Deep Learning with Gaussian Differential Privacy

Zhiqi Bu* Jinshuo Dong† Qi Long‡ Weijie J. Su§

University of Pennsylvania

November 25, 2019

Abstract

Deep learning models are often trained on datasets that contain sensitive information such as individuals' shopping transactions, personal contacts, and medical records. An increasingly important line of work therefore has sought to train neural networks subject to privacy constraints that are specified by differential privacy or its divergence-based relaxations. These privacy definitions, however, have weaknesses in handling certain important primitives (composition and subsampling), thereby giving loose or complicated privacy analyses of training neural networks. In this paper, we consider a recently proposed privacy definition termed *f-differential privacy* [17] for a refined privacy analysis of training neural networks. Leveraging the appealing properties of f-differential privacy in handling composition and subsampling, this paper derives analytically tractable expressions for the privacy guarantees of both stochastic gradient descent and Adam used in training deep neural networks, without the need of developing sophisticated techniques as [3] did. Our results demonstrate that the f-differential privacy framework allows for a new privacy analysis that improves on the prior analysis [3], which in turn suggests tuning certain parameters of neural networks for a better prediction accuracy without violating the privacy budget. These theoretically derived improvements are confirmed by our experiments in a range of tasks in image classification, text classification, and recommender systems.

1 Introduction

In many applications of machine learning, the datasets contain sensitive information about individuals such as location, personal contacts, media consumption, and medical records. Exploiting the output of the machine learning algorithm, an adversary may be able to identify some individuals in the dataset, thus presenting serious privacy concerns. This reality gave rise to a broad and pressing call for developing privacy-preserving data analysis methodologies. Accordingly, there have been numerous investigations in the scholarly literature of many fields—statistics, cryptography, machine learning, and law—for the protection of privacy in data analysis.

Along this line, research efforts have repeatedly suggested the necessity of a *rigorous* and *versatile* definition of privacy. Among other things, researchers have questioned whether the use of a privacy definition gives interpretable privacy guarantees, and if so, whether this privacy definition allows for high accuracy of the private model among alternative definitions. In particular,

*Graduate Group in Applied Mathematics and Computational Science. Email: zbu@sas.upenn.edu.

†Graduate Group in Applied Mathematics and Computational Science. Email: jinshuo@sas.upenn.edu.

‡Department of Biostatistics, Epidemiology and Informatics. Email: qlong@pennmedicine.upenn.edu.

§Department of Statistics. Email: suw@wharton.upenn.edu.

anonymization as a syntactic and *ad-hoc* privacy concept has been shown to generally fail to guarantee privacy. Examples include the identification of a homophobic individual in the anonymized Netflix Challenge dataset [43] and the identification of the health records of the then Massachusetts governor in public anonymized medical datasets [54].

In this context, (ε, δ)-*differential privacy* (DP) arose as a mathematically rigorous definition of privacy [22]. Today, this definition has developed into a firm foundation of private data analysis, with its applications deployed by Google [25], Apple [5], Microsoft [15], and the US Census Bureau [4]. Despite its impressive popularity in both the scholarly literature and the industry, (ε, δ)-DP is not versatile enough to handle *composition*, which is perhaps the most fundamental primitive in statistical privacy. For example, the training process of deep neural networks is in effect the composition of many primitive building blocks known as stochastic gradient descent (SGD). Under a modest privacy budget in the (ε, δ)-DP sense, however, it was not clear how to maintain a high prediction accuracy of deep learning. This requires a tight privacy analysis of composition in the (ε, δ)-DP framework. Indeed, the analysis of the privacy costs in deep learning was refined only recently using a sophisticated technique called the moments accountant [3].

Ideally, we hope to have a privacy definition that allows for refined privacy analyses of various algorithms in a principled manner, *without* resorting to sophisticated techniques. Having a refined privacy analysis not only enhances the trustworthiness of the models but can also be leveraged to improve the prediction accuracy by trading off privacy for utility. One possible candidate is f-*differential privacy*, a relaxation of (ε, δ)-DP that was recently proposed by Dong, Roth, and Su [17]. This new privacy definition faithfully retains the hypothesis testing interpretation of differential privacy and can losslessly reason about common primitives associated with differential privacy, including composition, privacy amplification by subsampling, and group privacy. In addition, f-DP includes a canonical single-parameter family that is referred to as *Gaussian differential privacy* (GDP). Notably, GDP is the focal privacy definition due to a central limit theorem that states that the privacy guarantees of the composition of private algorithms are approximately equivalent to telling apart two shifted normal distributions.

The main results of this paper show that f-DP offers a rigorous and versatile framework for developing private deep learning methodologies[1]. Our guarantee provides protection against an attacker with knowledge of the network architecture as well as the model parameters, which is in the same spirit as [51, 3]. In short, this paper delivers the following messages concerning f-DP:

> **Closed-form privacy bounds.** In the f-DP framework, the overall privacy loss incurred in training neural networks admits an amenable closed-form expression. In contrast, the privacy analysis via the moments accountant must be done by numerical computation [3], and the implicit nature of this earlier approach can hinder our understanding of how the tuning parameters affect the privacy bound. This is discussed in Section 3.1.

> **Stronger privacy guarantees.** The f-DP approach gives stronger privacy guarantees than the earlier approach [3], even in terms of (ε, δ)-DP. This improvement is due to the use of the central limit theorem for f-DP, which accurately captures the privacy loss incurred at each iteration in training the deep learning models. This is presented in Section 3.2 and illustrated with numerical experiments in Section 4.1.

[1] It is noteworthy that training deep learning models has served as an important benchmark in testing a privacy definition since the tightness of its privacy analysis crucially depends on whether the definition can tightly account for composition and subsampling.

Improved prediction accuracy. Leveraging the stronger privacy guarantees provided by f-DP, we can trade a certain amount of privacy for an improvement in prediction performance. This can be realized, for example, by appropriately reducing the amount of noise added during the training process of neural networks so as to match the target privacy level in terms of (ε, δ)-DP. See Section 3.2 and Section 4.2 for the development of this utility improvement.

The remainder of the paper is structured as follows. In Section 1.1 we provide a brief review of related literature. Section 2 introduces f-DP and its basic properties at a minimal level. Next, in Section 3 we analyze the privacy cost of training deep neural networks in terms of f-DP and compare it to the privacy analysis using the moments accountant. In Section 4, we present numerical experiments to showcase the superiority of the f-DP approach to private deep learning in terms of test accuracy and privacy guarantees. The paper concludes with a discussion in Section 5.

1.1 Related Work

There are continued efforts to understand how privacy degrades under composition. Developments along this line include the basic composition theorem and the advanced composition theorem [21, 23]. In a pioneering work, [31] obtained an optimal composition theorem for (ε, δ)-DP, which in fact served as one of the motivations for the f-DP work [17]. However, it is #P hard to compute the privacy bounds from their composition theorem [42]. More recently, [16] derived sharp composition bounds on the overall privacy loss for exponential mechanisms.

From a different angle, a substantial recent effort has been devoted to relaxing differential privacy using divergences of probability distributions to overcome the weakness of (ε, δ)-DP in handling composition [20, 10, 41, 11]. Unfortunately, these relaxations either lack a privacy amplification by subsampling argument or present a quite complex argument that is difficult to use [7, 55]. As subsampling is inherently used in training neural networks, therefore, it is difficult to directly apply these relaxations to the privacy analysis of deep learning.

To circumvent these technical difficulties associated with (ε, δ)-DP and its divergence-based relaxations, Abadi et al. [3] invented a technique termed the moments accountant to track detailed information of the privacy loss in the training process of deep neural networks. Using the moments accountant, their analysis significantly improves on earlier privacy analysis of SGD [13, 53, 9, 51, 58] and allows for meaningful privacy guarantees for deep learning trained on realistically sized datasets. This technique has been extended to a variety of situations by follow-up work [39, 46]. In contrast, our approach to private deep learning in the f-DP framework leverages some powerful tools of this new privacy definition, nevertheless providing a sharper privacy analysis, as seen both theoretically and empirically in Sections 3 and 4.

For completeness, we remark that different approaches have been put forward to incorporate privacy considerations into deep learning, without leveraging the iterative and subsampling natures of training deep learning models. This line of work includes training a private model by an ensemble of "teacher" models [45, 44], the development of noised federated averaging algorithms [40], and analyzing privacy costs through the lens of the optimization landscape of neural networks [57].

2 Preliminaries

2.1 f-Differential Privacy

In the differential privacy framework, we envision an adversary that is well-informed about the dataset except for a single individual, and the adversary seeks to determine whether this individual is in the dataset on the basis of the output of an algorithm. Roughly speaking, the algorithm is considered private if the adversary finds it hard to determine the presence or absence of any individual.

Informally, a dataset can be thought of as a matrix, whose rows each contain one individual's data. Two datasets are said to be *neighbors* if one can be derived by discarding an individual from the other. As such, the sizes of neighboring datasets differ by one[2]. Let S and S' be neighboring datasets, and $\varepsilon \geqslant 0, 0 \leqslant \delta \leqslant 1$ be two numbers, and denote by M a (randomized) algorithm that takes as input a dataset.

Definition 2.1 ([22, 21]). A (randomized) algorithm M gives (ε, δ)-differential privacy if for any pair of neighboring datasets S, S' and any event E,

$$\mathbb{P}(M(S) \in E) \leqslant \mathrm{e}^{\varepsilon}\mathbb{P}(M(S') \in E) + \delta.$$

To achieve privacy, the algorithm M is necessarily randomized, whereas the two datasets in Definition 2.1 are *deterministic*. This privacy definition ensures that, based on the output of the algorithm, the adversary has a limited (depending on how small ε, δ are) ability to identify the presence or absence of any individual, regardless of whether any individual opts in to or opts out of the dataset.

In essence, the adversary seeks to tell apart the two probability distributions $M(S)$ and $M(S')$ using a single draw. In light of this observation, it is natural to interpret what the adversary does as testing two simple hypotheses:

$$H_0 : \text{the true dataset is } S \quad \text{versus} \quad H_1 : \text{the true dataset is } S'.$$

The connection between differential privacy and hypothesis testing was, to our knowledge, first noted in [56], and was later developed in [31, 37, 8]. Intuitively, privacy is well guaranteed if the hypothesis testing problem is *hard*. Following this intuition, the definition of (ε, δ)-DP essentially uses the *worst-case* likelihood ratio of the distributions $M(S)$ and $M(S')$ to measure the hardness of testing the two simple hypotheses.

Is there a more *informative* measure of the hardness? In [17], the authors propose to use the *trade-off* between type I error and type II error in place of a few privacy parameters in (ε, δ)-DP or divergence-based DP definitions. To formally define this new privacy definition, let P and Q denote the distributions of $M(S)$ and $M(S')$, respectively, and let ϕ be any (possibly randomized) rejection rule for testing $H_0 : P$ against $H_1 : Q$. With these in place, [17] defines the *trade-off function* of P and Q as

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

[2] Alternatively, the neighboring relationship can be defined for datasets of the same size and differing by one individual.

Above, $\mathbb{E}_P[\phi]$ and $1 - \mathbb{E}_Q[\phi]$ are type I and type II errors of the rejection rule ϕ, respectively. Writing $f = T(P, Q)$, the definition says that $f(\alpha)$ is the minimum type II error among all tests at significance level α. Note that the minimum can be achieved by taking the likelihood ratio test, according to the Neymann–Pearson lemma. As is self-evident, the larger the trade-off function is, the more difficult the hypothesis testing problem is (hence more privacy). This motivates the following privacy definition.

Definition 2.2 ([17]). A (randomized) algorithm M is f-differentially private if

$$T\big(M(S), M(S')\big) \geqslant f$$

for all neighboring datasets S and S'.

In this definition, the inequality holds pointwise for all $0 \leqslant \alpha \leqslant 1$, and we abuse notation by identifying $M(S)$ and $M(S')$ with their associated distributions. This privacy definition is easily interpretable due to its inherent connection with the hypothesis testing problem. By adapting a result due to Wasserman and Zhou [56], (ε, δ)-DP is a special instance of f-DP in the sense that an algorithm is (ε, δ)-DP if and only if it is $f_{\varepsilon,\delta}$-DP with

$$f_{\varepsilon,\delta}(\alpha) = \max\left\{0, 1 - \delta - \mathrm{e}^{\varepsilon}\alpha, \mathrm{e}^{-\varepsilon}(1 - \delta - \alpha)\right\}. \tag{1}$$

The more intimate relationship between the two privacy definitions is that they are *dual* to each other: briefly speaking, f-DP ensures $(\varepsilon, \delta(\varepsilon))$-DP with $\delta(\varepsilon) = 1 + f^*(-\mathrm{e}^{\varepsilon})$ for every $\varepsilon \geqslant 0$[3].

