Individual Privacy Accounting for Differentially Private Stochastic Gradient Descent

Da Yu[†] Gautam Kamath[‡*] Janardhan Kulkarni[§*]

Tie-Yan Liu[§*] Jian Yin[†*] Huishuai Zhang[§*]

August 1, 2022

Abstract

Differentially private stochastic gradient descent (DP-SGD) is the workhorse algorithm for recent advances in private deep learning. It provides a single privacy guarantee to all datapoints in the dataset. We propose an efficient algorithm to compute privacy guarantees for individual examples when releasing models trained by DP-SGD. We use our algorithm to investigate individual privacy parameters across a number of datasets. We find that most examples enjoy stronger privacy guarantees than the worst-case bound. We further discover that the training loss and the privacy parameter of an example are well-correlated. This implies groups that are underserved in terms of model utility are simultaneously underserved in terms of privacy guarantee. For example, on CIFAR-10, the average ε of the class with the lowest test accuracy is 26.3% higher than that of the class with the highest accuracy. We also run membership inference attacks to show this reflects disparate empirical privacy risks.

1 Introduction

Differential privacy is a strong notion of data privacy, enabling rich forms of privacy-preserving data analysis [DMNS06, DR14a]. Informally speaking, it quantitatively bounds the maximum influence of any datapoint using a privacy parameter ε, where a small value of ε corresponds to stronger privacy guarantees. Training deep models with differential privacy is an active research area [ACG+16, PAE+17, BDLS20, YNB+22, AGG+21, LTLH22, GAW+22, MTKC22, DBH+22]. Models trained with differential privacy not only provide theoretical privacy guarantee to their data but also are more robust against empirical attacks [BGRK19, CLE+19, JUO20, NST+21].

Differentially private stochastic gradient descent (DP-SGD) is the de-facto choice for differentially private deep learning [SCS13, BST14, ACG+16]. DP-SGD first clips individual gradients and then adds Gaussian noise to the aggregated gradient. Standard privacy accounting takes a worst-case approach, and provides all examples with the same privacy parameter ε. However, from the perspective of machine learning, different examples can have very different impacts on a learning algorithm [KL17, FZ20]. For example, consider support vector machines: removing a non-support vector has no effect on the resulting model, and hence that example would have perfect privacy.

In this paper, we give an efficient algorithm to accurately estimate individual privacy parameters of models trained by DP-SGD. Our privacy guarantee takes the form of *ex-post* differential privacy that adapts to algorithms's outcomes, e.g., a training trajectory of DP-SGD, to provide a precise characterization of individual privacy costs [LNR+17, RW21]. More details about ex-post differential privacy are in Section 2.1. Inspecting individual privacy parameters allows us to better understand example-wise impacts. It turns out that, for common benchmarks, many examples experience much stronger privacy guarantee than the worst-case bound. To illustrate this, we plot the individual privacy parameters of MNIST [LBBH98], CIFAR-10 [Kri09], and UTKFace [ZSQ17] in Figure 1. Experimental details, as well as more

[†]Sun Yat-sen University. {yuda3@mail2,issjyin@mail}.sysu.edu.cn

[‡]Cheriton School of Computer Science, University of Waterloo. Supported by an NSERC Discovery Grant, an unrestricted gift from Google, and a University of Waterloo startup grant. g@csail.mit.edu

[§]Microsoft Research. {jakul,tyliu,huzhang}@microsoft.com

[*]Authors are listed in alphabetical order.

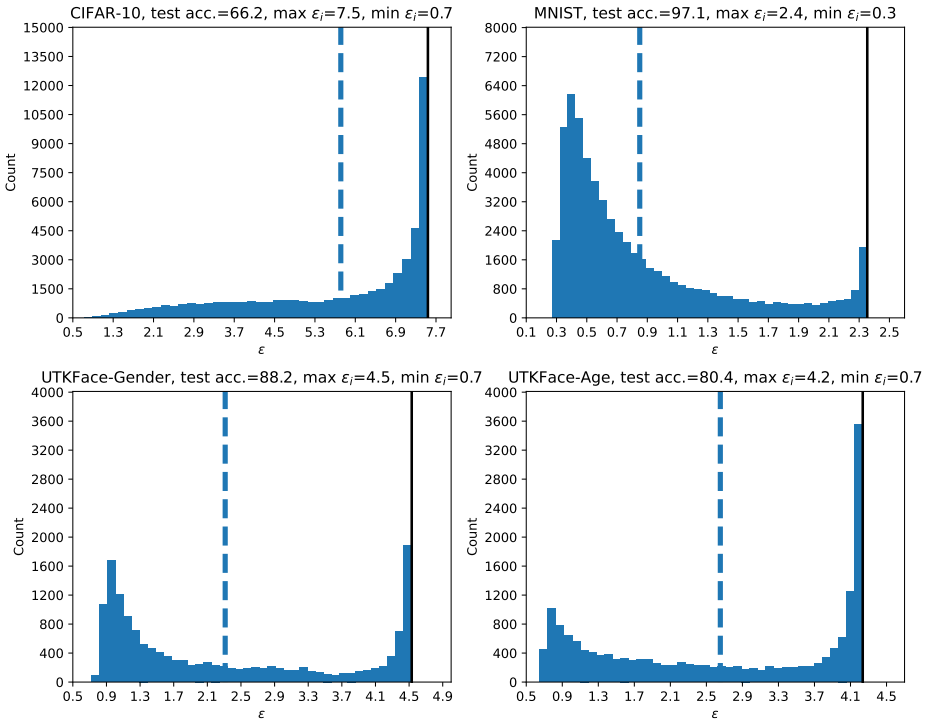

Figure 1: Individual privacy parameters of models trained by DP-SGD. The value of δ is 1×10^{-5}. The dashed line indicates the average of ε values. The black solid line indicates the privacy parameter of the classic analysis.

results, can be found in Section 4 and 5. The disparity in individual privacy guarantees naturally arise when running DP-SGD. To the best of our knowledge, our investigation is the first to explicitly reveal such disparity.

We propose two techniques to make individual privacy accounting viable for DP-SGD. First, we maintain estimates of the gradient norms for all examples so the individual privacy costs can be computed accurately at every update. Second, we round the gradient norms with a small precision r to control the number of different privacy costs, which need to be computed numerically. We explain why these two techniques are necessary in Section 2. More details of the proposed algorithm, as well as methods to release individual privacy parameters without additional privacy cost, are in Section 3.

We further demonstrate a strong correlation between individual privacy parameters and individual loss values. We find that the individual privacy parameters increase logarithmically with the final training losses. This suggests that the same examples suffer a simultaneous unfairness in terms of worse privacy and worse utility. While prior works have shown that underrepresented groups experience worse utility [BG18], and that these disparities are amplified when models are trained privately [BPS19, SPGG21, HNS+22, NHS22], we are the first to show that the privacy guarantee *and* utility are negatively impacted concurrently. This is in comparison to prior work in the differentially private setting which took a worst-case perspective for privacy accounting, resulting in a uniform privacy guarantee for all training examples. For instance, when running gender classification on the UTKFace dataset, the average ε of the group with the lowest test accuracy (Asian) is 25% higher than that of the group with the highest test accuracy (Indian). We also run membership inference attacks on those datasets and show the privacy parameters correlate well with the attack success rates.

1.1 Related Work

Several works have explored example-wise privacy guarantees in differentially private learning. [JYC15] propose *personalized differential privacy* that provides pre-specified individual privacy parameters which are independent of the learning algorithm, e.g., users can choose different levels of privacy guarantees based on their sensitivities to privacy leakage [MB22]. A recent line of works also uses the variation in example-wise sensitivities that naturally arise in learning to study example-wise privacy. *Per-instance differential privacy* captures the privacy parameter of a target example with respect to a fixed training set [Wan19, RW21, GAW+22]. [FZ21] design an individual *Rényi differential*

privacy filter and apply it on DP-GD to allow the algorithm to run more steps on some examples. In this work, we give the first algorithm to compute individual privacy parameters for DP-SGD. We also discover several new and interesting patterns in individual privacy parameters.

