DYNAMIC DIFFERENTIAL-PRIVACY PRESERVING SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially-Private Stochastic Gradient Descent (DP-SGD) prevents training-data privacy breaches by adding noise to the clipped gradient during SGD training to satisfy the differential privacy (DP) definition. On the other hand, the same clipping operation and additive noise across training steps results in unstable updates and even a ramp-up period, which significantly reduces the model's accuracy. In this paper, we extend the Gaussian DP central limit theorem to calibrate the clipping value and the noise power for each individual step separately. We, therefore, are able to propose the dynamic DP-SGD, which has a lower privacy cost than the DP-SGD during updates until they achieve the same target privacy budget at a target number of updates. Dynamic DP-SGD, in particular, improves model accuracy without sacrificing privacy by gradually lowering both clipping value and noise power while adhering to a total privacy budget constraint. Extensive experiments on a variety of deep learning tasks, including image classification, natural language processing, and federated learning, show that the proposed dynamic DP-SGD algorithm stabilizes updates and, as a result, significantly improves model accuracy in the strong privacy protection region when compared to DP-SGD.

1 INTRODUCTION

Data privacy protection is becoming increasingly important; not only are data breaches gaining public attention, but there are also data protection initiatives and data privacy laws in place, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Deep learning, on the other hand, poses a significant risk of data privacy leakage since the model embeds information about the training data. For example, both Zhu & Han (2020) and Zhao et al. (2020) provide paradigms for reconstructing training examples from published models. As a result, privacy-preserving algorithms are becoming increasingly important for preserving data privacy.

Differential privacy (DP) is a provable and quantifiable method for privacy protection (Dwork, 2006), which guarantees it nearly impossible for the adversary to differentiate from two *neighboring data sets*. However, applying DP to maintain the privacy of the training data in deep learning is extremely difficult due to the thousands to millions of repetitions[1] of privacy budget cost in stochastic gradient descent (SGD) iterations, resulting in a very high additive noise power that completely degrades the learning process within a reasonable privacy preserving region.

The first DP-SGD with a reasonable level of accuracy is proposed by Abadi et al. (2016). The calibrated noise added to the clipped gradient is much smaller than all previous methods due to their moments-accountant method for tight privacy accounting. Even so, the accuracy drop when compared to the non-DP model is still significant. In the MNIST dataset, for example, the moment-accountant based DP-SGD sacrifices accuracy by about 4.5% with the privacy protection level $\epsilon = 1.34$ (Bu et al., 2020). Performance will continue to suffer as the level of privacy protection increases. This motivates us to narrow the performance gap in this paper without compromising privacy.

To begin, motivated by the instability and even ramp-up of DP-SGD updates, as shown later in Fig. 1, we investigate how to perform a tight DP accounting to enable proactive DP budget allocation for each training step to avoid unstable updates. Second, under this new privacy accounting framework, we propose instances of algorithms for proactive privacy budget allocation, such as the sensitivity-decay method, the growing-μ_t method, and a hybrid of the two known as dynamic DP-SGD. Third, we thoroughly validate the performance of dynamic DP-SGD for a variety of deep learning tasks,

[1] Around 300,000 steps is needed to train large models like GPT3.

demonstrating consistently significant accuracy improvements over existing methods while preserving privacy. The technical contributions are summarized below.

- We broaden Gaussian DP's central limit theorem (CLT) to obtain a tight bound for the composition of Gaussian mechanisms with different privacy parameters considering subsampling amplification. The closed form of the CLT result facilitates the proactive calibration of different noise powers and clipping values at each training step.
- We investigate how to proactively allocate the DP cost based on the CLT result in such a way that the gradient updates are stabilized and thus model performance is improved. We propose the *sensitivity decay* and *growing-μ_t* methods, and demonstrate that combining these two methods results in the best performance, which is referred to as *dynamic DP-SGD*.
- We perform a series of experiments and ablation studies on a variety of neural network tasks, including image classification, natural language processing, and federated learning, and show that the proposed dynamic DP-SGD effectively stabilizes gradient updates and consistently outperforms existing methods. With a strong privacy guarantee, i.e., at $\epsilon = 1.2$, dynamic DP-SGD has a performance loss of only 2.49% when compared to the none-DP version on MNIST dataset, and the loss is reduced to 1.72% in the federated learning setting for a stronger privacy guarantee, i.e., $\epsilon = 1$ due to a larger DP amplification.

2 Related Work

The following section summarizes related work on algorithm design and DP accounting for DP-SGD.

