase, not as written in their paper. In our experiments we use the version shown in Alg. 4.

Our final baseline, referred to as “FairLens” was developed in [34] to reduce excessive risk from clipping,  $R_a^{\text{clip}}$ , and adding noise,  $R_a^{\text{noise}}$ . Regularization terms are added to the loss function in DPSGD that specifically target these sources of excessive risk. The source of  $R_a^{\text{noise}}$  was identified to involve the per-group Laplacian of the loss  $\ell$  with respect to model parameters - a second order derivative whose computation scales poorly with model size. To avoid this difficulty, the authors used a stand-in for the Laplacian based on the distance of a point to the decision boundary.

Our implementation is directly based off of code made available by the authors on OpenReview at [openreview.net/forum?id=7EFdodSWee4](https://openreview.net/forum?id=7EFdodSWee4). The version implemented in their code is shown in Alg.

5, and assumes there are only two mutually protected groups, denoted  $a$  and  $b$ . Hence, it is not applicable to the MNIST dataset.

---

**Algorithm 5** FairLens

---

**Require:** Iterations  $T$ , Dataset  $D$ , sampling rate  $q$ , clipping bound  $C_0$ , noise multiplier  $\sigma$ , learning rates  $\eta_t$ , regularization weights  $\gamma_1, \gamma_2$

```

Initialize  $\theta_0$  randomly
for  $t$  in  $0, \dots, T-1$  do
     $B \leftarrow$  Poisson sample of  $D$  with rate  $q$ 
    for  $(x_i, a_i, y_i)$  in  $B$  do
         $g_i \leftarrow \nabla_{\theta} \ell(f_{\theta_t}(x_i), y_i)$   $\triangleright$  Compute per-sample gradients of original loss
         $\bar{g}_i \leftarrow g_i \cdot \min\left(1, \frac{C_0}{\|g_i\|}\right)$ 
     $g_B \leftarrow \frac{1}{|B|} \sum_{i \in B} g_i$ 
     $\bar{g}_B \leftarrow \frac{1}{|B|} \sum_{i \in B} \bar{g}_i$ 
    for  $k$  in  $\{a, b\}$  do
         $g_{B_k} \leftarrow \frac{1}{|B_k|} \sum_{i \in B, a_i=k} g_i$ 
         $f_k \leftarrow \frac{1}{|B_k|} \sum_{i \in B, a_i=k} f_{\theta_t}(x_i)$ 
         $R_1 = |\langle g_{B_a} - g_{B_b}, \bar{g}_B - g_B \rangle|$ 
         $R_2 = \frac{1}{2}(f_a \cdot (1 - f_a) + f_b \cdot (1 - f_b))$ 
         $\mathcal{L} = \ell(f_{\theta_t}(x_i), y_i) + \gamma_1 R_1 + \gamma_2 R_2$   $\triangleright$  Define regularized loss
        for  $(x_i, a_i, y_i)$  in  $B$  do
             $g'_i \leftarrow \nabla_{\theta} \mathcal{L}(f_{\theta_t}(x_i), y_i)$   $\triangleright$  Compute per-sample gradients of regularized loss
             $\bar{g}'_i \leftarrow g'_i \cdot \min\left(1, \frac{C_0}{\|g'_i\|}\right)$   $\triangleright$  Clip to ensure finite sensitivity
         $\tilde{g}'_B \leftarrow \frac{1}{|B|} (\sum_{i \in B} \bar{g}'_i + \mathcal{N}(0, \sigma^2 C_0^2 \mathbb{I}))$ 
     $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{g}'_B$ 

```

---

## B.6 Additional results

In this section we complete the set of experimental results shown in Sec. 6 over all datasets and methods. All results are averaged over five random seeds with one standard error shown.