Next, we define a single-parameter family of privacy definitions within the f-DP class for a reason that will be apparent later. Let $G_\mu := T\big(\mathcal{N}(0,1), \mathcal{N}(\mu, 1)\big)$ for $\mu \geqslant 0$. Note that this trade-off function admits a closed-form expression $G_\mu(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \mu)$, where Φ is the cumulative distribution function of the standard normal distribution.

Definition 2.3 ([17]). A (randomized) algorithm M is μ-Gaussian differentially private (GDP) if

$$T\big(M(S), M(S')\big) \geqslant G_\mu$$

for all neighboring datasets S and S'.

In words, μ-GDP says that determining whether any individual is in the dataset is at least as difficult as telling apart the two normal distributions $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$ based on one draw. The Gaussian mechanism serves as a template to achieve GDP. Consider the problem of privately releasing a univariate statistic $\theta(S)$. The Gaussian mechanism adds $\mathcal{N}(0, \sigma^2)$ noise to the statistic θ, which gives μ-GDP if $\sigma = \mathrm{sens}(\theta)/\mu$. Here the *sensitivity* of θ is defined as $\mathrm{sens}(\theta) = \sup_{S,S'} |\theta(S) - \theta(S')|$, where the supremum is over all neighboring datasets.

2.2 Properties of f-Differential Privacy

Composition. Deep learning models are trained using the *composition* of many SGD updates. Broadly speaking, composition is concerned with a sequence of analyses on the *same* dataset where each analysis is informed by the explorations of prior analyses. A central question that every privacy definition is faced with is to pinpoint how the overall privacy guarantee degrades under composition. Formally, letting M_1 be the first algorithm and M_2 be the second, we define their composition algorithm M as $M(S) = (M_1(S), M_2(S, M_1(S)))$. Roughly speaking, the composition

[3]Here, f^* is the convex conjugate, which is defined as $f^*(x) = \sup_\alpha \alpha x - f(\alpha)$.

is to "release all information that is learned by the algorithms." Notably, the second algorithm M_2 can take as input the output of M_1 in addition to the dataset S. In general, the composition of more than two algorithms follows recursively.

To introduce the composition theorem for f-DP, [17] defines a binary operation $\otimes$ on trade-off functions. Given trade-off functions $f = T(P, Q)$ and $g = T(P', Q')$, let $f \otimes g = T(P \times P', Q \times Q')$. This definition depends on the distributions P, Q, P', Q' only through f and g. Moreover, $\otimes$ is commutative and associative. Now the composition theorem can be readily stated as follows. Let M_t be f_t-DP *conditionally* on any output of the prior algorithms for $t = 1, \ldots, T$. Then their T-fold composition algorithm is $f_1 \otimes \cdots \otimes f_T$-DP. This result shows that the composition of algorithms in the f-DP framework is reduced to performing the $\otimes$ operation on the associated trade-off functions. As an important fact, the privacy bound $f_1 \otimes \cdots \otimes f_T$ in general cannot be improved.

More profoundly, a *central limit theorem* phenomenon arises in the composition of many "very private" f-DP algorithms in the following sense: the trade-off functions of small privacy leakage accumulate to G_μ for some μ under composition. Informally, assuming each f_t is very close to $\mathrm{Id}(\alpha) = 1 - \alpha$, which corresponds to perfect privacy, then we have

$$f_1 \otimes f_2 \otimes \cdots \otimes f_T \text{ is approximately } G_\mu \tag{2}$$

if T is sufficiently large. The privacy parameter μ depends on some functionals such as the Kullback–Leibler divergence of the trade-off functions. The central limit theorem yields a very accurate approximation in the settings considered in Section 4 (see numerical confirmation in Appendix A). For a rigorous account of this central limit theorem for differential privacy, see Theorem 3.5 in [17]. We remark that a conceptually related article [52] developed a central limit theorem for privacy loss random variables.

At a high level, this convergence-to-GDP result brings GDP to the focal point of the family of f-DP guarantees, implying that GDP is to f-DP as normal random variables to general random variables. Furthermore, this result serves as an effective approximation tool for approximating the privacy guarantees of composition algorithms. In contrast, privacy loss cannot be losslessly tracked under composition in the (ε, δ)-DP framework.

Subsampling. In training neural networks, the gradient at each iteration is computed from a mini-batch that is *subsampled* from the training examples. Intuitively, an algorithm applied to a subsample gives stronger privacy guarantees than applied to the full sample. Looking closely, this privacy amplification is due to the fact that an individual enjoys perfect privacy if not selected in the subsample. A concrete and pressing question is, therefore, to precisely characterize how much privacy is amplified by subsampling in the f-DP framework.

Consider the following sampling scheme: for each individual in the dataset S, include his or her datum in the subsample independently with probability p, which is sometimes referred to as the Poisson subsampling [55]. The resulting subsample is denoted by $\texttt{Sample}_p(S)$. For the purpose of clearing up any confusion, we remark that the subsample $\texttt{Sample}_p(S)$ has a random size and as an intermediate step is not released. Given any algorithm M, denote by $M \circ \texttt{Sample}_p$ the subsampled algorithm.

The subsampling theorem for f-DP states as follows. Let M be f-DP, write f_p for $pf + (1-p)\mathrm{Id}$, and denote by f_p^{-1} the inverse[4] of f_p. It is proved in Appendix A that the subsampled algorithm

[4]For any trade-off function $f = T(P, Q)$, its inverse $f^{-1} = T(Q, P)$.

$M \circ \texttt{Sample}_p$ satisfies

$$T\big(M \circ \texttt{Sample}_p(S), M \circ \texttt{Sample}_p(S')\big) \geqslant f_p \tag{3}$$

if S can be obtained by removing one individual from S'. Likewise,

$$T\big(M \circ \texttt{Sample}_p(S'), M \circ \texttt{Sample}_p(S)\big) \geqslant f_p^{-1}.$$

As such, the two displays above say that the trade-off function of $M \circ \texttt{Sample}_p$ on any neighboring datasets is lower bounded by $\min\{f_p, f_p^{-1}\}$, which however is in general non-convex and thus is not a trade-off function. This suggests that we can boost the privacy bound by replacing $\min\{f_p, f_p^{-1}\}$ with its double conjugate $\min\{f_p, f_p^{-1}\}^{**}$, which is the greatest convex lower bound of $\min\{f_p, f_p^{-1}\}$ and is indeed a trade-off function. Taken together, all the pieces show that the subsampled algorithm $M \circ \texttt{Sample}_p$ is $\min\{f_p, f_p^{-1}\}^{**}$-DP.

Notably, the privacy bound $\min\{f_p, f_p^{-1}\}^{**}$ is larger than f and cannot be improved in general. In light of the above, the f-DP framework is flexible enough to nicely handle the analysis of privacy amplification by subsampling. In the case where the original algorithm M is (ε, δ)-DP, this privacy bound strictly improves on the subsampling theorem for (ε, δ)-DP [36].

3 Algorithms and Their Privacy Analyses

3.1 NoisySGD and NoisyAdam

SGD and Adam [32] are among the most popular optimizers in deep learning. Here we introduce a new privacy analysis of a private variant of SGD in the f-DP framework and then extend the study to a private version of Adam.

Letting $S = \{x_1, \ldots, x_n\}$ denote the dataset, we consider minimizing the empirical risk

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i),$$

where θ denotes the weights of the neural networks and $\ell(\theta, x_i)$ is a loss function. At iteration t, a mini-batch I_t is selected from $\{1, 2, \ldots, n\}$ with subsampling probability p, thereby having an approximate size of pn. Taking learning rate η_t and initial weights θ_0, the vanilla SGD updates the weights according to

$$\theta_{t+1} = \theta_t - \eta_t \cdot \frac{1}{|I_t|} \sum_{i \in I_t} \nabla_\theta \ell(\theta_t, x_i).$$

To preserve privacy, [13, 53, 9, 3] introduce two modifications to the vanilla SGD. First, a clip step is applied to the gradient so that the gradient is in effect bounded. This step is necessary to have a finite sensitivity. The second modification is to add Gaussian noise to the clipped gradient, which is equivalent to applying the Gaussian mechanism to the updated iterates. Formally, the private SGD algorithm is described in Algorithm 1. Herein I is the identity matrix and $\|\cdot\|_2$ denotes the ℓ_2 norm. Formally, we present this in Algorithm 1, which we henceforth refer to as `NoisySGD` and uses the Poisson subsampling. For completeness, we remark that there are two other possible subsampling methods: shuffling (randomly permuting and dividing data into folds at each epoch) and uniform sampling (sampling a batch of size L from the whole data at each iteration). We emphasize that different subsampling mechanisms produce different privacy guarantees.

Algorithm 1 `NoisySGD`

Input: Dataset $S = \{x_1, \dots, x_n\}$, loss function $\ell(\theta, x)$.
Parameters: initial weights θ_0, learning rate η_t, subsampling probability p, number of iterations T, noise scale σ, gradient norm bound R.

for $t = 0, \dots, T-1$ **do**
Take a Poisson subsample $I_t \subseteq \{1, \dots, n\}$ with subsampling probability p
for $i \in I_t$ **do**
$v_t^{(i)} \leftarrow \nabla_\theta \ell(\theta_t, x_i)$
$\bar{v}_t^{(i)} \leftarrow v_t^{(i)} / \max\left\{1, \|v_t^{(i)}\|_2 / R\right\}$ ▷ Clip gradient
$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \frac{1}{|I_t|}\Big(\sum_{i \in I_t} \bar{v}_t^{(i)} + \sigma R \cdot \mathcal{N}(0, I)\Big)$ ▷ Apply Gaussian mechanism

Output θ_T

The analysis of the overall privacy guarantee of `NoisySGD` makes heavy use of the compositional and subsampling properties of f-DP. We first focus on the privacy analysis of the step that computes θ_{t+1} from θ_t. Let M denote the gradient update and write $\texttt{Sample}_p(S)$ for the mini-batch I_t (we drop the subscript t for simplicity). This allows us to use $M \circ \texttt{Sample}_p(S)$ to represent what `NoisySGD` does at each iteration. Next, note that adding or removing one individual would change the value of $\sum_{i \in I_t} \bar{v}_t^{(i)}$ by at most R in the ℓ_2 norm due to the clipping operation, that is, $\sum_{i \in I_t} \bar{v}_t^{(i)}$ has sensitivity R. Consequently, the Gaussian mechanism with noise standard deviation σR ensures that M is $\frac{1}{\sigma}$-GDP. With a few additional arguments, in Appendix B we show that `NoisySGD` is $\min\{f, f^{-1}\}^{**}$-DP with $f = \big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T}$.

To facilitate the use of this privacy bound, we now derive an analytically tractable approximation of $\min\{f, f^{-1}\}^{**}$ using the privacy central limit theorem in a certain asymptotic regime, which further demonstrates the mathematical coherence and versatility of the f-DP framework. The central limit theorem shows that, in the asymptotic regime where $p\sqrt{T} \to \nu$ for a constant $\nu > 0$ as $T \to \infty$,

$$f = \big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T} \to G_\mu,$$

where $\mu = \nu\sqrt{\mathrm{e}^{1/\sigma^2} - 1}$. Thus, the overall privacy loss in the form of the double conjugate satisfies

$$\min\{f, f^{-1}\}^{**} \approx \min\{G_\mu, G_\mu^{-1}\}^{**} = G_\mu^{**} = G_\mu. \tag{4}$$

As such, the central limit theorem demonstrates that `NoisySGD` is approximately $p\sqrt{T(\mathrm{e}^{1/\sigma^2} - 1)}$-GDP. Denoting by $B = pn$ the mini-batch size, the privacy parameter $p\sqrt{T(\mathrm{e}^{1/\sigma^2} - 1)}$ equals $\frac{B}{n}\sqrt{T(\mathrm{e}^{1/\sigma^2} - 1)}$. Intuitively, this reveals that `NoisySGD` gives good privacy guarantees if $B\sqrt{T}/n$ is small and σ is not too small.

As an aside, we remark that this new privacy analysis is different from the one performed in Section 5 of [17]. Therein, the authors consider Algorithm 1 with uniform subsampling and obtain a privacy bound that is different from the one in the present paper.