DP-SGD Algorithm To improve the model's accuracy, previous work has concentrated on designing variations of DP-SGD by estimating the clipping bound and minimizing the bias introduced by gradient clipping. More precisely, Abadi et al. (2016) propose norm clipping and per-layer clipping, both of which select clipping values based on gradient differences between different layers. Pichapati et al. (2019) pioneer AdaClip, a coordinate-wise clipping method that significantly reduces the total amount of noise required. Thakkar et al. (2019) introduce gradient clipping based on the quantile statistics of the gradient, which requires additional DP cost to protect those quantiles. Recently, Chen et al. (2020) analyze the bias introduced by the gradient clipping operation and propose a method for reducing the bias error by first adding noise before clipping. It's worth noting that the proposed dynamic DP-SGD in this paper is compatible with the methods mentioned above and can be used in tandem to investigate accuracy improvement.

Furthermore, Yu et al. (2019) provide a means of reducing noise variance during the DP-SGD process, thereby improving model performance; and Zhang et al. (2021b) analyze the DP cost for the same method using the z-CDP privacy accounting. Due to the loose DP accountings, these methods have a large performance gap in the high privacy guaranteed region. As demonstrated later in the experiments, the proposed dynamic DP-SGD improves these results significantly due to the dynamic clipping operation and tight DP composition.

Privacy Accounting Each step in the DP-SGD process depletes the total privacy budget. Following T steps of training, DP accounting necessitates the composition of these mechanisms in order to calculate the total privacy cost in terms of (ϵ, δ). Simple composition (Dwork et al., 2006; Dwork & Lei, 2009) with $\epsilon = \mathcal{O}(T)$ and advance composition (Dwork et al., 2010) with $\epsilon = \mathcal{O}(\sqrt{T})$ have been proposed in the literature. If $M_1, M_2, \ldots, M_T$ are distinct DP mechanisms such that M_i is (ϵ_i, δ_i)-DP, then it is shown by Murtagh & Vadhan (2016) that computing the exact DP guarantees for the composition $M = M_1 \circ M_2 \circ \cdots \circ M_T$ is #P-complete. Abadi et al. (2016) propose the moment accountant of DP-SGD. By establishing a privacy cost function, the moment accountant imposes a tight constraint on the estimation of privacy loss, calibrating the noise power down to a practicable level for the first time in the privacy preserving deep learning context. From a broader perspective, Mironov (2017) proposes to use Rényi divergence to determine the distance between the outputs of two adjacent datasets, thereby establishing the RDP concept, and Balle et al. (2020) provide an advance conversion from RDP to (ϵ, δ)-DP. Due to the inherent use of subsampling in training neural networks, the DP composition with subsampling DP amplification is studied in detail by Wang et al. (2019); Balle et al. (2020), which reduce the noise power for the privacy preserving. Dong et al. (2019) recently propose the concept of f-DP to quantify the privacy cost from a hypothesis testing perspective, with Gaussian DP (GDP) as a major application. While GDP allows for a

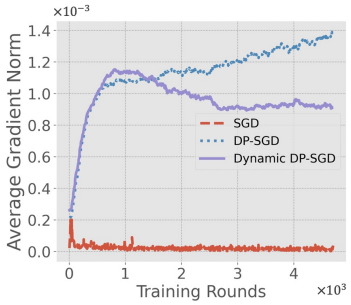

Figure 1: Experiments on MNIST with DP budget $(\epsilon, \delta) = (0.4, 10^{-5})$ after 5×10^3 steps for both DP-SGD and proposed dynamic DP-SGD. In comparison to SGD without DP protection, DP-SGD is unstable and has a ramp-up period. In contrast, the gradient norm is stabilized by the proposed dynamic DP-SGD.

tight composition, computing the exact composition of the Gaussian mechanism with subsampling amplification is computationally challenging. It is also demonstrated that a computationally efficient central limit theorem (CLT) can approximate the composition of multiple identical DP mechanisms. Very recently, Gopi et al. (2021); Zhu et al. (2021) propose optimal composition of DP mechanisms via numerical methods, but it is unclear how to calibrate different noise power and clipping value at each step as in the proposed dynamic DP-SGD.

3 An Inspiring Case of the Unstable DP-SGD

DP defines an upper bound, i.e., $(\epsilon, \delta(\epsilon))$ on the privacy budget. Lower ϵ values indicate better privacy protection. The value δ can be interpreted as the probability of failing to achieve DP. The neighboring data sets, i.e., X and X', which differs by one *data record* is denoted by $X \sim X'$. The widely accepted (ϵ, δ)-DP is defined as below (Dwork, 2006).

Definition 1. *($(\epsilon, \delta(\epsilon))$-DP Profile) A randomized algorithm $M(\cdot)$ gives $(\epsilon, \delta(\epsilon))$-differential privacy if for any pair of neighboring datasets $X \sim X'$ and any event E,*

$$\mathbb{P}(M(X) \in E) \leqslant \mathrm{e}^{\epsilon} \mathbb{P}\left(M\left(X'\right) \in E\right) + \delta,$$

where the probability $\mathbb{P}(\cdot)$ is taken over the randomness of M, and $\epsilon \geq 0$. When $\delta = 0$, the algorithm is ϵ-DP. Intuitively, this means that we can't tell whether M was run on X or X' based on the results. As a result, an adversary cannot infer the existence of any specific data record in the input data set.