Table 3: Performance and Fairness metrics for Adult dataset

METHOD	ACC M	ACC F	$\pi(M)$	$\pi(F)$	PCG	Loss M	Loss F	ER M	ER F	ERG
NON PRIVATE	80.5 $\pm$ 0.8	92.2 $\pm$ 0.2	-	-	-	0.40 $\pm$ 0.00	0.19 $\pm$ 0.00	-	-	-
DPSGD	69.9 $\pm$ 0.8	88.5 $\pm$ 0.1	10.6 $\pm$ 0.7	3.6 $\pm$ 0.2	6.9 $\pm$ 0.7	0.78 $\pm$ 0.01	0.40 $\pm$ 0.01	0.39 $\pm$ 0.01	0.21 $\pm$ 0.02	0.17 $\pm$ 0.02
FAIRLENS	68.8 $\pm$ 0.9	88.5 $\pm$ 0.1	11.7 $\pm$ 0.5	3.7 $\pm$ 0.2	7.9 $\pm$ 0.7	0.57 $\pm$ 0.00	0.42 $\pm$ 0.00	0.18 $\pm$ 0.00	0.23 $\pm$ 0.01	0.05 $\pm$ 0.01
DPSGD-F	78.0 $\pm$ 1.3	89.4 $\pm$ 0.2	2.5 $\pm$ 0.6	2.7 $\pm$ 0.3	0.2 $\pm$ 0.6	0.49 $\pm$ 0.01	0.31 $\pm$ 0.01	0.09 $\pm$ 0.00	0.12 $\pm$ 0.01	0.02 $\pm$ 0.01
DPSGD-G.	78.5 $\pm$ 1.1	89.9 $\pm$ 0.2	2.0 $\pm$ 0.3	2.2 $\pm$ 0.1	0.2 $\pm$ 0.4	0.43 $\pm$ 0.00	0.25 $\pm$ 0.01	0.04 $\pm$ 0.00	0.05 $\pm$ 0.01	0.02 $\pm$ 0.01
DPSGD-G.-A.	80.7 $\pm$ 0.8	92.3 $\pm$ 0.2	-0.1 $\pm$ 0.2	-0.1 $\pm$ 0.1	0.0 $\pm$ 0.0	0.39 $\pm$ 0.01	0.18 $\pm$ 0.01	0.00 $\pm$ 0.01	0.00 $\pm$ 0.01	0.00 $\pm$ 0.01

Table 4: Performance and Fairness metrics for Dutch dataset

METHOD	ACC M	ACC F	$\pi(M)$	$\pi(F)$	PCG	Loss M	Loss F	ER M	ER F	ERG
NON PRIVATE	80.4 $\pm$ 0.5	86.7 $\pm$ 0.6	-	-	-	0.41 $\pm$ 0.00	0.35 $\pm$ 0.00	-	-	-
DPSGD	74.3 $\pm$ 0.6	63.2 $\pm$ 2.9	6.1 $\pm$ 0.4	23.5 $\pm$ 2.6	17.3 $\pm$ 2.3	0.58 $\pm$ 0.01	0.64 $\pm$ 0.03	0.17 $\pm$ 0.01	0.28 $\pm$ 0.03	0.12 $\pm$ 0.03
FAIRLENS	73.1 $\pm$ 0.6	66.2 $\pm$ 4.2	7.3 $\pm$ 0.8	20.4 $\pm$ 3.9	13.1 $\pm$ 3.7	0.61 $\pm$ 0.01	0.64 $\pm$ 0.01	0.20 $\pm$ 0.01	0.29 $\pm$ 0.01	0.09 $\pm$ 0.01
DPSGD-F	77.9 $\pm$ 0.5	85.4 $\pm$ 0.6	2.5 $\pm$ 0.6	1.3 $\pm$ 0.4	1.2 $\pm$ 0.5	0.51 $\pm$ 0.01	0.42 $\pm$ 0.01	0.09 $\pm$ 0.01	0.06 $\pm$ 0.01	0.03 $\pm$ 0.01
DPSGD-G.	80.0 $\pm$ 0.6	86.6 $\pm$ 0.8	0.4 $\pm$ 0.4	0.1 $\pm$ 0.4	0.3 $\pm$ 0.5	0.42 $\pm$ 0.00	0.36 $\pm$ 0.01	0.00 $\pm$ 0.00	0.01 $\pm$ 0.01	0.00 $\pm$ 0.01
DPSGD-G.-A.	80.4 $\pm$ 0.5	86.9 $\pm$ 0.4	0.0 $\pm$ 0.3	0.0 $\pm$ 0.3	0.0 $\pm$ 0.4	0.42 $\pm$ 0.00	0.35 $\pm$ 0.01	0.00 $\pm$ 0.01	-0.01 $\pm$ 0.00	0.01 $\pm$ 0.01