Next, we present a private version of Adam [32] in Algorithm 2, which we refer to as `NoisyAdam` and can be found in [2]. This algorithm has the same privacy bound as `NoisySGD` in the f-DP framework. In short, this is because the momentum m_t and u_t are deterministic functions of the noisy gradients and no additional privacy cost is incurred due to the post-processing property of

differential privacy. In passing, we remark that the same argument applies to AdaGrad [19] and therefore it is also asymptotically GDP in the same asymptotic regime.

Algorithm 2 `NoisyAdam`

Input: Dataset $S = \{x_1, \ldots, x_n\}$, loss function $\ell(\theta, x)$.
Parameters: initial weights θ_0, learning rate η_t, subsampling probability p, number of iterations T, noise scale σ, gradient norm bound R, momentum parameters (β_1, β_2), initial momentum m_0, initial past squared gradient u_0, and a small constant $\xi > 0$.

for $t = 0, \ldots, T-1$ **do**
  Take a Poisson subsample $I_t \subseteq \{1, \ldots, n\}$ with subsample probability p
  for $i \in I_t$ **do**
    $v_t^{(i)} \leftarrow \nabla_\theta \ell(\theta_t, x_i)$
    $\bar{v}_t^{(i)} \leftarrow v_t^{(i)} / \max\left\{1, \|v_t^{(i)}\|_2 / R\right\}$ ▷ Clip gradient
  $\tilde{v}_t \leftarrow \frac{1}{|I_t|}\left(\sum_{i \in I_t} \bar{v}_t^{(i)} + \sigma R \cdot \mathcal{N}(0, I)\right)$ ▷ Apply Gaussian mechanism
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1)\tilde{v}_t$
  $u_t \leftarrow \beta_2 u_{t-1} + (1-\beta_2)\left(\tilde{v}_t \odot \tilde{v}_t\right)$ ▷ $\odot$ is the Hadamard product
  $w_t \leftarrow m_t / \left(\sqrt{u_t} + \xi\right)$ ▷ Component-wise division
  $\theta_{t+1} \leftarrow \theta_t - \eta_t w_t$

Output θ_T

3.2 Comparisons with the Moments Accountant

It is instructive to compare the moments accountant with our privacy analysis performed in Section 3.1 using the f-DP framework. Developed in [3], the moments accountant gives a tight one-to-one mapping between ε and δ for specifying the overall privacy loss in terms of (ε, δ)-DP under composition, which is beyond the reach of the advanced composition theorem [23]. As abuse of notation, this paper uses functions $\delta_{\texttt{MA}} = \delta_{\texttt{MA}}(\varepsilon)$ and $\varepsilon_{\texttt{MA}} = \varepsilon_{\texttt{MA}}(\delta)$ to denote the mapping induced by the moments accountant in both directions[5]. For self-containedness, the appendix includes a formal description of the two functions.

Although `NoisySGD` and `NoisyAdam` are our primary focus, our following discussion applies to general iterative algorithms where composition must be addressed in the privacy analysis. Let algorithm M_t be f_t-DP for $t = 1, \ldots, T$ and write M for their composition. On the one hand, the moments accountant technique ensures that M is $(\varepsilon, \delta_{\texttt{MA}}(\varepsilon))$-DP for any ε or, put equivalently, is $(\varepsilon_{\texttt{MA}}, \delta)$-DP[6]. On the other hand, the composition algorithm is $f_1 \otimes \cdots \otimes f_T$-DP from the f-DP viewpoint and, following from the central limit theorem (2), this composition can be shown to be approximately GDP in a certain asymptotic regime. For example, both `NoisySGD` and `NoisyAdam` presented in Algorithm 1 and Algorithm 2, respectively, asymptotically satisfy $\mu_{\texttt{CLT}}$-GDP with

[5] We omit the dependence of the functions on the specification of the composition algorithm such as p, σ, T as in `NoisySGD` and `NoisyAdam`.

[6] The moments accountant can be applied in the f-DP framework. In fact, the moments accountant is defined via a certain moment generating function, which is equivalent to the Rényi divergence. The Rényi divergence can be uniquely deduced from a trade-off function. See Section 2.3 in [17].

privacy parameter

$$\mu_{\texttt{CLT}} = p\sqrt{T(\mathrm{e}^{1/\sigma^2} - 1)}. \tag{5}$$

In light of the above, it is tempting to ask which of the two approaches yields a sharper privacy analysis. In terms of f-DP guarantees, it must be the latter, which we refer to as the `CLT` approach, because the composition theorem of f-DP is tight[7] and, more importantly, the privacy central limit theorem is asymptotic exact. To formally state the result, note that the moments accountant asserts that the private optimizer is $(\varepsilon, \delta_{\texttt{MA}}(\varepsilon))$-DP for all $\varepsilon \geqslant 0$, which is equivalent to $\sup_{\varepsilon \geqslant 0} f_{\varepsilon, \delta_{\texttt{MA}}(\varepsilon)}$-DP by recognizing (1) (see also Proposition 2.11 in [17]). Roughly speaking, the following theorem says that $\sup_{\varepsilon \geqslant 0} f_{\varepsilon, \delta_{\texttt{MA}}(\varepsilon)}$-DP (asymptotically) promises no more privacy guarantees than the bound of $\mu_{\texttt{CLT}}$-GDP given by the `CLT` approach. This simple result is summarized by the following theorem and see Appendix B for a formal proof of this result.

Theorem 1 (Comparison in f-DP). Assume that $p\sqrt{T}$ converges to a positive constant as $T \to \infty$. Then, both `NoisySGD` and `NoisyAdam` satisfy

$$\limsup_{T \to \infty} \left(\sup_{\varepsilon \geqslant 0} f_{\varepsilon, \delta_{\texttt{MA}}(\varepsilon)}(\alpha) - G_{\mu_{\texttt{CLT}}}(\alpha) \right) \leqslant 0$$

for every $0 \leqslant \alpha \leqslant 1$.

Remark 1. For ease of reading, we point out that, in the (ε, δ)-DP framework, the smaller ε, δ are, the more privacy is guaranteed. In contrast, in the f-DP framework, the smaller f is, the less privacy is guaranteed.

From the (ε, δ)-DP viewpoint, however, the question is presently unclear. Explicitly, the duality between f-DP and (ε, δ)-DP shows that μ-GDP implies $(\varepsilon, \delta(\varepsilon; \mu))$-DP for all $\varepsilon \geqslant 0$, where[8]

$$\delta(\varepsilon; \mu) = 1 + G_\mu^*(-\mathrm{e}^\varepsilon) = \Phi(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}) - \mathrm{e}^\varepsilon \Phi(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}). \tag{6}$$

The question is, therefore, reduced to the comparison between $\delta_{\texttt{MA}}(\varepsilon)$ and $\delta_{\texttt{CLT}}(\varepsilon) := \delta(\varepsilon; \mu_{\texttt{CLT}})$ or, equivalently, between $\varepsilon_{\texttt{MA}}(\delta)$ and $\varepsilon_{\texttt{CLT}}(\delta) := \varepsilon(\delta; \mu_{\texttt{CLT}})$[9].

Theorem 2 (Comparison in (ε, δ)-DP). Under the assumptions of Theorem 1, the f-DP framework gives an asymptotically sharper privacy analysis of both `NoisySGD` and `NoisyAdam` than the moments accountant in terms of (ε, δ)-DP. That is,

$$\limsup_{T \to \infty} \left(\delta_{\texttt{CLT}}(\varepsilon) - \delta_{\texttt{MA}}(\varepsilon)\right) < 0$$

for all $\varepsilon \geqslant 0$.

In words, the `CLT` approach in the f-DP framework allows for an asymptotically smaller δ than the moments accountant at the same ε. It is worthwhile mentioning that the inequality in this theorem holds for any finite T if δ is derived by directly applying the duality to the (exact) privacy bound $f_1 \otimes \cdots \otimes f_T$. Equivalently, the theorem says that $\limsup_{T \to \infty} (\varepsilon_{\texttt{CLT}}(\delta) - \varepsilon_{\texttt{MA}}(\delta)) < 0$ for any δ[10]. As such, by setting the same δ in both approaches, say $\delta = 10^{-5}$, the f-DP based `CLT` approach shall give a smaller value of ε.

[7] See the discussion following Theorem 3.2 in [17].

[8] See Section 2.4 of [17] for this result. See also [24, 6].

[9] Here, $\varepsilon(\delta; \mu)$ is the inverse function of $\delta(\varepsilon; \mu)$.

[10] Write $\delta^\star_{\texttt{CLT}} = \delta_{\texttt{CLT}}(0)$ and set $\varepsilon_{\texttt{CLT}}(\delta) = 0$ for $\delta \geqslant \delta^\star_{\texttt{CLT}}$. Apply the same adjustment for $\varepsilon_{\texttt{MA}}$.

From a practical viewpoint, this refined privacy analysis allows us to trade privacy guarantees for improvement in utility. More precisely, recognizing the aforementioned conclusion that $\delta(\varepsilon; \mu_{\texttt{CLT}}) \equiv \delta_{\texttt{CLT}}(\varepsilon) < \delta_{\texttt{MA}}(\varepsilon)$ (for sufficiently large T) and that $\delta(\varepsilon; \mu)$ increases as μ increases, one can find $\tilde{\mu}_{\texttt{CLT}} > \mu_{\texttt{CLT}}$ such that

$$\delta(\varepsilon; \tilde{\mu}_{\texttt{CLT}}) = \delta_{\texttt{MA}}(\varepsilon). \tag{7}$$

Put differently, we can carefully adjust some parameters in Algorithm 1 and Algorithm 2 in order to let the algorithms be $\tilde{\mu}_{\texttt{CLT}}$-GDP. For example, we can *reduce* the scale of the added noise from σ to a certain $\tilde{\sigma} < \sigma$, which can be solved from (7) and

$$\tilde{\mu}_{\texttt{CLT}} = p\sqrt{T(\mathrm{e}^{1/\tilde{\sigma}^2} - 1)}. \tag{8}$$

Note that this is adapted from (5).

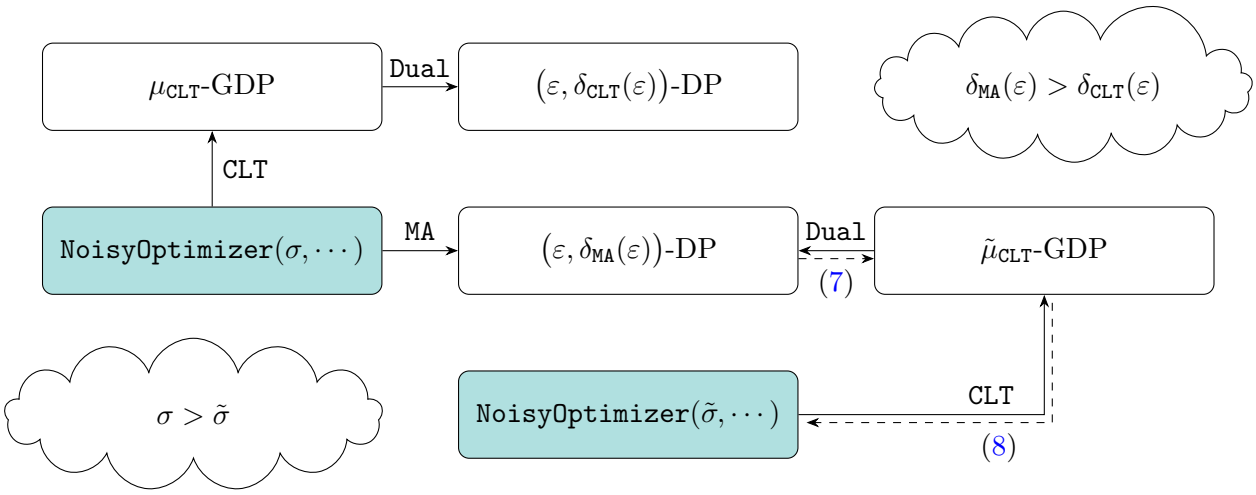

Figure 1: An illustration of the `CLT` approach in the f-DP framework and the moments accountant in the (ε, δ)-DP framework. `NoisyOptimizer`$(\sigma, \dots)$ using the moments accountant gives the same privacy guarantees in terms of (ε, δ)-DP as `NoisyOptimizer`$(\tilde{\sigma}, \dots)$ using the `CLT` approach (the ellipses denote omitted parameters). Note that the duality formula (6) is used in solving $\tilde{\mu}_{\texttt{CLT}}$ from (7).

Figure 1 shows the flowchart of the privacy analyses using the two approaches and their relationship. In addition, numerical comparisons are presented in Figure 2, consistently demonstrating the superiority of the `CLT` approach.