In a Gaussian mechanism, let Y be the random variable following Gaussian distribution with $Y \sim \mathcal{N}\left(0, \sigma^2 I_d\right)$ and $f : X^n \to \mathbb{R}^d$. The Gaussian mechanism $M(X) = f(X) + Y$ follows the DP profile (Wang et al., 2019):

$$\delta(\epsilon; \mu) = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - \mathrm{e}^{\epsilon} \Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right), \tag{1}$$

where

$$\mu = \frac{C}{\sigma} \tag{2}$$

with C the sensitivity of $f(X)$ and $\Phi(\cdot)$ the Gaussian cumulative distribution function. The DP-SGD is proposed below based on the Gaussian mechanism. Let the model parameters be denoted by θ. The gradient computed by a data sample x in the SGD is given by $g_x \triangleq \frac{\partial f_x}{\partial \theta}$, where f_x is the corresponding loss function. To calibrate the noise required for DP, a paradigm is to first clip the ℓ_2-norm of the gradient, i.e.,

$$\widetilde{g}_x \triangleq \mathrm{CL}\left(g_x; C\right) \triangleq g_x \cdot \min\left(1, \frac{C}{\|g_x\|}\right), \tag{3}$$

and in the subsequent, to add the calibrated noise $\xi_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = \frac{C}{\mu}$. Thereby, the DP-SGD updates at the t-th step is given by (Abadi et al., 2016):

$$\text{DP-SGD:} \quad \theta_t = \theta_{t-1} - \eta \frac{1}{|X_t|} \left(\sum_{x \in X_t} \widetilde{g}_x + \xi_t \right), \quad t \in [T]. \tag{4}$$

Because of DP's post-processing property, protecting gradients provides the same level of privacy protection on the output model. A closer examination of the DP-SGD process in (4) reveals that the clipping-per-sample operator introduces biases to the original unbiased gradient estimate in the SGD updates if C is not large enough. It is possible to have an unbiased gradient estimate by increasing C if C is greater than any $\frac{\partial f_x}{\partial \theta}$. It does, however, result in an over-calibrated noise power.

Motivations of this work: In standard SGD updates, the gradient norm is reduced to a very small value. However, because the DP-SGD is followed by evenly consuming the privacy budget, the ratio of noise power to the true gradient norm in Eqn. (4) would continue to rise, resulting in unstable updates. This hypothesis is supported by Fig. 1, which displays the average coordinate gradient norm for all the data records in each step during the iterative updates. It is worth noting that the DP-SGD gradient norm has a ramp-up period. This phenomenon also appeared in the experiments in (Thakkar et al., 2019). The observation motivates us to investigate a dynamic DP-SGD to reduce both the clipping value and the noise power in order to stabilize the updates. As a result, we can control the rate at which the privacy budget consumed. Despite the idea of a dynamic DP-SGD, existing DP accounting methodologies are inadequate to facilitate the study on under what criterior to allocate the privacy budget. As a result, in the following section, we first present an extended CLT of DP accounting based on Gaussian DP (GDP) (Dong et al., 2019) and then demonstrate how to achieve dynamic DP-SGD with the extended CLT of GDP to improve model accuracy.

4 DYNAMIC DP-SGD

The dynamic DP-SGD is presented in two parts: an extension of the CLT of GDP, specifically for dynamic noise and changing gradient clipping values, and the algorithms, which include growing-μ_t, sensitivity decay, and dynamic DP-SGD.

4.1 EXTENDED CLT FOR GDP

GDP Preliminary: We first introduce some background about GDP. Let P and Q denote the distributions of $M(X)$ and $M(X')$ with $X \sim X'$, and let ϕ be any (possibly randomized) rejection rule for testing $H_0 : P$ against $H_1 : Q$. With these in place, Dong et al. (2019) defines the trade-off function of P and Q as

$$\begin{aligned} \mathrm{T}(P, Q) : [0, 1] &\mapsto [0, 1] \\ \alpha &\mapsto \inf_{\phi} \left\{ 1 - \mathbb{E}_Q[\phi] : \mathbb{E}_P[\phi] \leqslant \alpha \right\}. \end{aligned} \tag{5}$$

Above, $\mathbb{E}_P[\phi]$ and $1 - \mathbb{E}_Q[\phi]$ are type I and type II errors of the rejection rule ϕ, respectively. It is shown that $\mathrm{T}(P, Q) \geq \mathrm{T}(\mathcal{N}(0, 1), \mathcal{N}(\mu, 1)) \triangleq G_\mu$, which is referred to as μ-GDP.