First we look at the final performance and fairness metrics on the test set for Adult in Tab. 3 and Dutch in Tab. 4 (cf. MNIST in Tab. 2). We see that FairLens reduces the excessive risk gap as its loss function was designed to do, but is inconsistent in reducing the privacy cost gap. DPSGD-F reduces both fairness metrics while achieving better performance. DPSGD-Global improves over DPSGD-F in all metrics, and does so without requiring access to protected group membership information.

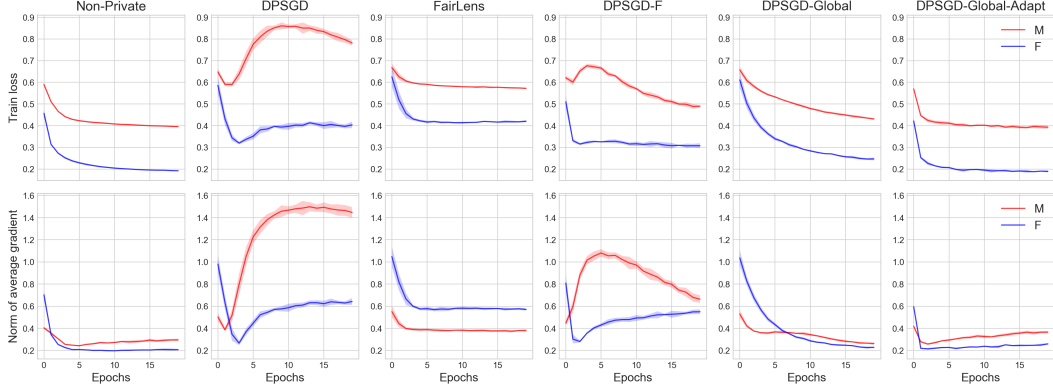


Figure 7: Adult dataset: (a) Train loss per epoch (b)  $\|g_B\|$ , averaged over batches per epoch

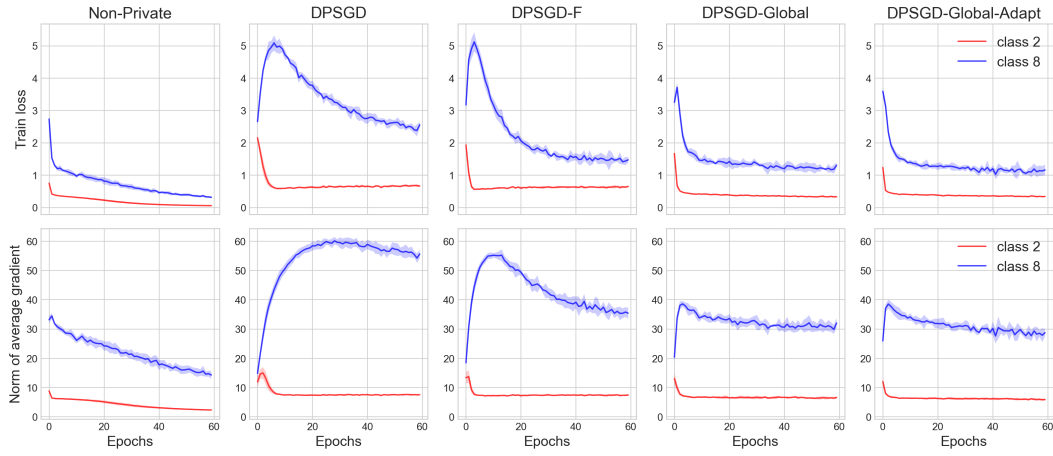


Figure 8: MNIST dataset (a) Train loss per epoch (b)  $\|g_B\|$ , averaged over batches per epoch

Our method DPSGD-Global-Adapt further improves both performance and fairness by clipping less aggressively and adaptively setting the upper clipping threshold  $Z$ .