4 Results

In this section, we use `NoisySGD` and `NoisyAdam` to train private deep learning models on datasets for tasks ranging from image classification (MNIST), text classification (IMDb movie review), recommender systems (MovieLens movie rating), to regular binary classification (Adult income). Note that these datasets all contain sensitive information about individuals, and this fact necessitates privacy consideration in the training process. A git repository with code to reproduce the results is available at `https://github.com/woodyx218/Deep-Learning-with-GDP`.

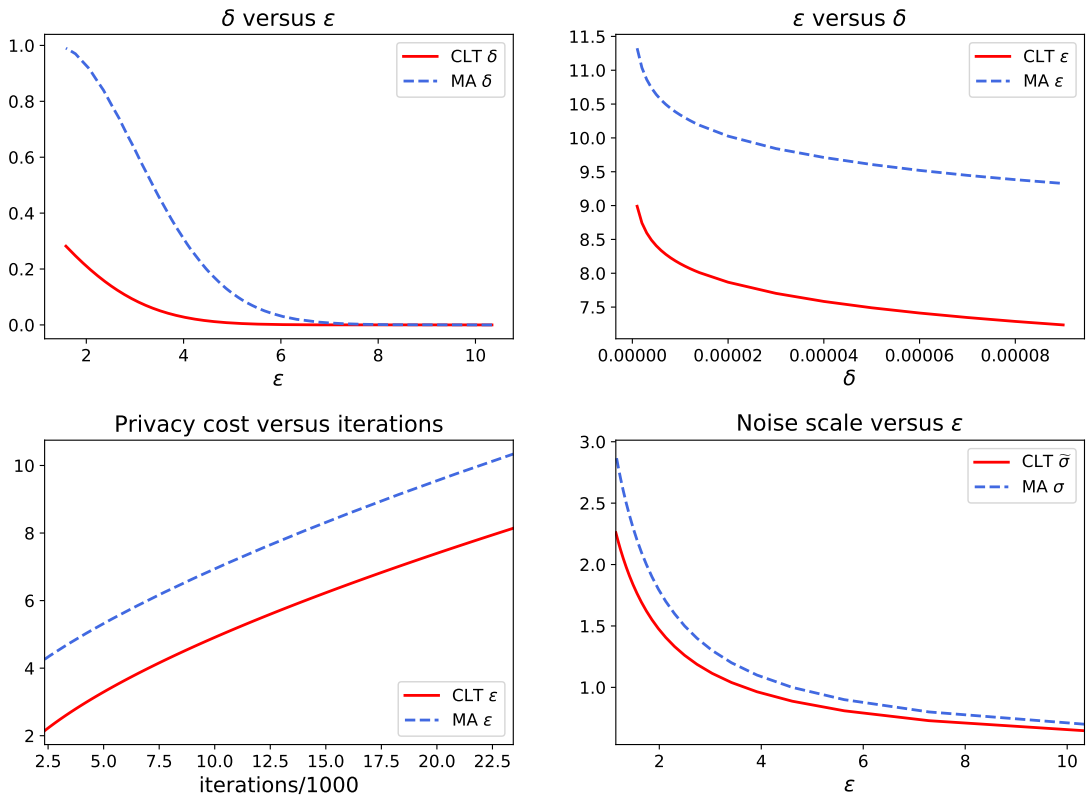

Figure 2: Tradeoffs between ε and δ for both `CLT` and `MA`, which henceforth denotes the moments accountant. The settings follows the MNIST experiment in Section 4 with $\sigma = 0.7, p = 256/60000$. The bottom two plots assume $\delta = 10^{-5}$. Note ε and δ in the `CLT` are related via (6) with $\mu = \mu_{\texttt{CLT}}$. The bottom right plot is consistent with the conclusion $\sigma > \tilde{\sigma}$ shown in the cloud icon of Figure 1.

4.1 The f-DP Perspective

This section demonstrates the utility and practicality of the private deep learning methodologies with associated privacy guarantees in terms of f-DP. In Section 4.2, we extend the empirical study to the (ε, δ)-DP framework.

MNIST. The MNIST dataset [34] contains 60,000 training images and 10,000 test images. Each image is in 28×28 gray-scale representing a handwritten digit ranging from 0 to 9. We train neural networks with the same architecture (two convolutional layers followed by one dense layer) as in [2, 3] on this dataset. Throughout the experiment, we set the subsampling probability to $p = 256/60000$ and use a constant learning rate η.

Table 1 displays the test accuracy of the neural networks trained by `NoisySGD` as well as the associated privacy analyses. The privacy parameters ε in the last two columns are both with respect to $\delta = 10^{-5}$. Over all six sets of experiments with different tuning parameters, the `CLT` approach gives a significantly smaller value of ε than the moments accountant, which is consistent with our discussion in Section 3.2. The point we wish to emphasize, however, is that f-DP offers a much more comprehensive interpretation of the privacy guarantees than (ε, δ)-DP. For instance,

η	R	σ	Epochs	Test accuracy (%)	CLT μ	CLT ε	MA ε
0.25	1.5	1.3	15	95.0	0.23	0.83	1.19
0.15	1.0	1.1	60	96.6	0.57	2.32	3.01
0.25	1.5	0.7	45	97.0	1.13	5.07	7.10
0.25	1.5	0.6	62	97.6	2.00	9.98	13.27
0.25	1.5	0.55	68	97.8	2.76	14.98	18.72
0.25	1.5	0.5	100	98.0	4.78	31.12	32.40

Table 1: Experimental results for `NoisySGD` and their privacy analyses on MNIST. The accuracy is averaged over 10 independent runs. The hyperparameters in the first three rows are the same as in [2]. The μ in the 6th row is calculated using (5), which carries over to the 7th row via (6) with $\delta = 10^{-5}$. The number of epochs is equal to $T \times$ mini-batch size$/n = pT$.

the model from the third row preserves a decent amount of privacy since it is *not* always easy to tell apart $\mathcal{N}(0,1)$ and $\mathcal{N}(1.13,1)$. In stark contrast, the (ε,δ)-DP viewpoint is too conservative, suggesting that for the *same* model not much privacy is left, due to a very large "likelihood ratio" e^{ε} in Definition 2.1: it equals $e^{7.10} = 1212.0$ or $e^{5.07} = 159.1$ depending on which approach is chosen. This shortcoming of (ε,δ)-DP cannot be overcome by taking a larger δ, which, although gives rise to a smaller ε, would undermine the privacy guarantee from a different perspective.

For all experiments described in Table 1, Figure 3 illustrates the privacy bounds given by the `CLT` approach and the moments accountant both in terms of trade-off functions. The six plots in the first and third rows are with respect to $\delta = 10^{-5}$, from which the f-DP framework is seen to provide an analyst with substantial improvements in the privacy bounds. For the model corresponding to 96.6% test accuracy, concretely, the minimum sum of type I and type II errors in the sense of hypothesis testing is (at least) 77.6% by the `CLT` approach, whereas it is merely (at least) 9.4% by the moments accountant. For completeness, we show the optimal trade-off functions over all pairs of ε, δ given by the moments accountant in the middle row. The gaps between the two approaches exist, as predicted by Theorem 1, and remain significant.

Next, we extend our experiments to other datasets to further test f-DP for training private neural networks. The experiments compare private models under the privacy budget $\mu \leq 2$ to their non-private counterparts and some popular baseline methods. For simplicity, we focus on shallow neural networks and leave the investigation of complex architectures for future research.

Adult income. Originally from the UCI repository [18], the Adult income dataset has been preprocessed into the LIBSVM format [12]. This dataset contains 32,561 examples, each of which has 123 features and a label indicating whether the individual's annual income is more than \$50,000 or not. We randomly choose 10% of the examples as the test set (3,256 examples) and use the remaining 29,305 examples as the training set.

Our model is a single-layer multi-perceptron with 16 neurons and the ReLU activation. We set $\sigma = 0.55$, $p = 256/29305$, $\eta = 0.15, R = 1$, and use `NoisySGD` as our optimizer. The results displayed in Table 2 show that our private model achieves comparable performance to the baselines in the MLC++ library [33] in terms of test accuracy.

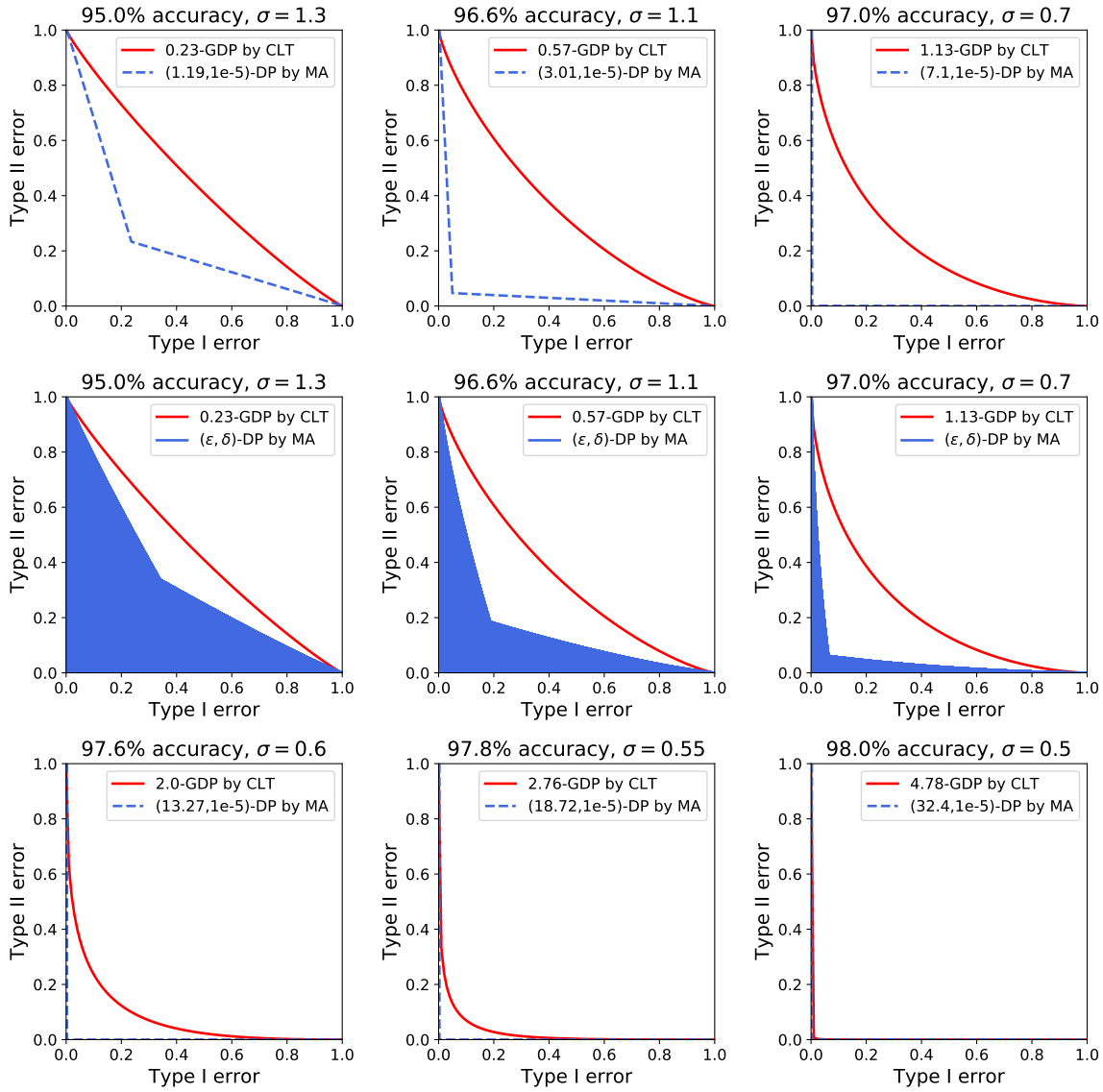

Figure 3: Comparisons between the two ways of privacy analysis on MNIST in terms of the trade-off between type I and type II errors, in the same setting as Table 1. The plots are different from Figure 7 in [17]. The (ε, δ)-DP guarantees are plotted according to (1). The blue regions in the plots from the second row correspond to all pairs of (ε, δ) computed by `MA`. The blue regions are not noticeable in the third row.