A recent line of work has found that some examples are more vulnerable to empirical attacks [SSSS17, LBG17, KYC+19, YMMS21, CCTCP21, CCN+21]. They show membership inference attacks have significantly higher success rates on certain examples, e.g., outlier datapoints. In this work, we show the disparity of individual privacy also exists theoretically when learning with DP-SGD. Moreover, we also show the disparity in privacy correlates well with the disparity in utility.

2 Preliminaries

We first give some background on differential privacy. We then briefly introduce the privacy analysis of DP-SGD and highlight the challenges in computing individual privacy guarantees. Finally, we argue that providing the same privacy bound to all examples is not ideal because of the variation in individual gradient norms.

2.1 Background on Differentially Private Learning

Differential privacy builds on the notion of neighboring datasets. A dataset $\mathbb{D} = \{\boldsymbol{d}_i\}_{i=1}^n$ is a neighboring dataset of $\mathbb{D}'$ (denoted as $\mathbb{D} \sim \mathbb{D}'$) if $\mathbb{D}'$ can be obtained by adding/removing one example from $\mathbb{D}$. The indiviudal privacy guarantees in this work take the form of (ε, δ)-differential privacy.

Definition 1. *[Individual (ε, δ)-DP] For a datapoint $\mathbf{d}$, let $\mathbb{D}$ be an arbitrary dataset and $\mathbb{D}' = \mathbb{D} \cup \{\boldsymbol{d}\}$ be its neighboring dataset. An algorithm $\mathcal{A}$ satisfies $(\varepsilon(\boldsymbol{d}), \delta)$-individual DP if for any subset of outputs $\mathbb{S} \subset Range(\mathcal{A})$ it holds that*

$$\Pr[\mathcal{A}(\mathbb{D}) \in \mathbb{S}] \leq e^{\varepsilon(\boldsymbol{d})} \Pr[\mathcal{A}(\mathbb{D}') \in \mathbb{S}] + \delta \text{ and } \Pr[\mathcal{A}(\mathbb{D}') \in \mathbb{S}] \leq e^{\varepsilon(\boldsymbol{d})} \Pr[\mathcal{A}(\mathbb{D}) \in \mathbb{S}] + \delta.$$

We further allow the privacy parameter ε to be a function of the subset of outcomes $\mathbb{S}$ to provide a sharper characterization of privacy. This type of variants are known as *ex-post* differential privacy [LNR+17, RW21]. One difference between our definition and the ex-post DP in [LNR+17, RW21] is that we define ex-post DP over a subset of outcomes while their definition is for a single outcome.

Definition 2. *[Ex-post individual (ε, δ)-DP] Fix a datapoint $\mathbf{d}$ and a set of outcomes $\mathbb{A} \subset Range(\mathcal{A})$, let $\mathbb{D}$ be an arbitrary dataset and $\mathbb{D}' = \mathbb{D} \cup \{\boldsymbol{d}\}$. An algorithm $\mathcal{A}$ satisfies $(\varepsilon(\mathbb{A}, \boldsymbol{d}), \delta)$-ex-post individual DP for $\mathbf{d}$ at $\mathbb{A}$ if for any $\mathbb{S} \subset \mathbb{A}$ it holds that*

$$\Pr[\mathcal{A}(\mathbb{D}) \in \mathbb{S}] \leq e^{\varepsilon(\mathbb{A}, \boldsymbol{d})} \Pr[\mathcal{A}(\mathbb{D}') \in \mathbb{S}] + \delta \text{ and } \Pr[\mathcal{A}(\mathbb{D}') \in \mathbb{S}] \leq e^{\varepsilon(\mathbb{A}, \boldsymbol{d})} \Pr[\mathcal{A}(\mathbb{D}) \in \mathbb{S}] + \delta.$$

Definition 2 has the same semantics as Definition 1 once the algorithm's outcome belongs to $\mathbb{A}$ is known. It is a strict generalization of (ε, δ)-DP as one can recover (ε, δ)-DP by maximizing $\varepsilon(\mathbb{A}, \boldsymbol{d})$ over $\mathbb{A}$ and $\boldsymbol{d}$. Making this generalization is crucial for us to precisely account the privacy guarantees of DP-SGD for specific trained models, i.e., some outcomes of a differentially private learning algorithm. The privacy risk of a training example highly depends on the trained model. For example, [TSJ+22] show one can adversarially modify the training data and hence change the trained models to maximize the privacy risk of a target example. In practice, people usually care more about the privacy of a target example at some instantiated models rather than at all possible models. Therefore, it is reasonable to provide fine-grained privacy guarantees as a complement to the worst-case bound.

A common approach for doing deep learning with differential privacy is to make each update differentially private instead of protecting the trained model directly. The composition property of differential privacy allows us to reason about the overall privacy of running several such steps. We give a simple example to illustrate how to privatize each update. Suppose we take the sum of all gradients $\boldsymbol{v} = \sum_{i=1}^n \boldsymbol{g}_i$ from dataset $\mathbb{D}$. Without loss of generality, further assume we add an arbitrary example $\boldsymbol{d}'$ to obtain a neighboring dataset $\mathbb{D}'$. The summed gradient becomes $\boldsymbol{v}' = \boldsymbol{v} + \boldsymbol{g}'$, where $\boldsymbol{g}'$ is the gradient of $\boldsymbol{d}'$. If we add independent Gaussian noise with variance σ^2 to each coordinate, then the output distributions of two neighboring datasets are

$$\mathcal{A}(\mathbb{D}) \sim \mathcal{N}(\boldsymbol{v}, \sigma^2 \boldsymbol{I}) \text{ and } \mathcal{A}(\mathbb{D}') \sim \mathcal{N}(\boldsymbol{v}', \sigma^2 \boldsymbol{I}).$$

We then can bound the difference between two Gaussian distributions to provide (ε, δ)-DP. The expectations of $\mathcal{A}(\mathbb{D})$ and $\mathcal{A}(\mathbb{D}')$ only differ by $\boldsymbol{g}'$. A larger gradient leads to a larger difference and hence worse privacy parameters. This approach is previously known as *Gaussian Mechanism* [DR+14b].

2.2 Challenges of Computing Individual Privacy Parameters for DP-SGD

Privacy accounting in DP-SGD is more complex than the simple example in Section 2.1 because the analysis involves *privacy amplification by subsampling* [ACG+16, BBG18, MTZ19, ZW19, WBK19]. Roughly speaking, randomly sampling a minibatch in DP-SGD strengthens the privacy guarantees since most points in the dataset are not involved in a single step. This complication makes direct computation of individual privacy parameters impractical.