In each step of the DP-SGD in (4) with the Gaussian mechanism, it achieves μ-GDP with $\mu = \frac{C}{\sigma}$. Consider the sampling scheme $\mathrm{PS}(X)$ that each individual data sample (x, y) is subsampled independently with probability p from the training set to construct X_t. It is shown in (Bu et al., 2020) that given two neighboring datasets X and X', if a randomized mechanism $\mathcal{M}$ is G_μ-DP, then

$$\mathrm{T}\left(\mathcal{M} \circ \mathrm{PS}(X), \mathcal{M} \circ \mathrm{PS}\left(X'\right)\right) \geqslant pG_\mu + (1 - p)\mathrm{Id}, \tag{6}$$

where $\mathrm{Id}(x) = 1 - x$. Then after a large enough T steps, a Berry-Esseen style CLT result is shown by Bu et al. (2020) that as $T \to +\infty$ and $p\sqrt{T} \to$ a constant, the composition of the r.h.s. of (6) converges to a $G_{\mu_{\text{tot}}}$-DP with

$$\mu_{\text{tot}} = p\sqrt{T(e^{\mu^2} - 1)}. \tag{7}$$

Extended CLT of GDP: The preceding CLT result is based on the assumption that each step satisfies the same G_μ-DP, which restricts the possibility to calibrate noise for each individual step. Following

that, in order to pave the way for DP accounting of a dynamic DP-SGD, we investigate the extended CLT in which each step t has a different G_{μ_t}-DP. Let $\mathcal{M}_A$ and $\mathcal{M}_B$ denote the compositions of T steps updates for the neighboring data sets X and X', respectively. According to (4), each step consists of subsampling and local updates. Then, we express $\mathcal{M}_A$ and $\mathcal{M}_B$ by

$$\mathcal{M}_A \triangleq M_T \circ \mathrm{PS}_T(X) \circ \cdots \circ \mathcal{M}_1 \circ \mathrm{PS}_1(X), \quad \mathcal{M}_B \triangleq M_T \circ \mathrm{PS}_T(X') \circ \cdots \circ \mathcal{M}_1 \circ \mathrm{PS}_1(X').$$

The composition theorem in (Bu et al., 2020) gives

$$\mathrm{T}(\mathcal{M}_A, \mathcal{M}_B) \geqslant \otimes_{t=1}^{T} \left(p \cdot G_{\mu_t} + (1-p)\,\mathrm{Id}\,\right). \tag{8}$$

The symbol $\otimes_{t=1}^{T}$ denotes the product of all the T trade-off functions with the form $p \cdot G_{\mu_t} + (1-p)\,\mathrm{Id}$, which is far from analytically computable. In the following theorem, we develop the CLT for the r.h.s of (8) with the proof provided in Appendix A.

Theorem 1. *Consider a series of adaptive composition mechanisms $\mathcal{M}_t$ for $t \in [T]$, where $\mathcal{M}_t$ is G_{μ_t}-DP, and each mechanism works only on a subsampled data sets by independent Bernoulli trial with probability p. The trade-off function for $\lim_{T\to\infty} \otimes_{t=1}^{T} \left(p \cdot G_{\mu_t} + (1-p)\,Id\,\right)$ in (8) approaches to $G_{\mu_{tot}}$-DP when $p\sqrt{T}$ is a constant, where*

$$\mu_{tot} = p \cdot \sqrt{\sum_{t=1}^{T} \left(\mathrm{e}^{\mu_t^2} - 1\right)}. \tag{9}$$

Notably, when $\mu_t = \mu$ is substituted for $t \in [T]$, Eqn.(9) reduces to the CLT result in (7).

Theorem 1 reflects that, given the target privacy budget (ϵ, δ) and the corresponding privacy parameter μ_{tot} obtained by (1), we can proactively allocate privacy cost to each step $t \in [T]$ for a predefined total number of steps T, according to (9). Because μ_t is determined by both the clipping value and the noise power, we can adjust both C_t and σ_t to control the privacy budget allocation. Sections 4.2-4.4 detail the corresponding algorithms.