IMDb. We use the IMDb movie review dataset [38] for binary sentiment classification (positive or negative reviews). The dataset contains 25,000 training and 25,000 test examples. In our experiments, we prepocess the dataset by only including the top 10,000 frequently used words and discard the rest. Next, we set every example to have 256 words by truncating the length or filling with zeros if necessary.

In our neural networks, the input is first embedded into 16 units and then is passed through a

Models	Epochs	Test accuracy (%)	CLT μ	CLT ε	MA ε
private networks	18	84.0	2.03	10.20	14.70
non-private networks	20	84.5	—	—	—
kNN [14]	—	79.7	—	—	—
naive Bayes	—	83.9	—	—	—
voted ID3 [47]	—	84.4	—	—	—
C4.5 [48]	—	84.5	—	—	—

Table 2: Results for `NoisySGD` on the Adult income dataset. The ε parameters are with respect to $\delta = 10^{-5}$.

global average pooling. The intermediate output is fed to a fully-connected layer with 16 neurons, followed by a ReLU layer. We set $\sigma = 0.56$, $p = 512/25000$, $\eta = 0.02, R = 1$, and use `NoisyAdam` as our optimizer, which is observed to converge much faster than `NoisySGD` in this training task. We use the (non-private) two-layer LSTM RNN model in the Tensorflow tutorials [1] as a baseline model. Table 3 reports the experimental results. Notably, the private neural networks perform comparably to the baseline model, at the cost of only one percent drop in test accuracy compared to the non-private counterpart.

Models	Epochs	Test accuracy (%)	CLT μ	CLT ε	MA ε
private networks	9	83.8	2.07	10.43	15.24
non-private networks	20	84.7	—	—	—
LSTM-RNN [29]	10	85.4	—	—	—

Table 3: Results for `NoisyAdam` on the IMDB dataset, with $\delta = 10^{-5}$ used in the privacy analyses.

MovieLens. The MovieLens movie rating dataset [27] is a benchmark dataset for recommendation tasks. Our experiments consider the MovieLens 1M dataset, which contains 1,000,209 movie ratings from 1 star to 5 stars. In total, there are 6,040 users who rated 3,706 different movies. For this multi-class classification problem, the root mean squared error (RMSE) is chosen as the performance measure. It is worthwhile to mention that, as each user only watched a small fraction of all the movies, most (user, movie) pairs correspond to missing ratings. We randomly sample 20% of the examples as the test set and take the remainder as the training set.

Our model is a simplified version of the neural collaborative filtering in [28]. The network architecture consists of two branches. The left branch applies generalized matrix factorization to embed the users and movies using five latent factors. The output of the user embedding is multiplied by the item embedding. In the right branch, we use 10 latent factors for embedding. The embedding from both branches are then concatenated, which is fed to a fully-connected output layer. We set $\sigma = 0.6, p = 1/80, \eta = 0.01$, and $R = 5$ in `NoisyAdam`.

Table 4 presents the numerical results of our neural networks as well as baseline models in the Suprise library [30] in their default settings. The difference in RMSE between the non-private networks and the private one is relatively large for the MovieLens 1M dataset. Nevertheless,

Models	Epochs	RMSE	CLT μ	CLT ε	MA ε
private networks	20	0.915	1.94	10.61	15.39
non-private networks	20	0.893	—	—	—
SVD	—	0.873	—	—	—
NMF	—	0.916	—	—	—
user-based CF [50]	—	0.923	—	—	—
global average	—	1.117	—	—	—

Table 4: Results for `NoisyAdam` on the MovieLens 1M dataset, with $\delta = 10^{-6}$ used in the privacy analyses. CF stands for collaborative filtering.

the private model still outperforms many popular non-private models, including the user-based collaborative filtering and nonnegative matrix factorization.

4.2 The (ε, δ)-DP Perspective

While we hope that the f-DP perspective has been conclusively demonstrated to be advantageous, this section shows that the `CLT` approach continues to bring considerable benefits even in terms of (ε, δ)-DP. Specifically, by making use of the comparisons between the `CLT` approach and the moments accountant in Section 3.2, we can add less noise to the gradients in `NoisySGD` and `NoisyAdam` while achieving the same (ε, δ)-DP guarantees provided by the moments accountant. With less added noise, conceivably, an optimizer would have a higher prediction accuracy.

Figure 4 illustrates the experimental results on MNIST. In the top two plots, we set the noise scales to $\sigma = 1.3, \tilde{\sigma} = 1.06$, which are both shown to give $(1.34, 10^{-5})$-DP at epoch 20 using the moments accountant and the `CLT` approach, respectively. The test accuracy associated with the `CLT` approach is almost always higher than that associated with the moments accountant. In addition, another benefit of taking the `CLT` approach is that it gives rise to stronger privacy protection before reaching epoch 20, as shown by the right plot. For the bottom plots, although the improvement in test accuracy at the end of training is less significant, the `CLT` approach leads to much faster convergence at early epochs. To be concrete, the numbers of epochs needed to achieve 95%, 96%, and 97% test accuracy are 18, 26, and 45, respectively, for the neural networks with less noise, whereas the numbers of epochs are 23, 33, and 64, respectively, using noise level that is computed by the moments accountant. In a similar vein, the moments accountant gives a test accuracy of 92% for the first time when $\varepsilon = 4$ and the `CLT` approach achieves 96% under the same privacy budget.

5 Discussion

In this paper, we have showcased the use of f-DP, a very recently proposed privacy definition, for training private deep learning models using SGD or Adam. Owing to its strength in handling composition and subsampling and the powerful privacy central limit theorem, the f-DP framework allows for a closed-form privacy bound that is sharper than the one given by the moments accountant in the (ε, δ)-DP framework. By numerical experiments, we show that the trained neural networks

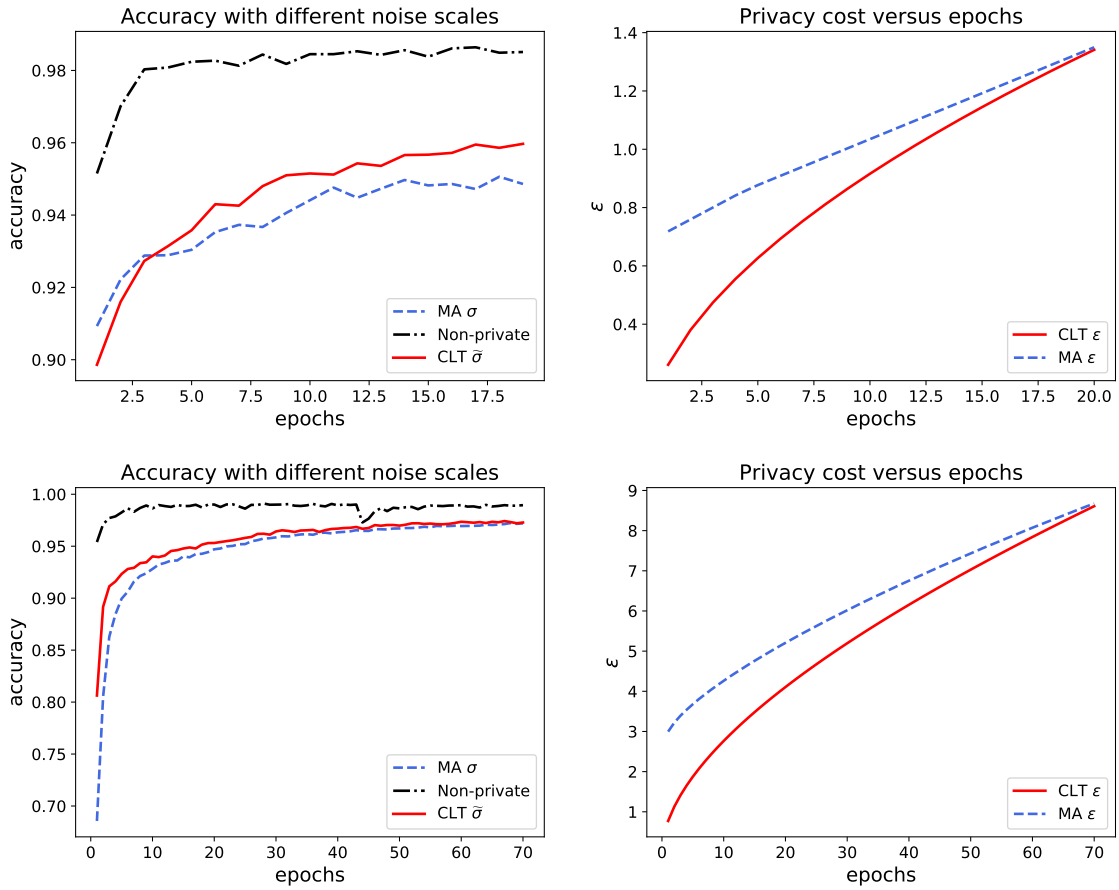

Figure 4: Experimental results from one run of `NoisySGD` on MNIST with different noise scales but the same (ε, δ)-DP guarantees. The top plots use $p = 256/60000, \eta = 0.15, R = 1.5$, and $\sigma = 1.3, \tilde{\sigma} = 1.06$. The `CLT` approach with $\tilde{\sigma} = 1.06$ and the moments accountant with $\sigma = 1.3$ give $(1.34, 10^{-5})$-DP at the 20th epoch ($\mu_{\texttt{CLT}} = 0.35$). The bottom plots use the same parameters except for $\sigma = 0.7, \tilde{\sigma} = 0.638$, and $\eta = 0.15$. Both approaches give $(8.68, 10^{-5})$-DP at epoch 70 ($\mu_{\texttt{CLT}} = 1.78$). The right plots show the privacy loss during the training process in terms of the ε spending with respect to $\delta = 10^{-5}$.

can be quite private from the f-DP viewpoint (for instance, 1.13-GDP[11]) but are *not* in the (ε, δ)-DP sense due to over conservative privacy bounds (for instance, $(7.10, 10^{-5})$-DP) computed in the (ε, δ)-DP framework. This in turn suggests that one can add less noise during the training process while having the same privacy guarantees as using the moments accountant, thereby improving model utility.

We conclude this paper by offering several directions for future research. As the first direction, we may consider using time-dependent noise scales and learning rates in `NoisySGD` and `NoisyAdam` for a better tradeoff between privacy loss and utility in the f-DP framework. Note that [35] has made considerable progress using concentrated differential privacy along this line. More generally,

[11]This means that undermining the privacy guarantee is harder than or of the same hardness as testing $H_0 : \mu = 0$ against $H_1 : \mu = 1.13$ based on the observation $\mu + \mathcal{N}(0, 1)$.

a straightforward but interesting problem is to extend this work to complex neural network architectures with a variety of optimization strategies. For example, can we develop some guidelines for choosing an optimizer among `NoisySGD`, `NoisyAdam`, and others for a given classification problem under some privacy constraint? Empirically, deep learning models are very sensitive to hyperparameters such as mini-batch size in terms of test accuracy. Therefore, from a practical standpoint, it would be of great importance to incorporate hyperparameter tuning into the f-DP framework [26]. Given f-DP's good interpretability and powerful toolbox, it is worthwhile investigating whether, from a broad perspective, its superiority over earlier differential privacy relaxations would hold in general private statistical and machine learning tasks. We look forward to more research efforts to further the theory and extend the use of f-DP.

Acknowledgments

We are grateful to David Durfee, Ryan Rogers, Aaron Roth, and Qinqing Zheng for stimulating discussions in the early stages of this work. This work was supported in part by NSF through CAREER DMS-1847415, CCF-1763314, and CCF-1934876, the Wharton Dean's Research Fund, and NIH through R01GM124111.

References

[1] IMDB Tensorflow RNN tutorial.
`www.tensorflow.org/tutorials/text/text_classification_rnn`.

[2] Tensorflow Privacy library.
`github.com/tensorflow/privacy`.

[3] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[4] J. M. Abowd. The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867. ACM, 2018.

[5] D. P. T. Apple. Learning with privacy at scale. Technical report, Apple, 2017.

[6] B. Balle and Y.-X. Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 403–412, 2018.

[7] B. Balle, G. Barthe, and M. Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, pages 6280–6290, 2018.

[8] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato. Hypothesis testing interpretations and renyi differential privacy. *arXiv preprint arXiv:1905.09982*, 2019.

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

[10] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[11] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86. ACM, 2018.

[12] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[13] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[14] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[15] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[16] J. Dong, D. Durfee, and R. Rogers. Optimal differential privacy composition for exponential mechanisms and the cost of adaptivity. *arXiv preprint arXiv:1909.13830*, 2019.

[17] J. Dong, A. Roth, and W. J. Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[18] D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[19] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[20] C. Dwork and G. N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[21] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[23] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 51–60. IEEE, 2010.