Before we expand on these difficulties, we first describe the output distributions of neighboring datasets in DP-SGD [ACG+16]. Poisson sampling is assumed, i.e., each example is sampled independently with probability p. Let $\boldsymbol{v} = \sum_{i \in \mathbb{M}} \boldsymbol{g}_i$ be the sum of the minibatch of gradients of $\mathbb{D}$, where $\mathbb{M}$ is the set of sampled indices. Consider also a neighboring dataset $\mathbb{D}'$ that has one datapoint with gradient $\boldsymbol{g}'$ added. Because of Poisson sampling, the output is exactly $\boldsymbol{v}$ with probability $1-p$ ($\boldsymbol{g}'$ is not sampled) and is $\boldsymbol{v}' = \boldsymbol{v} + \boldsymbol{g}'$ with probability p ($\boldsymbol{g}'$ is sampled). Suppose we still add isotropic Gaussian noise, the output distributions of two neighboring datasets are

$$\mathcal{A}(\mathbb{D}) \sim \mathcal{N}(\boldsymbol{v}, \sigma^2 \boldsymbol{I}), \tag{1}$$

$$\mathcal{A}(\mathbb{D}') \sim \mathcal{N}(\boldsymbol{v}, \sigma^2 \boldsymbol{I}) \text{ with prob. } 1-p, \ \mathcal{A}(\mathbb{D}') \sim \mathcal{N}(\boldsymbol{v}', \sigma^2 \boldsymbol{I}) \text{ with prob. } p. \tag{2}$$

With Equation (1) and (2), we explain the challenges in computing individual privacy parameters.

2.2.1 Full Batch Gradient Norms Are Required at Every Iteration

There is some privacy cost for $\boldsymbol{d}'$ *even if it is not sampled in the current iteration* because the analysis makes use of the subsampling process. For a given sampling probability and noise variance, the amount of privacy cost is determined by $\|\boldsymbol{g}'\|$. Therefore, we need accurate gradient norms of all examples to compute accurate privacy costs at every iteration. However, when running SGD, we only compute minibatch gradients. Previous analysis of DP-SGD evades this problem by simply assuming all examples have the maximum possible norm, i.e., the clipping threshold.

2.2.2 Computational Cost of Individual Privacy Parameters is Huge

The density function of $\mathcal{A}(\mathbb{D}')$ is a mixture of two Gaussian distributions. This makes computing the Rényi divergence between $\mathcal{A}(\mathbb{D})$ and $\mathcal{A}(\mathbb{D}')$ harder as there is no closed form solution. Although there are some asymptotic bounds, those bounds are looser than computing the divergence numerically, and thus such numerical computations are necessary [ACG+16, WBK19, MTZ19, GLW21]. In the classic analysis, there is only one numerical computation as all examples have the same privacy cost over all iterations. However, naive computation of individual privacy parameters would require up to $n \times T$ computations, where n is the dataset size and T is the number of iterations.

2.3 An Observation: Gradient Norms in Deep Learning Vary Significantly

We show gradient norms vary significantly to demonstrate that different examples experience very difference privacy costs when training with DP-SGD. We train the standard ResNet-20 model in [HZRS16] on CIFAR-10. The maximum clipping threshold is the median of gradient norms at initialization. More implementation details are in Section 4. We first sort all examples based on their average gradient norms across training. Then we divide them into five equally sized groups based on the quantiles. We plot the average norms of different groups across training in Figure 2.

The gradient norms of different groups in Figure 2 show significant stratification. Such stratification naturally leads to different individual privacy costs. This suggests that quantifying individual privacy parameters may be valuable despite the aforementioned challenges.

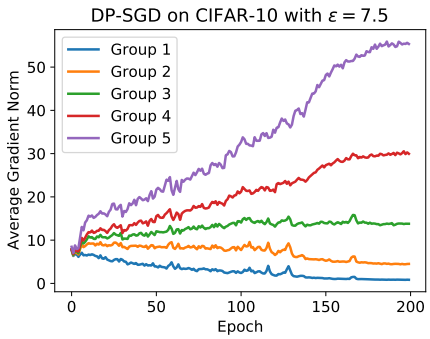

Figure 2: Gradient norms of different groups.

3 Deep Learning with Individual Privacy

We give an efficient algorithm (Algorithm 1) to estimate individual
privacy parameters for DP-SGD. The privacy analysis of Algorithm 1 is in Section 3.1. We perform two modifications to make individual privacy accounting feasible with small computational overhead. First, we compute full batch gradient norms once in a while, e.g., at the beginning of every epoch, and use the results to estimate the gradient norms for subsequent iterations. We show the estimated privacy parameters are accurate in Section 3.2. We note that the estimated privacy parameters themselves are strict differential privacy guarantees because we use the estimated norms to clip individual gradients.[1] Additionally, we round the gradient norms to a given precision so the number of numerical computations is independent of the dataset size and number of epochs. More details of this modification are in Section 3.3. Finally, we discuss how to make use of the individual privacy parameters in Section 3.4.

Algorithm 1: Deep Learning with Individual Privacy Accounting

Input : Maximum clipping threshold C, rounding precision r, noise variance σ^2, sampling probability p, frequency of updating full gradient norms at every epoch K.

Let $\{C^{(i)}\}_{i=1}^n$ be estimated gradient norms of all examples and initialize $C^{(i)} = C$.
Let $\mathbb{C} = \{r, 2r, 3r, \ldots, C\}$ be all possible norms under rounding.
Let $\{\boldsymbol{o}^{(i)} = \mathbf{0}\}_{i=1}^n$ be the accumulated Rényi divergences at different orders.
for $c \in \mathbb{C}$ **do**
    //Formulations of $\mathcal{A}(\mathbb{D})$ and $\mathcal{A}(\mathbb{D}')$ are in Equation (1) and (2).
    Compute Rényi divergences $\boldsymbol{\rho}_c$ between $\mathcal{A}(\mathbb{D})$ and $\mathcal{A}(\mathbb{D}')$ numerically with c, p, and σ^2.
end
for $e = 1$ *to* E **do**
    for $t' = 1$ *to* T **do**
        //T is the number of iterations per epoch, E is the number of epochs.
        $t = t' + T \times (e-1)$
        if $t \bmod \lfloor T/K \rfloor = 0$ **then**
            Compute full batch gradient norms and update $\{C^{(i)}\}_{i=1}^n$ with rounded norms.
        end
        Sample a minibatch of gradients $\{\boldsymbol{g}^{(I_j)}\}_{j=1}^m$ and clip them $\bar{\boldsymbol{g}}^{(I_j)} = clip(\boldsymbol{g}^{(I_j)}, C^{(I_j)})$.
        Update model $\theta_t = \theta_{t-1} - \eta(\sum \bar{\boldsymbol{g}}^{(I_j)} + z)$, where $z \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$.
        //Update privacy costs for the whole dataset.
        for $i = 1$ *to* n **do**
            Find corresponding $c \in \mathbb{C}$ for the i_{th} example and set $\boldsymbol{\rho}_t^{(i)} = \boldsymbol{\rho}_c$.
            $\boldsymbol{o}^{(i)} = \boldsymbol{o}^{(i)} + \boldsymbol{\rho}_t^{(i)}$.
        end
    end
end

3.1 Privacy Analysis of Algorithm 1

In Algorithm 1, we compose the privacy costs at different steps through Rényi differential privacy (RDP) [Mir17]. RDP uses the Rényi divergence at different orders to measure privacy. We use $D_\alpha(\mu||\nu) = \frac{1}{\alpha-1}\log\int(\frac{d\mu}{d\nu})^\alpha d\nu$ to denote the Rényi divergence at order α between μ and ν and $D_\alpha^\leftrightarrow(\mu||\nu) = \max(D_\alpha(\mu||\nu), D_\alpha(\nu||\mu))$ to denote the maximum divergence of two directions. The definition of Rényi differential privacy for a fixed datapoint is as follows.

Definition 3. *[Individual RDP [FZ21]] Fix a datapoint $\mathbf{d}$, let $\mathbb{D}$ be an arbitrary dataset and $\mathbb{D}' = \mathbb{D} \cup \{\boldsymbol{d}\}$. A randomized algorithm $\mathcal{A}$ satisfies $(\alpha, \rho(\boldsymbol{d}))$-individual RDP for $\mathbf{d}$ if it holds that $D_\alpha^\leftrightarrow(\mathcal{A}(\mathbb{D})||\mathcal{A}(\mathbb{D}')) \leq \rho(\boldsymbol{d})$.*

The RDP privacy parameters in Algorithm 1 themselves are random variables depending on previous outputs, i.e., different intermediate models lead to different gradients and hence different privacy parameters. This is different from

[1] Using individual clipping thresholds could lose more gradient signal if the estimates are inaccurate. In Appendix B, we show the individual clipping in Algorithm 1 does not affect accuracy.

the classic analysis where the privacy parameters are fixed before training. In this paper, we focus on the realizations of those random variables to precisely account the privacy cost of a realized run of Algorithm 1.