To go along with the training curves shown for Dutch in Fig. 3, we present the same for Adult in Fig. 7, and MNIST in Fig. 8. The trends appear to be consistent across datasets - whereas DPSGD produces large values and a large gap for the gradient norms and losses between protected groups, our method DPSGD-Global-Adapt reduces the values and gap at all stages of training.

We also present the values of terms  $R_a^{\text{dir}}$  and  $R_a^{\text{mag}}$  over training for Dutch in Fig. 9, and for MNIST in Fig. 10 as for Adult in Fig. 4. Both Global methods dramatically reduce  $R_a^{\text{dir}}$  compared to DPSGD at the cost of larger  $R_a^{\text{mag}}$ . Comparing to the final training results where Global methods also show the best performance, this provides further evidence for our hypothesis that gradient misalignment is the most significant cause of disparate impact in DPSGD.

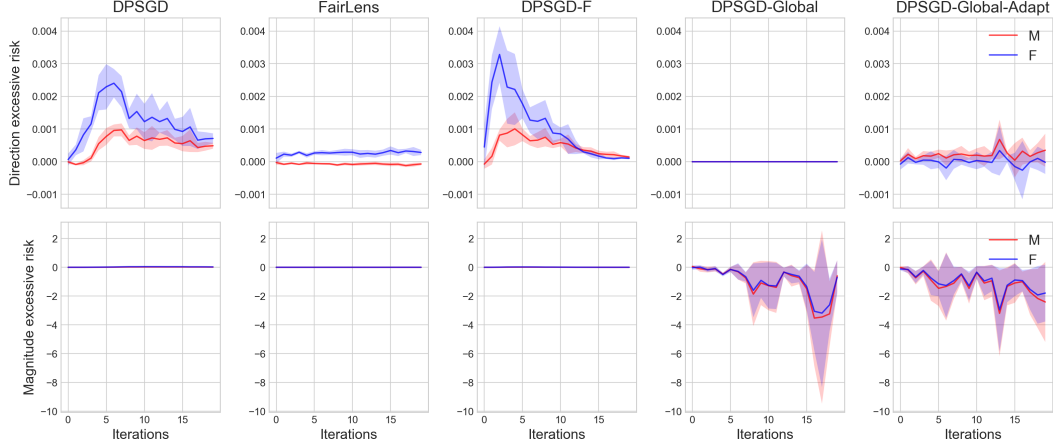


Figure 9: (a) Excessive risk due to gradient misalignment per group for the Dutch dataset (b) Excessive risk due to magnitude error per group for the Dutch dataset

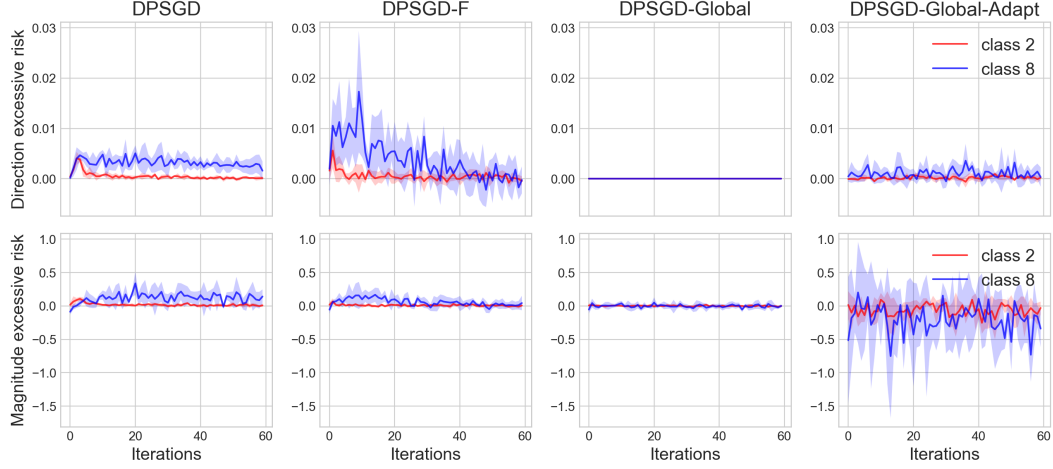


Figure 10: (a) Excessive risk due to gradient misalignment per group for the MNIST dataset (b) Excessive risk due to magnitude error per group for the MNIST dataset