4.2 Growing-μ_t Method

It is expected that the gradient should be decreased during training, thereby, it is natural to reduce the noise power to stabilize the gradient updates. As a result, we investigate the noise power decay rate in order to satisfy the μ grow rate and thus control the privacy cost. With (9) in mind, we control the privacy cost rate by adjusting a hyper-parameter ρ_μ with

$$\rho_\mu \triangleq \frac{\mu_T}{\mu_0}, \quad \rho_\mu \geq 1. \tag{10}$$

Then the dynamic μ_t, , which corresponds to the dynamic DP-SGD is given by

$$\mu_t = (\rho_\mu)^{t/T} \cdot \mu_0, \quad \forall t \in [T]. \tag{11}$$

Given the total privacy budget (ϵ, δ), the equivalent privacy parameter μ_{tot} of μ_{tot}-GDP is obtained according to (1). Then the rest problem is to determine the initial state μ_0. Once μ_0 is obtained, the whole $\{\mu_t\}$ sequence can be generated according to (11). By substituting (11) into (9), we obtain

$$\mu_{\text{tot}}^2 = p^2 \cdot \sum_{t=1}^{T} \left(\exp\left\{ \left((\rho_\mu)^{t/T} \cdot \mu_0 \right)^2 \right\} - 1 \right). \tag{12}$$

Algorithm 1 μ_t computation

Require: privacy budget (ϵ, δ), step #T, ρ_μ.

1: Compute μ_{tot} corresponding to (ϵ, δ) according to (1).
2: Compute μ_0 in (12) by binary search.
3: Compute σ_t according to (13) for all $t \in [T]$.

Because the above equation is transcendental when $\rho_\mu > 1$, there is no closed-form solution for μ_0. Nonetheless, the r.h.s of (12) is monotone increasing w.r.t. μ_0, and we can thus solve it efficiently using a numerical method such as binary search. The computation of μ_t is summarized in Algorithm 1. The noise power at iteration t can be calculated using a specific expression of μ_0 and ρ_μ:

$$\sigma_t = \frac{C}{\mu_0} (\rho_\mu)^{-\frac{t}{T}}, \quad \rho_\mu > 1. \tag{13}$$

One example is shown in Fig. 2. Given the total DP budget $\epsilon = 1.2$, growing-μ_t gives the freedom to adjust the privacy cost rate, which is the slope of the curve. In Fig 2, we demonstrate the privacy budget consumption curve for different ρ_μ. The solid line ($\rho_\mu = 1$) represents vanilla DP-SGD with evenly distributed noise power along with the step updates. With growing-μ_t, we can now realize any μ consumption process under the constraint of the total DP budget as shown by the dashed lines. Specifically, the growing-μ_t slows consumption in the early rounds and accelerates consumption in the later rounds.

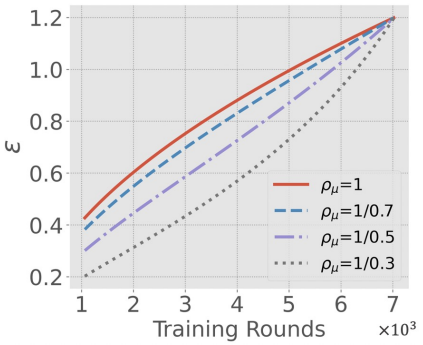

Figure 2: Consumption of privacy budget with different μ_t increasing rate.

4.3 Sensitivity-Decay Method

The gradient norm of a neural network converges to zero once the SGD training converges. This is not the case for DP-SGD. When the required privacy protection level is high, it requires a large calibrated noise promotional to the clipping value. The updated gradient norm tends to rise, as shown in Fig. 1. Based on this discovery, we propose to adjust the sensitivity across training iterations. We set the evolution of clipping values by

$$C_t = (\rho_c)^{-\frac{t}{T}} \cdot C_0, \quad \rho_c \geq 1. \tag{14}$$

Assuming a constant μ_t, and by substituting $\mu_t = \mu_0$ into (9) in Theorem 1, we have the closed form solution:

$$\mu_0 = \sqrt{\log\left(\frac{\mu_{\text{tot}}^2}{p^2 T} + 1\right)}, \quad t \in [T]. \tag{15}$$

With (14) and (15), we can calibrate the noise power at each round by:

$$\sigma_t = \frac{C_0}{\mu_0}(\rho_c)^{-\frac{t}{T}}, \quad \rho_c > 1, \quad t \in [T]. \tag{16}$$

4.4 Dynamic DP

Because the noise calibration is based on the clipping value, we incorporate the growing-μ_t method into the sensitivity-decay method and refer to this new one as dynamic DP-SGD. It maintains the same μ_t increasing rate as the previous growing-μ_t method while having a faster noise decay rate than the sensitivity-decay method. We summarize the dynamic DP-SGD algorithm in Algorithm 2 and conduct extensive experiments to show how dynamic DP-SGD improves performance.

Algorithm 2 Dynamic DP-SGD Algorithm

Require: DP budget (ϵ, δ), sampling rate p and hyper-parameters: ρ_μ, ρ_c and C_0.