We define a sequence of randomized algorithms $\hat{\mathcal{A}}^{(t)}(\theta_1, \ldots, \theta_{t-1}, \mathbb{D}) = (\mathcal{A}_1(\mathbb{D}), \mathcal{A}_2(\theta_1, \mathbb{D}), \ldots, \mathcal{A}_t(\theta_1, \ldots, \theta_{t-1}, \mathbb{D}))$ where $(\theta_1, \ldots, \theta_{t-1})$ are some fixed outcomes from the domain of $(\mathcal{A}_1, \ldots, \mathcal{A}_{t-1})$. Noting that $\hat{\mathcal{A}}^{(t)}$ is not adaptive as the input of each individual mechanism in $\hat{\mathcal{A}}^{(t)}$ does not depend on the outputs of previous mechanisms. Further let $\mathcal{A}^{(t)}(\mathbb{D}) = (\mathcal{A}_1(\mathbb{D}), \mathcal{A}_2(\mathcal{A}_1(\mathbb{D}), \mathbb{D}), \ldots, \mathcal{A}_t(\mathcal{A}_1(\mathbb{D}), \ldots, \mathbb{D}))$ be the adaptive composition of the randomized algorithms. Below we show a RDP bound on $\hat{\mathcal{A}}^{(t)}$ gives an ex-post DP bound on $\mathcal{A}^{(t)}$. We comment that the RDP parameters of the mechanisms in $\mathcal{A}^{(t)}$ are random variables that require additional care when being composed [FZ21, Léc21, WRRW22].

Theorem 3.1 (RDP of $\hat{\mathcal{A}}^{(t)}$ implies ex-post DP of $\mathcal{A}^{(t)}$)**.** *Let $\mathbb{A} = (\theta_1, \ldots, \theta_{t-1}, Range(\mathcal{A}_t)) \subset Range(\mathcal{A}^{(t)}) = Range(\hat{\mathcal{A}}^{(t)})$ where $\theta_1, \ldots, \theta_{t-1}$ are some arbitrary fixed outcomes. If $\hat{\mathcal{A}}^{(t)}(\cdot)$ satisfies o_α RDP at order α, then $\mathcal{A}^{(t)}(\mathbb{D})$ satisfies $(o_\alpha + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-ex-post differential privacy at $\mathbb{A}$.*

Proof. For any $\mathbb{D}$ and a given outcome $\theta^{(t)} = (\theta_1, \theta_2, \ldots, \theta_{t-1}, \theta_t) \in \mathbb{A}$, we have

$$\mathbb{P}\left[\mathcal{A}^{(t)}(\mathbb{D}) = \theta^{(t)}\right] = \mathbb{P}\left[\mathcal{A}^{(t-1)}(\mathbb{D}) = \theta^{(t-1)}\right] \mathbb{P}\left[\mathcal{A}_t(\mathcal{A}_1(\mathbb{D}), \ldots, \mathbb{D}) = \theta_t | \mathcal{A}^{(t-1)}(\mathbb{D}) = \theta^{(t-1)}\right], \tag{3}$$

$$= \mathbb{P}\left[\mathcal{A}^{(t-1)}(\mathbb{D}) = \theta^{(t-1)}\right] \mathbb{P}\left[\mathcal{A}_t(\theta_1, \ldots, \theta_{t-1}, \mathbb{D}) = \theta_t\right], \tag{4}$$

by the product rule of conditional probability. Apply the product rule recurrently on $\mathbb{P}\left[\mathcal{A}^{(t-1)}(\mathbb{D}) = \theta^{(t-1)}\right]$,

$$\mathbb{P}\left[\mathcal{A}^{(t)}(\mathbb{D}) = \theta^{(t)}\right] = \mathbb{P}\left[\mathcal{A}^{(t-2)}(\mathbb{D}) = \theta^{(t-2)}\right] \mathbb{P}\left[\mathcal{A}_{t-1}(\theta_1, \ldots, \theta_{t-2}, \mathbb{D}) = \theta_{t-1}\right] \mathbb{P}\left[\mathcal{A}_t(\theta_1, \ldots, \theta_{t-1}, \mathbb{D}) = \theta_t\right], \tag{5}$$

$$= \mathbb{P}\left[\mathcal{A}_1(\mathbb{D}) = \theta_1\right] \mathbb{P}\left[\mathcal{A}_2(\theta_1, \mathbb{D}) = \theta_2\right] \ldots \mathbb{P}\left[\mathcal{A}_t(\theta_1, \ldots, \theta_{t-1}, \mathbb{D}) = \theta_t\right], \tag{6}$$

$$= \mathbb{P}\left[\hat{\mathcal{A}}^{(t)}(\theta_1, \ldots, \theta_{t-1}, \mathbb{D}) = \theta^{(t)}\right]. \tag{7}$$

In words, $\mathcal{A}^{(t)}$ and $\hat{\mathcal{A}}^{(t)}$ are identical in $\mathbb{A}$. Therefore, $\mathcal{A}^{(t)}$ satisfies (ε, δ)-DP at any $\mathbb{S} \subset \mathbb{A}$ if $\hat{\mathcal{A}}^{(t)}$ satisfies (ε, δ)-DP. Converting the RDP bound on $\hat{\mathcal{A}}^{(t)}(\mathbb{D})$ into a (ε, δ)-DP bound with Lemma 3.2 then completes the proof.

Lemma 3.2 (Conversion from RDP to (ε, δ)-DP [Mir17])**.** *If $\mathcal{A}$ satisfies (α, ρ)-RDP, then $\mathcal{A}$ satisfies $(\rho + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP for all $0 < \delta < 1$.*

□

We then give the privacy guarantee of Algorithm 1 as a corollary of Theorem 3.1.

Corollary 3.2.1 (Ex-post privacy guarantee of Algorithm 1)**.** *The model at step t in Algorithm 1 satisfies $(o_\alpha^{(i)} + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-ex-post individual DP for the i_{th} example at the subset of outcomes specified by the first $t-1$ models, where $o_\alpha^{(i)}$ is the accumulated RDP at order α.*

Proof. We instantiate $\hat{\mathcal{A}}^{(t)}(\mathbb{D})$ with a sequence of realized steps of Algorithm 1. The basic composition theorem of RDP (Lemma 3.3) gives the accumulated RDP of $\hat{\mathcal{A}}^{(t)}(\mathbb{D})$. The claim then follows as a direct corollary of Theorem 3.1.

