DEEP LEARNING WITH GAUSSIAN DIFFERENTIAL PRIVACY

Zhiqi Bu[†], Jinshuo Dong[‡], Qi Long[‡], Weijie J. Su^{◇,*}[†] Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania[‡] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania[◇] Department of Statistics, University of Pennsylvania

ABSTRACT. Deep learning models are often trained on data sets 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 article, we consider a recently proposed privacy definition termed *f-differential privacy* (Dong, Roth, et al., 2019) for a refined privacy analysis of training neural networks. Leveraging the appealing properties of *f*-differential privacy in handling composition and subsampling, this article 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 Abadi et al. (2016) did. Our results demonstrate that the *f*-differential privacy framework allows for a new privacy analysis that improves on the prior analysis (Abadi et al., 2016), 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. Python code to calculate the privacy cost for these experiments is publicly available in the `TensorFlow Privacy` library.

Keywords: Gaussian differential privacy, deep learning, noisy gradient descent, central limit theorem, privacy accounting

MEDIA SUMMARY

In the era of big data, it is crucial and even urgent to develop algorithms that preserve the privacy of sensitive individual data while maintaining high utility. However, these two goals are generally competing with each other in the sense that privacy guarantees inevitably lead to utility loss. What is the optimal tradeoff between privacy and utility for a given machine learning task? This is a question that any data scientist caring about data privacy would ask.

*Corresponding author: suw@wharton.upenn.edu

In this article, the authors offer deep learning models with privacy guarantees in the framework of f -differential privacy, which was recently proposed as a unifying generalization of several earlier privacy definitions. The authors train their private neural networks on multiple tasks that involve different types of data, including images, texts, and numeric data. Owing to the versatility of the f -differential privacy framework, the experimental results demonstrate that this approach to private deep learning outperforms existing approaches in terms of the tradeoff between privacy and utility.

1. INTRODUCTION

In many applications of machine learning, the data sets 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 data set, 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, 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 data set (Narayanan & Shmatikov, 2008) and the identification of the health records of the then-Massachusetts governor in public anonymized medical data sets (Sweeney, 1997).

In this context, (ϵ, δ) -differential privacy (DP) arose as a mathematically rigorous definition of privacy by Dwork, McSherry, et al. (2006). Today, this definition has developed into a firm foundation of private data analysis, with its applications deployed by Google (Erlingsson et al., 2014), Apple (Differential-Privacy-Team, 2017), Microsoft (Ding et al., 2017), and the U.S. Census Bureau (Abowd, 2018). Despite its impressive popularity both in the scholarly literature and in 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 by Abadi et al. (2016).

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, et al. (2019). 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 (CLT) that states that the privacy guarantees of the composition of private algorithms are approximately equivalent to distinguish two shifted normal distributions.

The main results of this article 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 Shokri and Shmatikov (2015) and Abadi et al. (2016). In short, this article 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 (see Abadi et al. (2016)), 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 by Abadi et al. (2016), 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.

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 article 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 article 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 by Dwork, Kenthapadi, et al. (2006) and Dwork et al. (2010). In a pioneering work, Kairouz et al. (2017) obtained an optimal composition theorem for (ϵ, δ) -DP, which in fact served as one of the motivations for the f -DP work by Dong, Roth, et al. (2019). However, it is $\#P$ hard to compute the privacy bounds from their composition theorem (see Murtagh and Vadhan

¹It 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.

(2016)). More recently, Dong, Durfee, et al. (2019) 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 (see Bun et al. (2018), Bun and Steinke (2016), Dwork and Rothblum (2016), Mironov (2017)). Unfortunately, these relaxations either lack a privacy amplification by subsampling argument or present a quite complex argument that is difficult to use as in Balle et al. (2018) and Wang et al. (2019). 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. (2016) 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 (e.g., by Bassily et al. (2014), Chaudhuri et al. (2011), Shokri and Shmatikov (2015), Song et al. (2013), Zhang et al. (2017)) and allows for meaningful privacy guarantees for deep learning trained on realistically sized data sets. This technique has been extended to a variety of situations by follow-up work from McMahan et al. (2018), Pichapati et al. (2019). 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 by Papernot et al. (2018) and Papernot et al. (2016), the development of noised federated averaging algorithms by McMahan et al. (2017), and analyzing privacy costs through the lens of the optimization landscape of neural networks by Xiang et al. (2019).

2. PRELIMINARIES

2.1. f -Differential Privacy. In the differential privacy framework, we envision an adversary that is well-informed about the data set except for a single individual, and the adversary seeks to determine whether this individual is in the data set 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 data set can be thought of as a matrix, whose rows each contain one individual's data. Two data sets are said to be *neighbors* if one can be derived by discarding an individual from the other. As such, the sizes of neighboring data sets differ by one.² Let S and S' be neighboring data sets, and $\epsilon \geq 0, 0 \leq \delta \leq 1$ be two numbers, and denote by M a (randomized) algorithm that takes as input a data set.