[24] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM Symposium on Theory of Computing*, pages 11–20. ACM, 2014.

[25] Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067. ACM, 2014.

[26] A. Gupta, K. Ligett, F. McSherry, A. Roth, and K. Talwar. Differentially private combinatorial optimization. In *Proceedings of the twenty-first annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1106–1125. Society for Industrial and Applied Mathematics, 2010.

[27] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Acm Transactions on Interactive Intelligent Systems (TIIS)*, 5(4):19, 2016.

[28] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.

[29] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[30] N. Hug. Surprise, a Python library for recommender systems. `http://surpriselib.com`, 2017.

[31] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[32] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] R. Kohavi, D. Sommerfield, and J. Dougherty. Data mining using/spl mscr//spl lscr//spl cscr/++ a machine learning library in c++. In *Proceedings Eighth IEEE International Conference on Tools with Artificial Intelligence*, pages 234–245. IEEE, 1996.

[34] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

[35] J. Lee and D. Kifer. Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1656–1665. ACM, 2018.

[36] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33. ACM, 2012.

[37] C. Liu, X. He, T. Chanyaswad, S. Wang, and P. Mittal. Investigating statistical privacy frameworks from the perspective of hypothesis testing. *Proceedings on Privacy Enhancing Technologies*, 2019(3): 233–254, 2019.

[38] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics, 2011.

[39] B. McMahan, G. Andrew, I. Mironov, N. Papernot, P. Kairouz, S. Chien, and Ú. Erlingsson. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210*, 2018.

[40] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[41] I. Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[42] J. Murtagh and S. Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2016.

[43] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pages 111–125. IEEE, 2008.

[44] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[45] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[46] V. Pichapati, A. T. Suresh, F. X. Yu, S. J. Reddi, and S. Kumar. AdaCliP: Adaptive clipping for private SGD. *arXiv preprint arXiv:1908.07643*, 2019.

[47] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[48] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[49] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton University Press, 1970.

[50] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl, et al. Item-based collaborative filtering recommendation algorithms. *Www*, 1:285–295, 2001.

[51] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321. ACM, 2015.

[52] D. Sommer, S. Meiser, and E. Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. 2018.

[53] S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

[54] L. Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

[55] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan. Subsampled Renyi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235, 2019.

[56] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[57] L. Xiang, J. Yang, and B. Li. Differentially-private deep learning from an optimization perspective. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 559–567. IEEE, 2019.

[58] J. Zhang, K. Zheng, W. Mou, and L. Wang. Efficient private ERM for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3922–3928. AAAI Press, 2017.

A Omitted Details in Section 2

We present Equation (3) as the following proposition, which is given in Section 2 but not in the foundational work [17].

Proposition A.1. If M is f-DP, and $S' = S \cup \{x_0\}$, then

$$T\big(M \circ \texttt{Sample}_p(S), M \circ \texttt{Sample}_p(S')\big) \geqslant pf + (1-p)\mathrm{Id}.$$

Proof. We first write the two distributions $M \circ \texttt{Sample}_p(S)$ and $M \circ \texttt{Sample}_p(S')$ as mixtures.

Without loss of generality, we can assume $S = \{x_1, \dots, x_n\}$ and $S' = \{x_0, x_1, \dots, x_n\}$. An outcome of the process $\texttt{Sample}_p$ when applied to S is a bit string $\vec{b} = (b_1, \dots, b_n) \in \{0,1\}^n$. Bit b_i dependes on whether x_i is selected into the subsample. We use $S_{\vec{b}} \subseteq S$ to denote the subsample determined by $\vec{b}$. When each b_i is sampled from a Bernoulli(p) distribution independently, $S_{\vec{b}}$ can be identified with $\texttt{Sample}_p(S)$. Let $\theta_{\vec{b}}$ be the probability that $\vec{b}$ appears. More specifically, if k out of n entries of $\vec{b}$ is one, then $\theta_{\vec{b}} = p^k(1-p)^{n-k}$. With this notation, $M \circ \texttt{Sample}_p(S)$ can be written as the following mixture:

$$M \circ \texttt{Sample}_p(S) = \sum_{\vec{b} \in \{0,1\}^n} \theta_{\vec{b}} \cdot M(S_{\vec{b}}).$$

Similarly, $M \circ \texttt{Sample}_p(S)$ can also be written as a mixture, with an additional bit indicating the presence of x_0. Alternatively, we can divide the components into two groups: one with x_0 present, and the other with x_0 absent. Namely,

$$M \circ \texttt{Sample}_p(S') = \sum_{\vec{b} \in \{0,1\}^n} p \cdot \theta_{\vec{b}} \cdot M(S_{\vec{b}} \cup \{x_0\}) + \sum_{\vec{b} \in \{0,1\}^n} (1-p) \cdot \theta_{\vec{b}} \cdot M(S_{\vec{b}}).$$

Note that $S_{\vec{b}} \cup \{x_0\}$ and $S_{\vec{b}}$ are neighbors, i.e. $M \circ \texttt{Sample}_p(S')$ is the mixture of neighboring distributions. The following lemma is the perfect tool to deal with it.

Lemma A.2. Let I be an index set. For all $i \in I$, P_i and Q_i are distributions that reside on a common sample space. $(\theta_i)_{i \in I}$ is a collection of non-negative numbers that sums to 1. If f is a trade-off function and $T(P_i, Q_i) \geqslant f$ for all i, then

$$T\Big(\sum \theta_i P_i, (1-p)\sum \theta_i P_i + p \sum \theta_i Q_i\Big) \geqslant pf + (1-p)\mathrm{Id}.$$

To apply the lemma, let the index be $\vec{b} \in \{0,1\}^n$, P_i be $M(S_{\vec{b}})$ and Q_i be $M(S_{\vec{b}} \cup \{x_0\})$. Condition $T(P_i, Q_i) \geqslant f$ is the consequence of M being f-DP. The conclusion simply translates to

$$T\big(M \circ \texttt{Sample}_p(S), M \circ \texttt{Sample}_p(S')\big) \geqslant pf + (1-p)\mathrm{Id},$$

which is what we want. The proof is complete. □

Proof of Lemma A.2. Let $P = \sum \theta_i P_i$ and $Q = (1-p)\sum \theta_i P_i + p \sum \theta_i Q_i$. Suppose ϕ satisfies $\mathbb{E}_P \phi = \alpha$. That is,

$$\sum \theta_i \mathbb{E}_{P_i} \phi = \alpha.$$

It is easy to see that

$$\mathbb{E}_Q \phi = (1-p)\alpha + p \sum \theta_i \mathbb{E}_{Q_i} \phi.$$

We know that $T(P_i, Q_i) \geqslant f$. Hence $\mathbb{E}_{Q_i}\phi \leqslant 1 - f(\mathbb{E}_{P_i}\phi)$. So

$$\sum \theta_i \mathbb{E}_{Q_i}\phi \leqslant 1 - \sum \theta_i f(\mathbb{E}_{P_i}\phi).$$

Since f is convex, Jensen's inequality implies

$$\sum \theta_i f(\mathbb{E}_{P_i}\phi) \geqslant f(\sum \theta_i \mathbb{E}_{P_i}\phi) = f(\alpha).$$

□

Next we use a figure to justify the claim we made in Section 2.2 that "CLT approximation works well for SGD". Recall that we argued in Section 3 that Algorithms 1 and 2 are $\min\{f, f^{-1}\}^{**}$-DP where

$$f = \big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T}.$$

This function converges to G_μ with $\mu = \nu\sqrt{\mathrm{e}^{1/\sigma^2} - 1}$ as $T \to \infty$ provided $p\sqrt{T} \to \nu$. In the following figure, we numerically compute f (blue dashed) and compare it with the predicted limit G_μ (red solid). More specifically, the configuration is designed to illustrate the fast convergence in the setting of the second line of Table 1, i.e. noise scale $\sigma = 1.1$, final GDP parameter $\mu = 0.57$ and test accuracy 96.6%. Originally the algorithm runs 60 epochs, i.e. $\approx$ 14k iterations. To best illustrate that convergence appears in early stage, the numerical evaluation uses a much smaller $T_{\text{numeric}} = 234$, i.e. only *one* epoch. In order to make the final limit consistent, we also enlarge the sample probability to p_{numeric} so that $p_{\text{numeric}} \cdot \sqrt{T_{\text{numeric}}}$ remain the same.

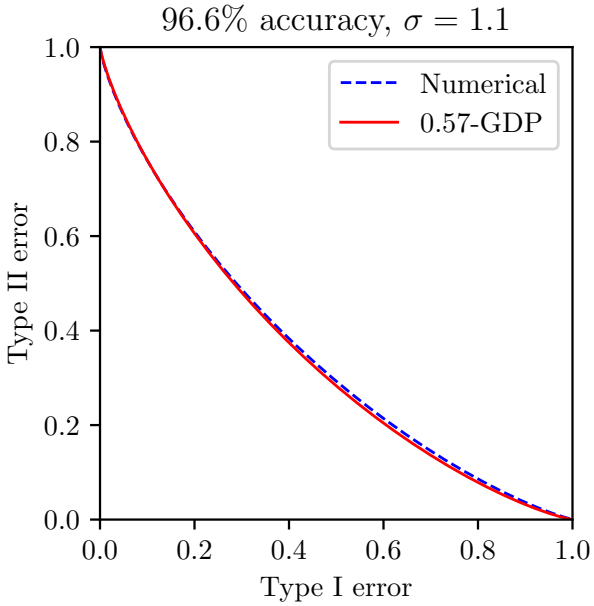

Figure 5: $\big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T}$ (blue dashed) is numerically computed and compared with the GDP limit (red solid) predicted by CLT. The two are almost identical at merely epoch one.

We have to remark that when σ is small, $\mu = \nu\sqrt{\mathrm{e}^{1/\sigma^2} - 1}$ gets large and yields challenges in the numerical computation of $\big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T}$. We leave rigorous and complete study to future work.

B Omitted Details in Section 3

B.1 Privacy Property of Algorithms 1 and 2

Theorem 3. Algorithms 1 and 2 are both $\min\{f, f^{-1}\}^{**}$-DP with $f = \big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T}$.

Proof. The proof is mostly done in the main text, except the composition step. Let V be the vector space that all θ_t live in and $\widetilde{M} = M \circ \texttt{Sample}_p : X^n \times V \to V$ be the gradient update. We have already proved (using Proposition A.1) that for both Algorithms 1 and 2, if $S' = S \cup \{x_0\}$, then $\widetilde{M}$ satisfies

$$T\big(M(S), M(S')\big) \geqslant f_p := pG_{1/\sigma} + (1-p)\mathrm{Id}.$$

Note that we cannot say M is f_p-DP because $T\big(M(S'), M(S)\big)$ is not necessarily lower bounded by f_p. So we need a more specific composition theorem than stated in [17].

Theorem 4 (Refined Composition). Suppose $M_1 : X \to Y, M_2 : X \times Y \to Z$ satisfy the following conditions for any S, S' such that $S' = S \cup \{x_0\}$:

1. $T\big(M_1(S), M(S')\big) \geqslant f$;
2. $T\big(M_2(S, y), M_2(S', y)\big) \geqslant g$ for any $y \in Y$.

Then the composition $M_2 \circ M_1 : X \to Y \times Z$ satisfies

$$T\big(M_2 \circ M_1(S), M_2 \circ M_1(S')\big) \geqslant f \otimes g$$

for any S, S' such that $S' = S \cup \{x_0\}$.

The theorem can be identically proved as Theorem 3.2 in [17].

Taking Algorithm 1 as an example, since

$$\begin{aligned} \texttt{NoisySGD} : X^n &\to V \times V \times \cdots \times V \\ S &\mapsto (\theta_1, \theta_2, \ldots, \theta_T) \end{aligned}$$

is simply the composition of T copies of $\widetilde{M}$, the above composition theorem implies that

$$T\big(\texttt{NoisySGD}(S), \texttt{NoisySGD}(S')\big) \geqslant \big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T} = f.$$

Moreover, $T\big(\texttt{NoisySGD}(S), \texttt{NoisySGD}(S')\big) \geqslant f^{-1}$. The two inequality let us conclude that any trade-off function of neighboring distributions must be lower bounded by at least one of f and f^{-1}, hence $\min\{f, f^{-1}\}$, hence $\min\{f, f^{-1}\}^{**}$. In other words, `NoisySGD` is $\min\{f, f^{-1}\}^{**}$-DP.