Lemma 3.3 (Composition of RDP [Mir17])**.** *If $\mathcal{A}_1 : \mathbb{D} \to Range(\mathcal{A}_1)$ is (α, ρ_1)-RDP and $\mathcal{A}_2 : \mathbb{D} \times Range(\mathcal{A}_1) \to Range(\mathcal{A}_2)$ is (α, ρ_2)-RDP for some constants ρ_1 and ρ_2, then their composition is $(\alpha, \rho_1 + \rho_2)$-RDP.*

□

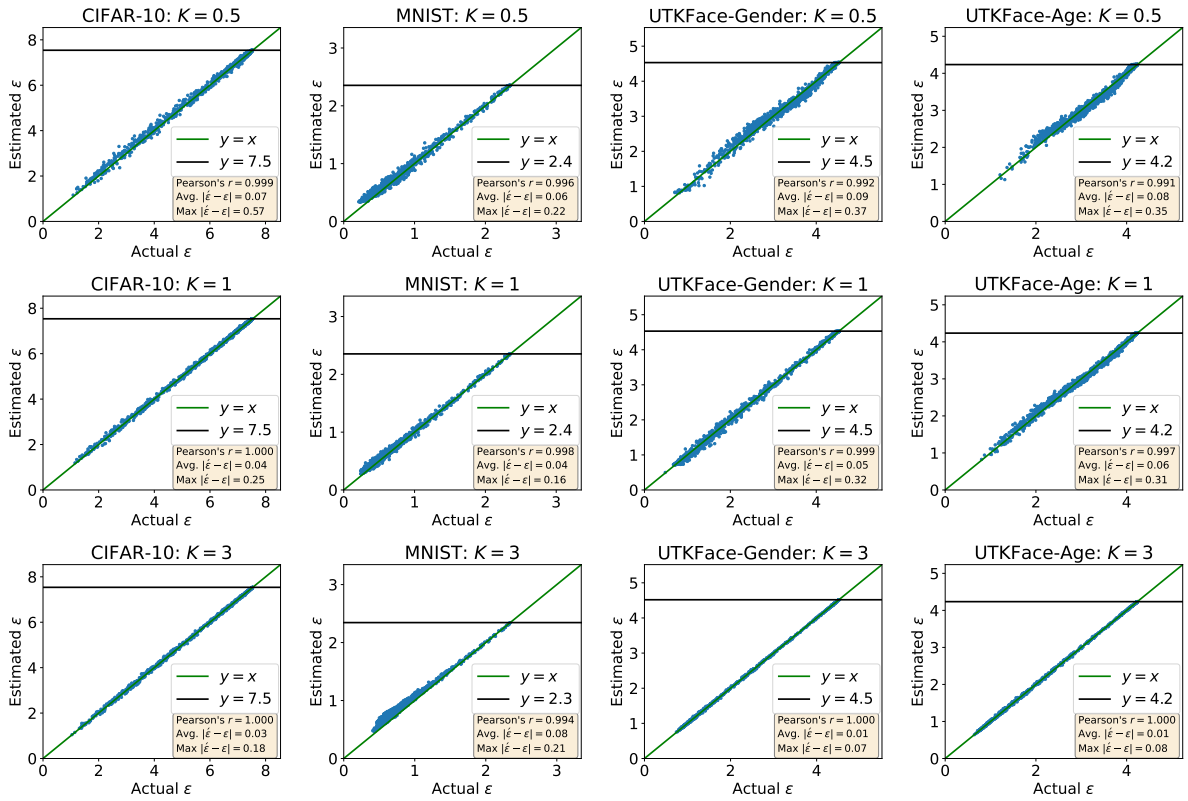

Figure 3: Estimated ε values versus actual ε values. The value of K is the number of full batch norms update at every epoch. The solid black line indicates ε of the original analysis for every example.

3.2 Estimated Privacy Parameters Are Accurate

Although the gradient norms used for privacy accounting are updated only occasionally, we show that the computed individual privacy parameters are very close to the actual ones (Figure 3). This indicates that, in general, the gradient norms do not change rapidly during training.

To compute the actual privacy parameters, we randomly sample 1000 examples and compute the exact gradient norms at every iteration. We compute the Pearson correlation coefficient between the estimated and actual privacy parameters as well as the average and the worst absolute errors. In additional to CIFAR-10 and MNIST, we also include the UTKFace dataset and run age/gender classification tasks on it. Details about the experiments are in Section 4. We plot the results in Figure 3. The estimated ε values are very close to the actual ones (Pearson's $r > 0.99$) even we only update the gradient norms every two epochs. Updating full batch gradient norms more times further improves the estimation, though doing so would increase the computational overhead.

It is worth noting that C affects the computed privacy parameters. Large C increases the variation of gradient norms but leads to large worst-case privacy (or large gradient variance if keeping the worst-case privacy unchanged) while small C suppresses the variation and leads to large gradient bias [CWH20, SSTT21]. In this work, we set the maximum clipping threshold as the median of gradient norms at initialization unless otherwise mentioned, which is a common choice in practice and has been observed to achieve good accuracy [ACG+16]. In Appendix C, we show the influence of using different C on both accuracy and privacy.

3.3 Rounding Individual Gradient Norms

The rounding operation in Algorithm 1 is essential to make the computation of individual privacy parameters feasible. To compute the privacy cost of one example at each step, we need to upper bound the Rényi divergence between Equation (1) and (2). For a fixed sampling probability, the privacy cost is different for every different gradient norm c.

Table 1: Computational costs of computing individual privacy parameters for CIFAR-10 with $K = 1$.

	With Rounding ($r = 0.1$)	Without Rounding
# of numerical computations	$< 1.5 \times 10^2$	1×10^7
Time (in seconds)	< 3	$\sim 2.6 \times 10^4$

Consequently, there are at most $n \times E$ different privacy costs because individual gradient norms vary across different examples and epochs (n is the number of datapoints and E is the number of training epochs). In order to make the number of different privacy costs tractable, we round individual gradient norms with a prespecified precision r. Because the maximum clipping threshold is usually a small constant, then, by the pigeonhole principle, there are at most $\lceil C/r \rceil$ different values. Throughout this paper we set $r = 0.1$ that has almost no impact on the precision of privacy accounting.

We give a concrete comparison on the computational costs in Table 1. We run the numerical method in [MTZ19] once for every different privacy cost (with the default setup in the Opacus library [YSS+21]). We run DP-SGD on CIFAR-10 for 200 epochs with $K = 1$. All results in Table 1 use parallelization with 5 cores of an AMD EPYC™ 7V13 CPU. With rounding, the overhead of computing individual privacy parameters is negligible. In contrast, the computational cost without rounding is more than 7 hours.

3.4 What Can We Do with Individual Privacy Parameters?

Note that individual privacy parameters are dependent on the private data and thus sensitive, and consequently may not be released publicly without care. We describe some approaches to safely make use of individual privacy parameters. The first is to only release the privacy parameter to the rightful data owner. The second is to release some statistics of the individual privacy parameters to the public. Both approaches offer more granular and tighter privacy guarantees than the single worst-case guarantee. Another approach is for a trusted data curator to improve the model quality based on the individual parameters.

The first approach is to release ε_i to the owner of the ith example. Although we use gradient norms without adding noise, this approach does not incur additional privacy loss for two reasons. First, it is safe for the ith example because only the rightful owner sees ε_i. Second, releasing ε_i does not increase the privacy loss of any other examples. This is because the *post-processing* property of differential privacy and the fact that computing ε_i only involves a differentially private model and the ith example itself. The second reason is important and it may not hold for other private learning algorithms. For example, the individual privacy parameters for objective perturbation are interdependent and require additional delicate analysis before publication [RW21].

The second approach is to privately release aggregate statistics of the population, e.g., the average or quantiles of the ε values. Recent works have demonstrated such statistics can be published accurately with minor privacy cost [ATMR21]. We show the statistics can be released accurately with a very small privacy parameter ($\varepsilon = 0.1$) in Appendix D.

Finally, individual privacy parameters can also serve as a powerful tool for a trusted data curator to improve the model quality. By analysing the individual privacy parameters of a dataset, a trusted curator can focus on collecting more data representative of the groups that have higher privacy risks to mitigate the disparity in privacy.

4 Individual Privacy Parameters on Different Datasets

We first show the distribution of individual privacy parameters of running DP-SGD on four classification tasks in Section 4.1. Then we investigate how individual privacy parameters correlate with training loss in Section 4.2. Experimental setup is as follows.