We note that by tuning the learning rate in DPSGD-Global and DPSGD-Global-Adapt, there is a trade-off between magnitude error and noise error. Referring to Proposition 1,  $R_a^{\text{noise}} = \frac{\eta_t^2}{2} \text{Tr}(H_\ell^a) C_0^2 \sigma^2$ , we see that the excessive risk due to noise is affected by the learning rate  $\eta$ , the noise multiplier  $\sigma$ , clipping bound  $C_0$  and the trace of the Hessian for group  $a$ . In choosing a larger learning rate for the global methods to offset the magnitude error, we increase the noise error quadratically. Refer to values of  $R_k^{\text{noise}}$  over training for MNIST in Fig. 11, Adult in Fig. 12, and Dutch in Fig. 13. While the excessive risk due to noise is significantly larger for the global methods, these methods outperform all local private methods at the end of training, see Tables 2, 3, 4. This further validates that direction error is the core cause of disparate impact, and minimizing gradient misalignment should be prioritized over other sources of unfairness.

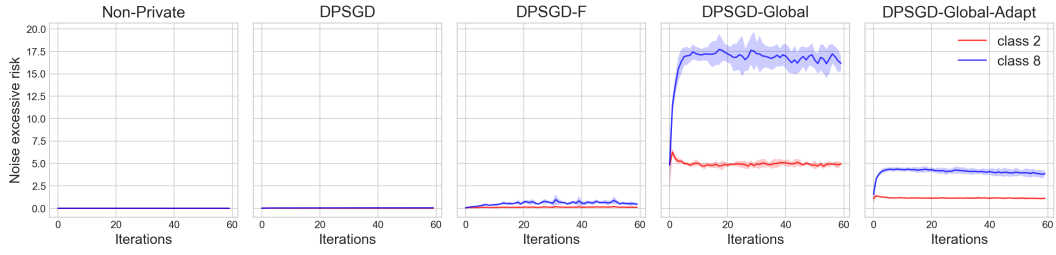


Figure 11: Excessive risk due to noise error per group for the MNIST dataset

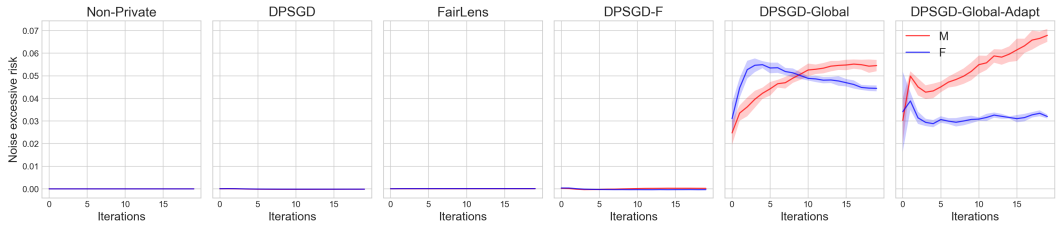


Figure 12: Excessive risk due to noise error per group for the Adult dataset

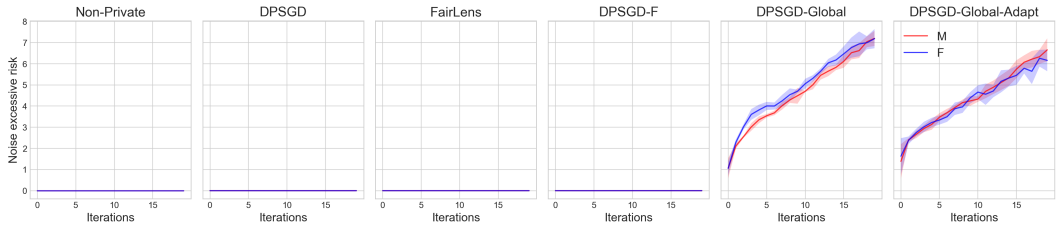


Figure 13: Excessive risk due to noise error per group for the Dutch dataset