For `NoisyAdam`, we argued that its privacy property is the same as `NoisySGD` in each iteration, so the above argument also applies, and we have the same conclusion. □

B.2 Justifying CLT for Algorithms 1 and 2

The main purpose of this section is to show the following theorem

Theorem 5. Suppose p depends on T and $p\sqrt{T} \to \nu$. Then we have the following uniform convergence as $T \to \infty$

$$\big(pG_{1/\sigma} + (1-p)\mathrm{Id}\big)^{\otimes T} = G_\mu,$$

where $\mu = \nu \cdot \sqrt{T(\mathrm{e}^{1/\sigma^2} - 1)}$.

This theorem is the corollary of the following more general CLT on composition of subsample mechanisms and Lemma B.1 below.

Theorem 6. Suppose f is a trade-off function such that (1) $f(0) = 1$, (2) $f(x) > 0$, for all $x < 1$ and (3) $\int_0^1 (f'(x)+1)^4 \,\mathrm{d}x < +\infty$. Let $f_p = pf + (1-p)\mathrm{Id}$ as usual. Furthermore, assume $p\sqrt{T} \to \nu$ as $T \to \infty$ for some constant $\nu > 0$. Then we have the uniform convergence

$$f_p^{\otimes T} \to G_{\nu\sqrt{\chi^2(f)}}$$

as $T \to \infty$, where $\chi^2(f) = \int_0^1 \left(f'(x)\right)^2 \mathrm{d}x - 1$.

Lemma B.1. We have

$$\chi^2(G_{1/\sigma}) = \mathrm{e}^{1/\sigma^2} - 1.$$

In order to prove Theorem 6, we need an even more general CLT. The first privacy CLT was introduced in [17]. However, that version is valid only when each component trade-off function is symmetric, which is not true for $pG_{1/\sigma} + (1-p)\mathrm{Id}$. In order to state the general CLT that applies to asymmetric trade-off functions, we need to introduce the following functionals:

$$\begin{aligned}
\mathrm{kl}(f) &:= -\int_0^1 \log|f'(x)| \,\mathrm{d}x \\
\tilde{\mathrm{kl}}(f) &:= \int_0^1 |f'(x)| \log|f'(x)| \,\mathrm{d}x \\
\kappa_2(f) &:= \int_0^1 \log^2|f'(x)| \,\mathrm{d}x \\
\tilde{\kappa}_2(f) &:= \int_0^1 |f'(x)| \log^2|f'(x)| \,\mathrm{d}x \\
\kappa_3(f) &:= \int_0^1 \left|\log|f'(x)|\right|^3 \mathrm{d}x \\
\tilde{\kappa}_3(f) &:= \int_0^1 |f'(x)| \cdot \left|\log|f'(x)|\right|^3 \mathrm{d}x.
\end{aligned}$$

Theorem 7. Let $\{f_{ni} : 1 \leqslant i \leqslant n\}_{n=1}^{\infty}$ be a triangular array of (possibly asymmetric) trade-off functions and assume the following limits for some constants $K \geqslant 0$ and $s > 0$ as $n \to \infty$:

1. $\sum_{i=1}^n \mathrm{kl}(f_{ni}) + \tilde{\mathrm{kl}}(f_{ni}) \to K$;
2. $\max_{1\leqslant i\leqslant n} \mathrm{kl}(f_{ni}) \to 0, \quad \max_{1\leqslant i\leqslant n} \tilde{\mathrm{kl}}(f_{ni}) \to 0$;
3. $\sum_{i=1}^n \kappa_2(f_{ni}) \to s^2, \quad \sum_{i=1}^n \tilde{\kappa}_2(f_{ni}) \to s^2$;
4. $\sum_{i=1}^n \kappa_3(f_{ni}) \to 0, \quad \sum_{i=1}^n \tilde{\kappa}_3(f_{ni}) \to 0$.

Then, we have

$$\lim_{n\to\infty} f_{n1} \otimes f_{n2} \otimes \cdots \otimes f_{nn}(\alpha) = G_{K/s}(\alpha)$$

uniformly for all $\alpha \in [0, 1]$.

Proof of this theorem exactly mimics that of Theorem 3.5 in [17], which we omit here for its length and tediousness.

Next, we apply the asymmetric CLT to $\big(pf + (1-p)\mathrm{Id}\big)^{\otimes T}$ and prove Theorem 6. We start by collecting the necessary expressions into the following lemma. All of them are straightforward.

Lemma B.2. Let $g(x) = -f'(x) - 1 = |f'(x)| - 1$. Then

$$\begin{aligned}
\mathrm{kl}(f_p) &= -\int_0^1 \log(1 + pg(x))\,\mathrm{d}x \\
\tilde{\mathrm{kl}}(f_p) &= \int_0^1 (1 + pg(x))\log(1 + pg(x))\,\mathrm{d}x \\
\kappa_2(f_p) &= \int_0^1 \big[\log\big(1 + pg(x)\big)\big]^2\,\mathrm{d}x \\
\tilde{\kappa}_2(f_p) &= \int_0^1 \big(1 + pg(x)\big)\big[\log\big(1 + pg(x)\big)\big]^2\,\mathrm{d}x \\
\kappa_3(f_p) &= \int_0^1 \big[\log\big(1 + pg(x)\big)\big]^3\,\mathrm{d}x \\
\tilde{\kappa}_3(f_p) &= \int_0^1 \big(1 + pg(x)\big)\big[\log\big(1 + pg(x)\big)\big]^3\,\mathrm{d}x.
\end{aligned}$$

Proof of Theorem 6. It suffices to compute the limits in the asymmetric Central Limit Theorem 7, namely

$$T \cdot \big(\mathrm{kl}(f_p) + \tilde{\mathrm{kl}}(f_p)\big),\ T \cdot \kappa_2(f_p),\ T \cdot \tilde{\kappa}_2(f_p),\ T \cdot \kappa_3(f_p) \text{ and } T \cdot \tilde{\kappa}_3(f_p).$$

Since $T \sim p^{-2}$, we can consider $p^{-2}\big(\mathrm{kl}(f_p) + \tilde{\mathrm{kl}}(f_p)\big)$ and so on.

As in Lemma B.2, let $g(x) = -f'(x) - 1 = |f'(x)| - 1$. The assumption expressed in terms of g is simply

$$\int_0^1 g(x)^4\,\mathrm{d}x < +\infty.$$

In particular, it implies $|g(x)|^k$ is integrable in $[0,1]$ for $k = 2, 3, 4$. In addition, by Lemma B.2,

$$\chi^2(f) = \int_0^1 \big(f'(x)\big)^2\,\mathrm{d}x - 1 = \int_0^1 \big(f'(x) + 1\big)^2\,\mathrm{d}x = \int_0^1 g(x)^2\,\mathrm{d}x.$$

For the functional kl, by Lemma B.2,

$$\begin{aligned}
\lim_{p\to 0^+} \frac{1}{p^2}\big(\mathrm{kl}(f_p) + \tilde{\mathrm{kl}}(f_p)\big) &= \lim_{p\to 0^+} \int_0^1 g(x) \cdot \frac{1}{p}\log\big(1 + pg(x)\big)\,\mathrm{d}x \qquad (9) \\
&= \int_0^1 g(x) \cdot \lim_{p\to 0^+} \frac{1}{p}\log\big(1 + pg(x)\big)\,\mathrm{d}x \\
&= \int_0^1 g(x)^2\,\mathrm{d}x = \chi^2(f)
\end{aligned}$$

Changing the order of the limit and the integral in (9) is approved by the dominated convergence theorem. To see this, notice that $\log(1 + x) \leqslant x$. The integrand in (9) satisfies

$$0 \leqslant g(x) \cdot \frac{1}{p}\log\big(1 + pg(x)\big) \leqslant g(x)^2.$$

We already argued that $g(x)^2$ is integrable, so it works as a dominating function and the limit is justified. When $p\sqrt{T} \to \nu$, we have

$$T \cdot \mathrm{kl}(f_p) \to \nu^2 \cdot \chi^2(f).$$

So the constant K in Theorem 7 is $\nu^2 \cdot \chi^2(f)$.

For the functional κ_2 we have

$$\frac{1}{p^2}\kappa_2(f_p) = \int_0^1 \Big[\frac{1}{p}\log\big(1+pg(x)\big)\Big]^2 \,\mathrm{d}x.$$

By a similar dominating function argument,

$$\lim_{p\to 0^+} \frac{1}{p^2}\,\kappa_2(f_p) = \lim_{p\to 0^+} \frac{1}{p^2}\,\tilde{\kappa}_2(f_p) = \int_0^1 g(x)^2\,\mathrm{d}x = \chi^2(f).$$

Adding in the limit $p\sqrt{T} \to \nu$, we know s^2 in Theorem 7 is $\nu^2 \cdot \chi^2(f)$.

The same argument involving $\big|g(x)\big|^3$ and $g(x)^4$ applies to the functional κ_3 and $\tilde{\kappa}_3$ respectively and yields

$$\lim_{p\to 0^+} \frac{1}{p^3}\,\kappa_3(f_p) = \lim_{p\to 0^+} \frac{1}{p^3}\,\tilde{\kappa}_3(f_p) = \int_0^1 g(x)^3\,\mathrm{d}x.$$

Note the different power in p in the denominator. It means $\kappa_3(f_p) = o(p^2)$ and hence $T\cdot\kappa_3(f_p) \to 0$ when $p\sqrt{T} \to \nu$.

Hence all the limits in Theorem 7 check and we have a G_μ limit where

$$\mu = K/s = s = \sqrt{\nu^2 \cdot \chi^2(f)} = \nu \cdot \sqrt{\chi^2(f)}.$$

This completes the proof. □

We finish the section by proving the formula in Lemma B.1.

Proof of Lemma B.1. The best calculation is done via better understanding. We point out that the functional χ^2 is doing nothing more than computing the famous χ^2-divergence. Recall that Neyman χ^2-divergence (reverse Pearson) of P, Q is defined as

$$\chi^2(P\|Q) := \mathbb{E}_P\big[(\tfrac{\mathrm{d}Q}{\mathrm{d}P} - 1)^2\big]$$

Lemma B.3. If $f = T(P,Q)$ and $f(0) = 1$, $f(x) > 0$, for all $x < 1$, then $\chi^2(f) = \chi^2(P\|Q)$.

This lemma is a straightforward corollary of Proposition B.4 in [17], which gives expressions for all F-divergence[12]. In particular, if $f = T(P,Q)$ and $f(0) = 1$, $f(x) > 0, \forall x < 1$, then F-divergence of P, Q can be computed from their trade-off function as follows:

$$D_F(P\|Q) = \int_0^1 F\big(\big|f'(x)\big|^{-1}\big) \cdot \big|f'(x)\big|\,\mathrm{d}x.$$

[12] We use capital F to avoid confusion with the notation of trade-off function.

Neyman χ^2-divergence corresponds to $F(t) = \frac{1}{t} - 1$, so

$$\begin{aligned}\chi^2(P\|Q) &= \int_0^1 \left(\frac{1}{\left|f'(x)\right|^{-1}} - 1\right) \cdot \left|f'(x)\right| \mathrm{d}x \\ &= \int_0^1 \left(f'(x)\right)^2 \mathrm{d}x - \int_0^1 \left|f'(x)\right| \mathrm{d}x \\ &= \int_0^1 \left(f'(x)\right)^2 \mathrm{d}x - 1.\end{aligned}$$

With this formula, computing $\chi^2(G_{1/\sigma})$ is straightforward:

$$\chi^2(G_{1/\sigma}) = \chi^2\big(\mathcal{N}(\tfrac{1}{\sigma}, 1)\|\mathcal{N}(0,1)\big) = \mathrm{e}^{1/\sigma^2} - 1.$$

□

B.3 Proof of Theorems 1 and 2

Recall that Theorems 1 and 2 compare our `CLT` approach to moments accountant (`MA`) from two different perspectives: f-DP perspective in Theorem 1 and (ε, δ)-DP perspective in Theorem 2. We first show that Theorem 1 can be derived from Theorem 2. Then we prove a refined version of Theorem 2. To be more precise about the statement, let us first expand the notations used in the main text.

Let $\delta_{\texttt{MA}}(\varepsilon; \sigma, p, T)$ be the δ value computed by moment accountant method (described in detail below) for `NoisySGD` algorithm with subsampling probability p, iteration T and noise scale σ. Similarly, $\delta_{\texttt{CLT}}(\varepsilon; \sigma, \nu)$ denotes the δ value computed for the same algorithm using central limit theorem assuming $p\sqrt{T} \to \nu$.