Datasets. We use two benchmark datasets MNIST ($n = 60000$) and CIFAR-10 ($n = 50000$) [LBBH98, Kri09] as well as the UTKFace dataset ($n \simeq 15000$) [ZSQ17] that contains the face images of four different races (White, $n \simeq 7000$; Black, $n \simeq 3500$; Asian, $n \simeq 2000$; Indian, $n \simeq 2800$). We construct two tasks on UTKFace: predicting gender, and predicting whether the age is under 30.[2] We slightly modify the dataset between these two tasks by randomly removing

[2]We acknowledge that predicting gender and age from images may be problematic. Nonetheless, as facial images have previously been highlighted as a setting where machine learning has disparate accuracy on different groups, we revisit this domain through a related lens. The labels are provided

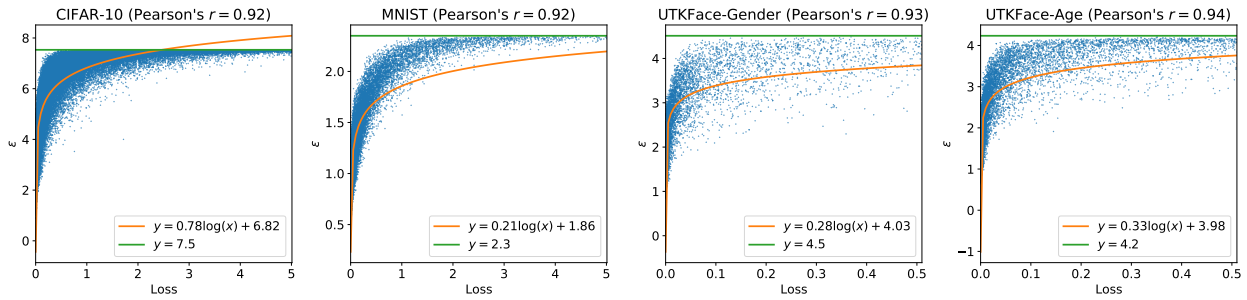

Figure 4: Individual privacy parameters and final training losses. Each point shows the loss and privacy parameter of one example. The Pearson correlation coefficient is computed with privacy parameters predicted by the fitted curve.

a few examples to ensure each race has balanced positive and negative labels.

Models and hyperparameters. We train ResNet-20 models on all datasets. For CIFAR-10 and MNIST, we train the models from scratch. For UTKFace, we fine-tune a model from the PyTorch library[3] that is pre-trained on ImageNet. For all datasets, the maximum clipping threshold is the median of gradient norms at initialization. We set $K = 3$ for all experiments in this section. More details about the models and hyperparameters are in Appendix A.

4.1 Individual Privacy Parameters Vary Significantly

Figure 1 shows the individual privacy parameters on all datasets. The privacy parameters vary across a large range on all tasks. On the CIFAR-10 dataset, the maximum ε_i value is 7.5 while the minimum ε_i value is only 0.7. We also observe that, for easier tasks, more examples enjoy stronger privacy guarantees. For example, $\sim$30% of examples reach the worst-case ε on CIFAR-10 while only $\sim$3% do so on MNIST. This may because the loss decreases quickly when the task is easy, resulting in gradient norms also decreasing and thus a reduced privacy loss.

4.2 Privacy Parameters And Loss Are Positively Correlated

We study how individual privacy parameters correlate with individual losses of the trained model. The analysis in Section 2 suggests that the privacy parameter of one example depends on its gradient norms across training. Intuitively, an example would have high loss after training if its gradient norms are large. We visualize individual privacy parameters and the final training losses in Figure 4. The individual privacy parameters on all datasets increase with losses until they reach the maximum ε. To quantify the order of correlation, we further fit the points with one-dimensional logarithmic functions and compute the Pearson correlation coefficients with the privacy parameters predicted by the fitted curves. The Pearson correlation coefficients are larger than 0.9 on all datasets, suggesting the privacy parameter of one example correlates with its final training loss logarithmically.

5 Groups Are Simultaneously Underserved in Both Accuracy and Privacy

It is well-documented that machine learning models may have large differences in accuracy on different groups [BG18, BPS19]. Our finding demonstrates that this disparity may be simultaneous in terms of both accuracy *and* privacy. We empirically verify this by plotting the average ε and training/test accuracy of different groups. For CIFAR-10 and MNIST, the groups are the data from different classes, while for UTKFace, the groups are the data from different races.

We plot the results in Figure 5. The groups are sorted based on the average ε. Both training and test accuracy correlate well with ε. Groups with worse accuracy do tend to have worse privacy guarantee. On CIFAR-10, the average ε of the 'Cat' class (which has the worst test accuracy) is 26.3% higher than the average ε of the 'Automobile' class (which has the highest test accuracy). On UTKFace-Gender, the average ε of the group with the lowest test accuracy ('Asian') is

by the dataset curators.

[3] https://pytorch.org/vision/stable/models.html

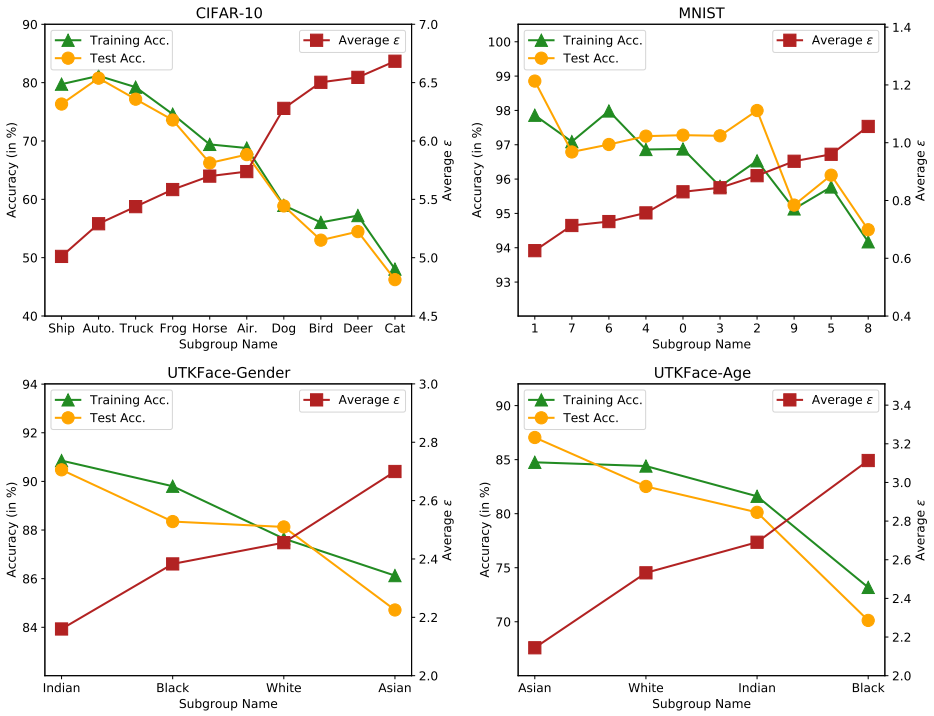

Figure 5: Accuracy and ε of different groups. Groups with worse accuracy also have worse privacy in general.

25.0% higher than the average ε of the group with the highest accuracy ('Indian'). Similar observation also holds on other tasks. To the best of our knowledge, our work is the first to reveal this simultaneous disparity.