1: Compute μ_0 in Algorithm 1
2: **for** $t = 1, \ldots, T$ **do**
3: Compute $C_t = (\rho_c)^{-\frac{t}{T}} \cdot C_0$ in (14)
4: Calibrate noise : $\sigma_t = \frac{C_0}{\mu_0}(\rho_\mu \cdot \rho_c)^{-\frac{t}{T}}$
5: Sample $X_t \in X$ with sampling rate p and sample noise $\xi_t \sim \mathcal{N}(0, \sigma_t^2 I)$.
6: Compute: $\theta_t = \theta_{t-1} - \frac{\eta}{|X_t|}\left[\xi_t + \sum_{x \in X_t} \text{CL}\left(g_x; C_t\right)\right]$ with CL$(\cdot)$ in (3)
7: **end for**

5 Experiments

5.1 Datasets, Models and Benchmarks

Datasets: To conduct a comprehensive test of the dynamic DP-SGD performance, we run experiments on the following 5 datasets: MNIST, FashionMNIST, IMDB, NAME, and InfiniteMNIST, using neural network models such as MLP, CNN, LSTM, and Federated Learning. In Appendix A.2, we describe each data set, the corresponding neural network model, and parameter settings for each experiment separately. We also go into details about the dynamic DP-SGD federated learning algorithm to make the paper self-contained.

Benchmarks: Using the aforementioned datasets and models, we compare our proposed dynamic DP-SGD, growing-μ_t method, and sensitivity-decay method to the four benchmarks listed below.

(i) *SGD without DP*: The SGD method serve as the upper bound of model inference accuracy in the absence of clipping and additive noise. Please note that the DP-SGD and proposed dynamic DP-SGD can also be used to obtain the DP-Adam counterparts of an Adam optimizer due to DP's postprocessing properties (Dwork & Lei, 2009). For the IMDB data set, we apply the Adam to compute the performance upper bound, and DP-Adam counterparts for comparison.

(ii) *DP-SGD with the CLT of GDP Acountant (Bu et al., 2020)*: GDP, a recently developed DP accounting framework, provides a simple, explicit, and tight privacy accounting CLT bound for the classical DP-SGD with evenly calibrated noise power. It is proved by Dong et al. (2019) that the CLT of GDP provides a tighter composition bound for DP-SGD than the moment accountant method (Abadi et al., 2016). As a result, it serves as a standard for our dynamic DP-SGD methods.

(iii) *Noise Power Decay with ρ-zCDP Accountant* (Yu et al., 2019): This paper proposes decaying the noise power during the SGD and computing the privacy loss using the ρ-zCDP. However, only parallel composition is considered, with no regard for DP amplification by subsampling.

(iv) *Noise Power Decay with tCDP Accountant (Zhang et al., 2021a)*: Zhang et al. (2021a) recently proposed decaying the noise power for DP-SGD training as well as analyzing the DP composition and subsampling amplification under the truncated concentrated differential privacy (tCDP) framework.

Parameters: We concentrate on the strong privacy guarantee, with ϵ set to be in the range $[0.4, 9]$ and $\delta = 1/(10|X|)$, where $|X|$ is the training data sample size. It is worth noting that we observed a significant performance degradation of DP-SGD when performing IMDB tasks. As a result, we also test large ϵ values in this case, i.e., $\epsilon = 9$. The hyperparameters ρ_c and ρ_μ are swept in the following predefined sets: $1/\rho_c, 1/\rho_\mu \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Other hyper-parameters are detailed in Appendix A.2[2].

5.2 Results Analysis

Experiment results for the above five different datasets on MLP, CNN, LSTM, and Federated Learning models are shown in Table 1-Table 4, along with benchmarks and proposed methods' performance. It is worth noting that in the majority of cases, we concentrate on the strong privacy protection region with $\epsilon < 1$. As a result, if the DP accountant is not tight enough, it will lead to an overestimation of the noise power needed, resulting in the noise dominating the gradient and causing the network to fail to learn. For example, the method in Yu et al. (2019) fails to leverage the subsampling DP amplification in the DP accounting, resulting in significantly worse performance in the high privacy guarantee region due to overestimated noise. As a result, we only replicate its CNN model results in the MNIST and FashionMNIST datasets.

In contrast, the GDP framework proposed by Bu et al. (2020) provides a detailed examination of DP amplification through subsampling as well as DP composition. Thereby, more precise noise power is calibrated for the same privacy budget. Specifically, even when there are no dynamics for DP-SGD updates, it outperforms the accuracy of the noise decay method by Yu et al. (2019) and Zhang et al. (2021a).

The proposed extended CLT supports the privacy accounting for dynamic clipping and noise power decay. Separate experiments are carried out with the proposed growing-μ_t method, sensitivity-decay method, and dynamic DP-SGD method. The results show that all the three proposed methods improve performance when compared to the static noise GDP method Bu et al. (2020). In particular, μ_t grows at the expense of early convergence speed in order to achieve higher accuracy, while sensitivity decay ensures more stable convergence. This explains why the performance of sensitivity decay outperforms that of growing-μ_t. The dynamic DP-SGD, a combination of the two, improves performance while causing no additional privacy loss, as demonstrated by the Dynamic DP results in each table.