Let $f_T(\alpha) = \sup_{\varepsilon \geqslant 0} f_{\varepsilon, \delta_{\texttt{MA}}(\varepsilon)}(\alpha)$. It is supported by $f_{\varepsilon_T, \delta_{\texttt{MA}}(\varepsilon_T)}$ at α. Theorem 2 says this supporting function is smaller than that of $G_{\mu_{\texttt{CLT}}}$ at α by a strict gap. Taking the limit, $\limsup_{T\to\infty} f_T(\alpha)$ has at least that much gap from $G_{\mu_{\texttt{CLT}}}(\alpha)$, which proves Theorem 1.

Theorem 2 is a straightforward corollary of the following proposition. Note that the inequality is reversed compared to the statement of Theorem 2 so that the gap is positive, which also turns lim sup into lim inf.

Proposition B.4.

$$\liminf_{T\to\infty} \ \delta_{\texttt{MA}}(\varepsilon; \sigma, \tfrac{\nu}{\sqrt{T}}, T) - \delta_{\texttt{CLT}}(\varepsilon; \sigma, \nu) \geqslant \mathrm{e}^{\varepsilon} \cdot \Phi(-\frac{\varepsilon}{\mu} - \frac{\mu}{2})$$

where $\mu = \nu \cdot \sqrt{\mathrm{e}^{1/\sigma^2} - 1}$.

Let us first describe how the two methods compute δ from ε.

$$\begin{aligned}\delta_{\texttt{nMA}}(\varepsilon; \sigma, p, T) &:= \inf_{\lambda \in \text{orders}} \exp\left(T \cdot \alpha_{\mathrm{GM}}(\lambda; \sigma, p) - \lambda\varepsilon\right) \\ \delta_{\texttt{MA}}(\varepsilon; \sigma, p, T) &:= \inf_{\lambda > 0} \exp\left(T \cdot \alpha_{\mathrm{GM}}(\lambda; \sigma, p) - \lambda\varepsilon\right)\end{aligned}$$

where $\alpha_{\mathrm{GM}}(\lambda; \sigma, p)$ is a scaled version of the Rényi divergence of Gaussian mixtures. More specifically, let $P = \mathcal{N}(0,1)$ and $Q = \mathcal{N}(\frac{1}{\sigma}, 1)$. We further denote the Gaussian mixture $pQ + (1-p)P$ by $\mathrm{GM}_{p,\sigma}$. The

$$\alpha_{\mathrm{GM}}(\lambda; \sigma, p) = \max\Big\{\lambda D_{\lambda+1}(\mathrm{GM}_{p,\sigma}\|P), \lambda D_{\lambda+1}(P\|\mathrm{GM}_{p,\sigma})\Big\}.$$

In [3], it has been shown that Algorithm 1 (hence also the Adam variant, Algorithm 2) with subsampling probability p, iteration T and noise scale σ is (ε, δ)-DP for each $\varepsilon \geqslant 0$ if $\delta = \delta_{\texttt{MA}}(\varepsilon; \sigma, p, T)$. To evaluate the infimum, the domain is discretized[13]. This results in the numerical moment accountant method that is actually implemented. Since $\delta_{\texttt{nMA}}(\varepsilon; \sigma, p, T) \geqslant \delta_{\texttt{MA}}(\varepsilon; \sigma, p, T)$, Algorithm 1 is also (ε, δ)-DP with $\delta = \delta_{\texttt{nMA}}(\varepsilon; \sigma, p, T)$.

On the other hand, $\delta_{\texttt{CLT}}(\varepsilon; \sigma, \nu)$ is obtained by first observing Algorithm 1 is asymptotically $\mu_{\texttt{CLT}}$-GDP with $\mu_{\texttt{CLT}} = \nu \cdot \sqrt{\mathrm{e}^{1/\sigma^2} - 1}$ and then convert GDP to (ε, δ)-DP via Equation (6), i.e. Algorithm 1 asymptotically satisfies (ε, δ)-DP where

$$\delta = \delta_{\texttt{CLT}}(\varepsilon; \sigma, \nu) = 1 + G^*_{\mu_{\texttt{CLT}}}(-\mathrm{e}^{\varepsilon}).$$

We have just explained how `MA` and `CLT` works. Next we prove Proposition B.4

Proof of Proposition B.4. Let $f_T = \big(pG_\mu + (1-p)\mathrm{Id}\big)^{\otimes T}$. We need a lemma (whose proof is provided later) that relates the Rényi divergence to the trade-off function f_T.

Lemma B.5.

$$T \cdot \alpha_{\mathrm{GM}}(\lambda; \sigma, p) \geqslant \log \int_0^1 |f_T'(x)|^{\lambda+1} \, \mathrm{d}x.$$

Let $x_T \in (0,1)$ be the point such that $f_T'(x_T) = -\mathrm{e}^{\varepsilon}$ (or $x_T \in \partial f_T^*(-\mathrm{e}^{\varepsilon})$ if readers worry about differentiability). We have

$$1 + f_T^*(-\mathrm{e}^{\varepsilon}) = \sup_{0 \leqslant x \leqslant 1} \{1 - f_T(x) - \mathrm{e}^{\varepsilon} x\} = 1 - f_T(x_T) - \mathrm{e}^{\varepsilon} x_T$$

It is clear that $|f_T'(x)| \geqslant \mathrm{e}^{\varepsilon}$ for $0 \leqslant x \leqslant x_T$.

On the other hand, using Lemma B.5, we get

$$\begin{aligned}
\delta_{\texttt{MA}}(\varepsilon; \sigma, p, T) &= \inf_{\lambda > 0} \exp\left(T \cdot \alpha_{\mathrm{GM}}(\lambda; \sigma, p) - \lambda\varepsilon\right) \\
&\geqslant \inf_{\lambda > 0} \mathrm{e}^{-\lambda\varepsilon} \int_0^1 |f_T'(x)|^{\lambda+1} \, \mathrm{d}x \\
&> \inf_{\lambda > 0} \int_0^{x_T} |f_T'(x)|^{\lambda+1} \mathrm{e}^{-\lambda\varepsilon} \, \mathrm{d}x \\
&= \inf_{\lambda > 0} \int_0^{x_T} |f_T'(x)| \cdot |f_T'(x)|^{\lambda} \mathrm{e}^{-\lambda\varepsilon} \, \mathrm{d}x \\
&\geqslant \inf_{\lambda > 0} \int_0^{x_T} |f_T'(x)| \cdot (\mathrm{e}^{\varepsilon})^{\lambda} \mathrm{e}^{-\lambda\varepsilon} \, \mathrm{d}x \\
&= f_T(0) - f_T(x_T) \\
&= 1 - f_T(x_T) \\
&= \big(1 - f_T(x_T) - \mathrm{e}^{\varepsilon} x_T\big) + \mathrm{e}^{\varepsilon} x_T \\
&= 1 + f_T^*(-\mathrm{e}^{\varepsilon}) + \mathrm{e}^{\varepsilon} x_T.
\end{aligned}$$

In summary, we have

$$\delta_{\texttt{MA}}(\varepsilon; \sigma, p, T) > 1 + f_T^*(-\mathrm{e}^{\varepsilon}) + \mathrm{e}^{\varepsilon} x_T. \tag{10}$$

[13] Code in tensorflow/privacy discretizes at $[1.25, 1.5, 1.75, 2., 2.25, 2.5, 3., 3.5, 4., 4.5, 5, 6, 7, \ldots, 63, 64, 128, 256, 512]$.

Setting $p = \frac{\nu}{\sqrt{T}}$, we would like to take limit on both sides of (10). First notice that f_T converge pointwise to $G_{\mu_{\mathtt{CLT}}}$, which we have already proven in Appendix B.2. The limit of x_T is taken care of in the following lemma:

Lemma B.6.

$$\lim_{T\to\infty} x_T = x^* := \Phi(-\frac{\varepsilon}{\mu_{\mathtt{CLT}}} - \frac{\mu_{\mathtt{CLT}}}{2}).$$

Combining these results, we can take limits on both sides of (10):

$$\begin{aligned}\liminf_{T\to\infty} \delta_{\mathtt{MA}}(\varepsilon;\sigma,p,T) &\geqslant \lim_{T\to\infty} 1 + f_T^*(-\mathrm{e}^{\varepsilon}) + \mathrm{e}^{\varepsilon} x_T \\ &= 1 + G_{\mu_{\mathtt{CLT}}}^*(-\mathrm{e}^{\varepsilon}) + \mathrm{e}^{\varepsilon} x^* \\ &= \delta_{\mathtt{CLT}}(\varepsilon;\sigma,\nu) + \mathrm{e}^{\varepsilon}\cdot\Phi(-\frac{\varepsilon}{\mu_{\mathtt{CLT}}} - \frac{\mu_{\mathtt{CLT}}}{2}).\end{aligned}$$

This finishes the proof. □

Proof of Lemma B.5. The Rényi divergence can also be computed from the trade-off function, just like the χ^2-divergence. In fact, under the same assumptions as in Lemma B.1, we have

$$D_\alpha(Q\|P) = \frac{1}{\alpha-1}\log\int_0^1 |f'(x)|^{\alpha-1}\,\mathrm{d}x.$$

Alternatively,

$$\lambda D_{\lambda+1}(Q\|P) = \log\int_0^1 |f'(x)|^{\lambda}\,\mathrm{d}x. \tag{11}$$

This identity will be the bridge between α_{GM} and f_T.

On one hand, $\alpha_{\mathrm{GM}}(\lambda;\sigma,p)$ is the maximum of two Rényi divergences, so

$$\alpha_{\mathrm{GM}}(\lambda;\sigma,p) \geqslant \lambda D_{\lambda+1}(pQ + (1-p)P\|P)$$

Consequently,

$$\begin{aligned}T\cdot\alpha_{\mathrm{GM}}(\lambda;\sigma,p) &\geqslant T\lambda D_{\lambda+1}(pQ+(1-p)P\|P) \\ &= \lambda D_{\lambda+1}\big((pQ+(1-p)P)^T\|P^T\big).\end{aligned}$$

The last step is the tensorization identity of Rényi divergence.

On the other hand, notice that $pG_\mu + (1-p)\mathrm{Id} = T\big(P, \mathrm{GM}_{p,\sigma}\big)$ where we continue the use of notations $P = \mathcal{N}(0,1), Q = \mathcal{N}(\frac{1}{\sigma},1)$ and $\mathrm{GM}_{p,\sigma} = pQ + (1-p)P$. We have

$$f_T = \big(pG_\mu + (1-p)\mathrm{Id}\big)^{\otimes T} = T\big(P, (pQ+(1-p)P)\big)^{\otimes T} = T\big(P^T, (pQ+(1-p)P)^T\big)$$

Using (11), we have

$$\begin{aligned}T\cdot\alpha_{\mathrm{GM}}(\lambda;\sigma,p) &\geqslant \lambda D_{\lambda+1}\big((pQ+(1-p)P)^T\|P^T\big) \\ &= \log\int_0^1 |f_T'(x)|^{\lambda}\,\mathrm{d}x.\end{aligned}$$

□

Proof of Lemma B.6. By definition, $f_T'(x_T) = -\mathrm{e}^{\varepsilon}$. The convexity of f_T implies $\nabla f_T^*(-\mathrm{e}^{\varepsilon}) = x_T$.

Since f_T converges uniformly to $G_{\mu_{\mathtt{CLT}}}$ in $[0,1]$, we have uniform convergence $f_T^* \to G_{\mu_{\mathtt{CLT}}}^*$. By convexity of these functions, the convergence also implies the convergence of derivatives (See Theorem 25.7 of [49]), namely,

$$\nabla f_T^* \to \nabla G_{\mu_{\mathtt{CLT}}}^*.$$

Therefore,

$$x_T = \nabla f_T^*(-\mathrm{e}^{\varepsilon}) \to \nabla G_{\mu_{\mathtt{CLT}}}^*(-\mathrm{e}^{\varepsilon}).$$

Let $x^* = \nabla G_{\mu_{\mathtt{CLT}}}^*(-\mathrm{e}^{\varepsilon})$ be the limit. Using the convexity again, we have

$$-\mathrm{e}^{\varepsilon} = G_{\mu_{\mathtt{CLT}}}'(x^*).$$

We can solve for x^* using the expression of G_μ (6). After some algebra, we have

$$x^* = \Phi\left(-\frac{\varepsilon}{\mu_{\mathtt{CLT}}} - \frac{\mu_{\mathtt{CLT}}}{2}\right).$$

The proof is complete. □