6 Privacy Parameters Reflect Empirical Privacy Risks

We run membership inference (MI) attacks to verify whether examples with larger privacy parameters have higher privacy risks in practice. We use a simple loss-threshold attack that predicts an example is a member if its loss value is smaller than a prespecified threshold [SDS+19]. Previous works show that even large privacy parameters are sufficient to defend against such attacks [CLE+19, YZC+21]. In order to better observe the difference in privacy risks, we also include models trained without differential privacy. We do not show the results on MNIST because the attack success rates are close to random guessing for both DP and non-DP models. This is probably because the generalization gap of MNIST is very small even models are trained without DP. For each group, we use its whole test set and a random subset of training set so the numbers of training and test losses are balanced. We further split the data into two subsets evenly to find the optimal threshold on one and report the success rate on another.

The results are in Figure 6. The groups are sorted based on their average ε. When the models are trained with DP, all attack success rates are close to random guessing (50%), as anticipated. Although the attack we use can not show the disparity in this case, we note that there are more powerful attacks whose success rates are closer to the theoretical bound that DP offers [JUO20, NST+21]. On the other hand, the difference in privacy risks is clear when models are trained without DP. On CIFAR-10, the MI success rate is 79.7% on the 'Cat' class (which has the worst average ε) while is only 61.4% on the 'Ship' class (which has the best average ε). These results suggest that the ε values reflect empirical privacy risks which could vary significantly on different groups.

7 Conclusion

We propose an algorithm to compute individual privacy parameters for DP-SGD. The algorithm can give accurate estimations of individual privacy parameters. We use this new algorithm to examine individual privacy guarantees for

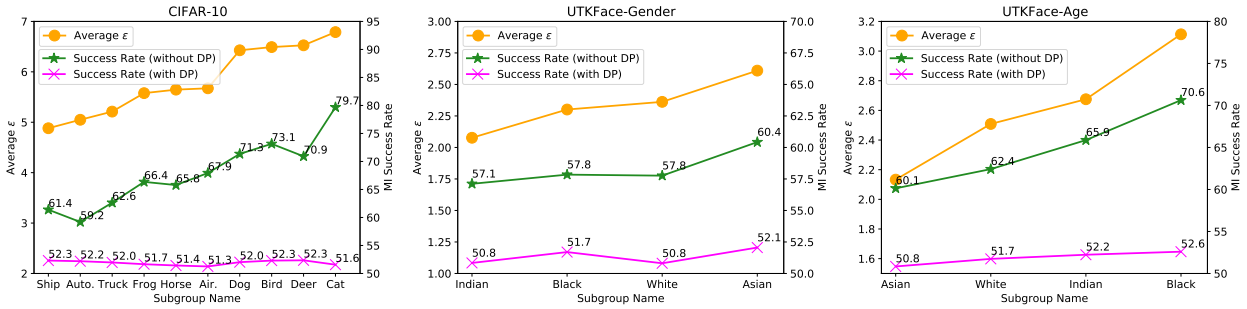

Figure 6: Average ε and membership inference success rates on different groups.

examples in several datasets. Significantly, we find that groups with worse utility also suffer from worse privacy. Our paper reveals the complex while interesting relation among utility, fairness, and privacy, which may inspire new studies of jointly considering these factors.

Acknowledgments

The authors would like to thank Yu-Xiang Wang and Saeed Mahloujifar for pointing out a flaw in the privacy guarantee in the previous version. The authors would also like to thank Yu-Xiang Wang for suggestions about ex-post DP.

References

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, 2016. ACM.

[AGG+21] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private BERT. *arXiv preprint arXiv:2108.01624*, 2021.

[ATMR21] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[BBG18] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pages 6277–6287. Curran Associates, Inc., 2018.

[BDLS20] Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J. Su. Deep learning with Gaussian differential privacy. *Harvard Data Science Review*, 2(3), 2020.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency*, FAT* '18, pages 77–91. JMLR, Inc., 2018.

[BGRK19] Daniel Bernau, Philip-William Grassal, Jonas Robl, and Florian Kerschbaum. Assessing differentially private deep learning with membership inference. *arXiv preprint arXiv:1912.11328*, 2019.

[BPS19] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 15479–15488. Curran Associates, Inc., 2019.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 464–473, Washington, DC, USA, 2014. IEEE Computer Society.

[CCN+21] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021.

[CCTCP21] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, ICML '21, pages 1964–1974. JMLR, Inc., 2021.

[CLE+19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, USENIX Security '19, pages 267–284. USENIX Association, 2019.

[CWH20] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 13773–13782. Curran Associates, Inc., 2020.

[DBH+22] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[DR14a] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DR+14b] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 2014.

[FZ20] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 2881–2891. Curran Associates, Inc., 2020.

[FZ21] Vitaly Feldman and Tijana Zrnic. Individual privacy accounting via a Renyi filter. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[GAW+22] Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the 2022 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '22, Washington, DC, USA, 2022. IEEE Computer Society.

[GLW21] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[HNS+22] Victor Petrén Bach Hansen, Atula Tejaswi Neerkaje, Ramit Sawhney, Lucie Flek, and Anders Søgaard. The impact of differential privacy on group disparity mitigation. *arXiv preprint arXiv:2203.02745*, 2022.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778, Washington, DC, USA, 2016. IEEE Computer Society.

[JUO20] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? In *Advances in Neural Information Processing Systems 33*, NeurIPS '20. Curran Associates, Inc., 2020.

[JYC15] Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? personalized differential privacy. In *International Conference on Data Engineering (ICDE)*, 2015.

[KL17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, 2017.

[Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[KYC+19] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *arXiv preprint arXiv:1906.00389*, 2019.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBG17] Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*, 2017.

[Léc21] Mathias Lécuyer. Practical privacy filters and odometers with r\'enyi differential privacy and applications to differentially private deep learning. *arXiv preprint arXiv:2103.01379*, 2021.

[LNR+17] Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Steven Z Wu. Accuracy first: Selecting a differential privacy level for accuracy constrained erm. *Advances in Neural Information Processing Systems*, 2017.

[LTLH22] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *Proceedings of the 10th International Conference on Learning Representations*, ICLR '22, 2022.

[MB22] Christopher Mühl and Franziska Boenisch. Personalized PATE: Differential privacy for machine learning with individual privacy guarantees. *arXiv preprint arXiv:2202.10517*, 2022.

[Mir17] Ilya Mironov. Rényi differential privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium*, CSF '17, pages 263–275, Washington, DC, USA, 2017. IEEE Computer Society.

[MTKC22] Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:*, 2022.

[MTZ19] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[NHS22] Frederik Noe, Rasmus Herskind, and Anders Søgaard. Exploring the unfairness of dp-sgd across settings. *arXiv preprint arXiv:2202.12058*, 2022.

[NST+21] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. *arXiv preprint arXiv:2101.04535*, 2021.

[PAE+17] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017.

[RW21] Rachel Redberg and Yu-Xiang Wang. Privately publishable per-instance privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[SCS13] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing*, GlobalSIP '13, pages 245–248, Washington, DC, USA, 2013. IEEE Computer Society.

[SDS+19] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pages 5558–5567. JMLR, Inc., 2019.

[SPGG21] Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. Chasing your long tails: Differentially private prediction in health care settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 723–734. JMLR, Inc., 2021.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, SP '17, pages 3–18, Washington, DC, USA, 2017. IEEE Computer Society.

[SSTT21] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private GLMs. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, AISTATS '21, pages 2638–2646. JMLR, Inc., 2021.

[TSJ+22] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. *arXiv preprint arXiv:2204.00032*, 2022.

[Wan19] Yu-Xiang Wang. Per-instance differential privacy. *The Journal of Privacy and Confidentiality*, 2019.

[WBK19] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS '19, pages 1226–1235. JMLR, Inc., 2019.

[WRRW22] Justin Whitehouse, Aaditya Ramdas, Ryan Rogers, and Zhiwei Steven Wu. Fully adaptive composition in differential privacy. *arXiv preprint arXiv:2203.05481*, 2022.