Experiments with different privacy budgets were conducted, and the results show that the stronger the privacy protection required (lower ϵ value), the more noticeable the improvement by the proposed dynamic DP-SGD method. For example, when $\epsilon = 0.4$ for the MNIST, even though the noise decay is adopted by Yu et al. (2019) and Zhang et al. (2021b), however, their model fail to learn due to large calibrated noise power by the loose DP compositions. In contrast, our method outperforms GDP method by a large margin, achieving 3.17% when $\epsilon = 0.4$ and even 4.6% for the LSTM network on

[2]The code will be available at github.com/dynamic-dp once the paper is accepted.

NAME. Without compromising privacy, our method consistently outperforms all other benchmarks on MNIST, FashionMNIST, IMDN, and NAME datasets for CNN, MLP, LSTM, and federated learning models, as shown in Table1-4.

Table 1: CNN on MNIST and FashionMNIST datasets.

DP Accountant	Dynamic Noise	MNIST			FashionMNIST		
		ϵ = 0.4	ϵ =0.6	ϵ =1.2	ϵ =0.4	ϵ =1.2	ϵ =2.0
Non-private	-	98.83			87.92		
ρ-zCDP (Yu et al., 2019)	Noise Decay	10.28	10.12	65.33	9.86	63.30	72.18
tCDP (Zhang et al., 2021a)	Noise Decay	26.93	83.28	92.60	53.69	76.48	77.58
CLT for GDP (Bu et al., 2020)	-	91.18	93.80	95.50	76.77	80.45	82.55
Extended CLT for GDP (Ours)	growing-μ_t	91.67	94.49	96.06	77.81	80.95	83.10
	Sensitivity Decay	93.95	95.17	96.17	78.11	82.83	83.64
	Dynamic DP	**94.35**	**95.21**	**96.34**	**78.50**	**83.22**	**83.81**

Table 2: MLP on the IMDB dataset.

DP Framework	Dynamic Noise	IMDB				
		ϵ = 0.5	ϵ =1	ϵ =3	ϵ =6	ϵ =9
Non-private	-	82.85				
tCDP (Zhang et al., 2021a)	Noise Decay	56.67	58.24	62.15	65.88	70.16
CLT for GDP (Bu et al., 2020)	-	63.62	69.71	75.64	77.75	78.60
Extended CLT for GDP(Ours)	growing-μ_t	64.92	69.85	76.00	78.16	78.56
	Sensitivity Decay	65.44	70.25	76.23	78.47	79.42
	Dynamic GDP	**65.63**	**70.77**	**76.64**	**78.61**	**79.61**

Table 3: LSTM on NAME dataset.

DP Framework	Dynamic Noise	NAME			
		ϵ = 1	ϵ =2	ϵ =4	ϵ =8
Non-private	-	80.14			
tCDP(Zhang et al., 2021a)	Noise Decay	48.13	51.20	57.67	68.53
CLT for GDP (Bu et al., 2020)	-	62.71	69.64	73.04	74.15
Extended CLT for GDP(Ours)	growing-μ_t	64.10	71.25	74.68	75.50
	Sensitivity Decay	66.66	71.78	73.67	74.78
	Dynamic DP	**67.30**	**72.01**	**75.03**	**75.75**

Table 4: Federated learning on InfiniteMNIST dataset.

DP Framework	Dynamic Noise	MNIST-250K			MNIST-500K		
		ϵ = 0.1	ϵ =0.4	ϵ =1	ϵ =0.1	ϵ =0.4	ϵ = 1
Non-private	-	98.89			98.96		
CLT for GDP (Bu et al., 2020)	-	93.47	95.71	96.02	94.82	96.55	96.89
Extended CLT for GDP(Ours)	growing-μ_t	93.75	95.93	96.13	95.65	96.76	97.05
	Sensitivity Decay	94.46	95.90	96.40	95.75	96.83	97.06
	Dynamic DP	**94.72**	**96.00**	**96.55**	**95.88**	**96.95**	**97.22**

5.3 HYPER-PARAMETER SENSITIVITY

We then test the robustness of dynamic DP-SGD performance to different values of ρ_μ and ρ_c as shown in Fig. 3. We use grid search to demonstrate the impact of these parameters on model performance. The dynamic method can consistently improve model performance across a wide range.