[YMMS21] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. Enhanced membership inference attacks against machine learning models. *arXiv preprint arXiv:2111.09679*, 2021.

[YNB+22] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *Proceedings of the 10th International Conference on Learning Representations*, ICLR '22, 2022.

[YSS+21] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

[YZC+21] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale privacy learning via low-rank reparametrization. In *Proceedings of the 38th International Conference on Machine Learning*, ICML '21, pages 12208–12218. JMLR, Inc., 2021.

[ZSQ17] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the 2017 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '17, pages 5810–5818, Washington, DC, USA, 2017. IEEE Computer Society.

[ZW19] Yuqing Zhu and Yu-Xiang Wang. Poission subsampled Rényi differential privacy. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pages 7634–7642. JMLR, Inc., 2019.

Table 2: Comparison between the test accuracy of using individual clipping thresholds and that of using a single maximum clipping threshold.

	CIFAR-10	MNIST
Individual	65.42 ($\pm$0.37)	97.17 ($\pm$0.12)
Maximum	65.66 ($\pm$0.68)	97.26 ($\pm$0.11)

A More Details on Hyperparameters

The noise multipliers are 2.2, 4.0, and 1.5 for CIFAR-10, MNIST, and UTKFace, respectively. The standard deviation of noise in Algorithm 1 is the noise multiplier times the maximum clipping threshold. The batchsize is 2000 for CIFAR-10/MNIST and 200 for UTKFace. The training epoch is 200 for CIFAR-10 and 100 for MNIST and UTKFace. For ResNet-20 models on CIFAR-10 and MNIST, we replace batch normalization with group normalization. For ResNet-20 models on UTKFace, we use the pre-trained batch normalization layers in evaluation mode. All experiments are run on a single Tesla V100 GPU.

B Individual Clipping Does Not Affect Accuracy

Here we run experiments to check the influence of using individual clipping thresholds on utility. Algorithm 1 uses individual clipping thresholds to ensure the computed privacy parameters are valid privacy guarantees. If the clipping thresholds are close to the actual gradient norms, then the clipped results are close to those of using a single maximum clipping threshold. However, if the estimations of gradient norms are not accurate, individual thresholds would clip more signal than using a single maximum threshold.

We compare the accuracy of using the maximum clipping threshold and that of using individual clipping thresholds. The results on CIAFR-10 and MNIST are in Table 2. The individual clipping thresholds are updated once per epoch. We repeat the experiment four times with different random seeds. Other setups are the same as those in Section 4. The results suggest that using individual clipping thresholds in Algorithm 1 does not affect the accuracy.

C The Influence of the Maximum Clipping on Individual Privacy

As discussed in Section 3.2, the value of the maximum clipping threshold C would affect individual privacy parameters in Algorithm 1. Here we run experiments with different values of C on CIFAR-10. Let M be the median of gradient norms at initialization, we choose C from the list $[0.2M, 0.5M, M, 1.5M, 2M, 5M]$. Other experimental setup is the same as that in Section 4.

We plot the results in Figure 7. The variation in privacy parameters increases with the value of C. When $C = 0.2M$, nearly 70% datapoints reach the worst privacy parameter while only 3% datapoints reach the worst parameter when $C = 1.5M$. When $C = 5M$, the maximum privacy parameter is only 2.3.

D Release Populational Statistics of Individual Privacy Parameters

The individual privacy parameters computed by Algorithm 1 are sensitive and hence can not be directly released to the public. Here we show the populational statistics of individual parameters can be released with minor privacy cost. Specifically, we compute the average and quantiles of the ε values with differential privacy. For average, we release the noisy aggregation through Gaussian Mechanism. For quantiles, we solve the objective function in [ATMR21] with 20 steps of full batch gradient descent. The results on MNIST and CIFAR-10 are in Table 3 and Table 4 respectively. The released statistics are close to the actual values on both datasets with $(0.1, 10^{-5})$-DP.

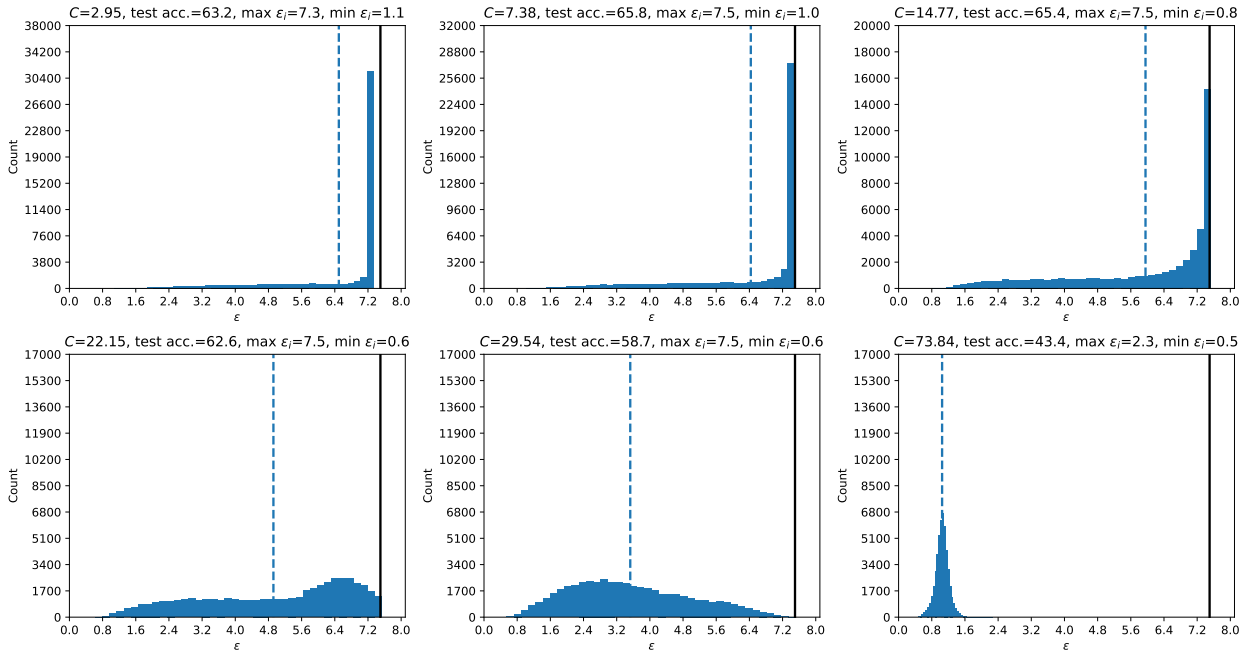

Figure 7: Distributions of individual privacy parameters on CIFAR-10 with different values of C. The median of gradient norms at initialization is 14.77. The black solid line indicates the original privacy parameter of DP-SGD for all examples. The maximum ε_i value in the first plot does not match the maximum ε because we round individual gradient norms with precision 0.1.

Table 3: Populational statistics of individual privacy parameters on MNIST. The average estimation error rate is 1.19%. The value of δ is 1×10^{-5}.

MNIST	Average	0.1-quantile	0.3-quantile	Median	0.7-quantile	0.9-quantile
Non-private	0.850	0.362	0.467	0.626	0.931	1.840
$\varepsilon = 0.1$	0.852	0.361	0.469	0.628	0.935	1.742

Table 4: Populational statistics of individual privacy parameters on CIFAR-10. The average estimation error rate is 1.51%. The value of δ is 1×10^{-5}.

CIFAR-10	Average	0.1-quantile	0.3-quantile	Median	0.7-quantile	0.9-quantile
Non-private	5.813	2.870	4.957	6.602	7.290	7.488
$\varepsilon = 0.1$	5.811	2.960	4.975	6.609	7.357	7.828