5.4 EXACT PRIVACY COST

Though Dong et al. (2019) have shown that the CLT of GDP approximate the true privacy cost with negligible error, Gopi et al. (2021) recently discovere that the CLT of GDP may underestimate the privacy cost. The RDP accountant (Wang et al., 2019), on the other hand, overestimates the true cost. To evaluate the exact privacy cost, thereby, we plot the privacy cost curves for both the proposed extended CLT of GDP and RDP in Fig. 4. Specifically, we set a target training round and conduct

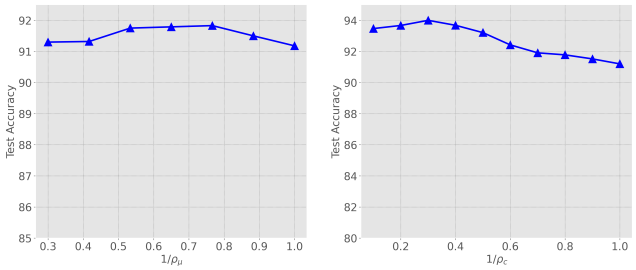

Figure 3: Dynamic DP-SGD performance is robust to ρ_μ and ρ_c.

privacy accounting for the dynamic DP-SGD based on the extended CLT of GDP as well as RDP[3] with the advanced translation method between RDP and (ϵ, δ)-DP (Balle et al., 2020). Consequently, the true privacy cost curve must lie somewhere between these two limits. The result of extended CLT for GDP is reasonable because the privacy cost differences between these two limits are small, particularly in the high privacy protection region.

It is worth noting that, in addition to the proposed extended CLT for dynamic DP-SGD accounting, exact accounting can be performed using the recently proposed methods by Gopi et al. (2021) and Zhu et al. (2021), separately. However, because they both require numerical computation, the computation of μ_0 in (15) becomes much involved.

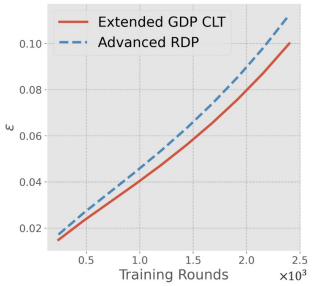

Figure 4: Upper bound (advanced RDP) and lower bound (extended GDP CLT) of the dynamic DP-SGD privacy budget cost curves. The true privacy cost curve must lie somewhere between these two limits. The result of extended CLT for GDP is reasonable because the privacy cost differences between these two limits are small, especially in the high privacy protection region.

6 Conclusions

In this paper, we extend the central limit theorem (CLT) of Gaussian DP to perform tight privacy accounting in order to calibrate dynamic noise for each individual step of stochastic gradient descent (SGD) updates. We, therefore, are able to allocate a lower privacy cost than the DP-SGD during updates based on this extended CLT until both methods consume the same target privacy budget at the predefined update number. Extensive testing on a variety of datasets and models demonstrates that the dynamic DP-SGD consistently and clearly outperforms existing methods.

For the future work, we intend to investigate how to leverage the recently achievement of numerical composition methods (Gopi et al., 2021; Zhu et al., 2021) to calibrate the noise power and clipping value for each individual training step of dynamic DP-SGD.

References

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Edgar R. Weippl, Stefan Katzenbeisser,

[3]For RDP, we use autodp library by Wang for DP accounting.

Christopher Kruegel, Andrew C. Myers, and Shai Halevi (eds.), *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 308–318. ACM, 2016.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), 2020.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J. Su. Deep learning with Gaussian differential privacy. *Harvard data science review*, 2020 23, 2020.

Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (eds.), *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380, 2009.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *arXiv preprint arXiv:2106.02848*, 2021.

Y. Lecun and L. Bottou. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, pp. 263–275. IEEE Computer Society, 2017.

Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pp. 157–175. Springer, 2016.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *ArXiv*, abs/1908.07643, 2019.

Om Thakkar, Galen Andrew, and H. Brendan McMahan. Differentially private learning with adaptive clipping. *CoRR*, abs/1905.03871, 2019.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019.

Yuxiang Wang. autodp: A flexible and easy-to-use package for differential privacy. `https://github.com/yuxiangw/autodp`. Accessed: 2021-09-30.

Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. Differentially private model publishing for deep learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 332–349. IEEE, 2019.

Xinyue Zhang, Jiahao Ding, Maoqiang Wu, Stephen T. C. Wong, Hien Van Nguyen, and Miao Pan. Adaptive privacy preserving deep learning algorithms for medical data. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pp. 1168–1177. IEEE, 2021a.

Xinyue Zhang, Jiahao Ding, Maoqiang Wu, Stephen TC Wong, Hien Van Nguyen, and Miao Pan. Adaptive privacy preserving deep learning algorithms for medical data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1169–1178, 2021b.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *ArXiv*, abs/2001.02610, 2020.

Ligeng Zhu and Song Han. Deep leakage from gradients. In Qiang Yang, Lixin Fan, and Han Yu (eds.), *Federated Learning - Privacy and Incentive*, volume 12500 of *Lecture Notes in Computer Science*, pp. 17–31. Springer, 2020.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. *arXiv preprint arXiv:2106.08567*, 2021.