Definition 2.1 (Dwork, Kenthapadi, et al. (2006), Dwork, McSherry, et al. (2006)). A (randomized) algorithm M gives (ϵ, δ) -differential privacy if for any pair of neighboring data sets S, S' and any

²Alternatively, the neighboring relationship can be defined for data sets of the same size and differing by one individual.

event E ,

$$\mathbb{P}(M(S) \in E) \leq e^\epsilon \mathbb{P}(M(S') \in E) + \delta.$$

To achieve privacy, the algorithm M is necessarily randomized, whereas the two data sets 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 into or opts out of the data set.

In essence, the adversary seeks to distinguish 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:

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

The connection between differential privacy and hypothesis testing was, to our knowledge, first noted in Wasserman and Zhou (2010), and was later developed in Balle et al. (2019), Kairouz et al. (2017), Liu et al. (2019). 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 Dong, Roth, et al. (2019), the authors propose to use the *trade-off* between type I error (the probability of erroneously rejecting H_0 when H_0 is true) and type II error (the probability of erroneously accepting H_0 when H_1 is true) 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, Dong, Roth, et al. (2019) define the *trade-off function* of P and Q as

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

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 (Dong, Roth, et al. (2019)). A (randomized) algorithm M is f -differentially private if

$$T(M(S), M(S')) \geq f$$

for all neighboring data sets S and S' .

In this definition, both T and f are functions that take $\alpha \in [0, 1]$ as input and the inequality holds pointwise for all $0 \leq \alpha \leq 1$, 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

(2010), (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

$$(2.1) \quad f_{\varepsilon, \delta}(\alpha) = \max \{0, 1 - \delta - e^\varepsilon \alpha, e^{-\varepsilon}(1 - \delta - \alpha)\}.$$

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 $\delta(\varepsilon) = 1 + f^*(-e^\varepsilon)$ for every $\varepsilon \geq 0$.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 $G_\mu := T(\mathcal{N}(0, 1), \mathcal{N}(\mu, 1))$ for $\mu \geq 0$. 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 (Dong, Roth, et al. (2019)). A (randomized) algorithm M is μ -Gaussian differentially private (GDP) if

$$T(M(S), M(S')) \geq G_\mu$$

for all neighboring data sets S and S' .

In words, μ -GDP says that determining whether any individual is in the data set is at least as difficult as distinguishing 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 $\sigma = \text{sens}(\theta)/\mu$. Here the *sensitivity* of θ is defined as $\text{sens}(\theta) = \sup_{S, S'} |\theta(S) - \theta(S')|$, where the supremum is over all neighboring data sets.

2.2. Properties of f -Differential Privacy. 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* data set where each analysis is informed by the explorations of prior analyses. A central issue 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 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 data set S . In general, the composition of more than two algorithms follows recursively.

To introduce the composition theorem for f -DP, Dong, Roth, et al. (2019) define 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 $t = 1, \dots, T$. Then their T -fold composition algorithm is $f_1 \otimes \dots \otimes f_T$ -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 $f_1 \otimes \dots \otimes f_T$ in general cannot be improved. Put more precisely, one can find an f_t -DP mechanism M_t for $t = 1, \dots, T$ such that their composition is precisely $f_1 \otimes \dots \otimes f_T$ -DP (see the discussion following Theorem 4 in Dong, Roth, et al. (2019)).

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

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

accumulate to G_μ for some μ under composition. Informally, assuming each f_t is very close to $\text{Id}(\alpha) = 1 - \alpha$, which corresponds to perfect privacy, then we have

$$(2.2) \quad f_1 \otimes f_2 \otimes \cdots \otimes f_T \text{ is approximately } G_\mu$$

if the number of iterations T is sufficiently large.⁴ This central limit theorem approximation is especially suitable for the privacy analysis of deep learning, where the training process typically takes at least tens of thousands of iterations. 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 6 in Dong, Roth, et al. (2019). We remark that a conceptually related article by Sommer et al. (2018) 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 are 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.

Next, we turn to the subsampling property of f -DP. In training neural networks, the gradient at each iteration is computed from a mini-batch that is *subsampling* from the training examples. Intuitively, an algorithm applied to a subsample gives stronger privacy guarantees than when 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 data set S , include his or her datum in the subsample independently with probability p , which is sometimes referred to as the Poisson subsampling (see Wang et al. (2019)). The resulting subsample is denoted by $\text{Sample}_p(S)$. For the purpose of clearing up any confusion, we remark that the subsample $\text{Sample}_p(S)$ has a random size and as an intermediate step is not released. Given any algorithm M , denote by $M \circ \text{Sample}_p$ the subsampled algorithm.

The subsampling theorem for f -DP states as follows. Let M be f -DP, write f_p for $pf + (1-p)\text{Id}$, and denote by f_p^{-1} the inverse⁵ of f_p . It is proved in Appendix A that the subsampled algorithm $M \circ \text{Sample}_p$ satisfies

$$(2.3) \quad T(M \circ \text{Sample}_p(S), M \circ \text{Sample}_p(S')) \geq f_p$$

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

$$T(M \circ \text{Sample}_p(S'), M \circ \text{Sample}_p(S)) \geq f_p^{-1}.$$

As such, the two displays above say that the trade-off function of $M \circ \text{Sample}_p$ on any neighboring data sets 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 $\min\{f_p, f_p^{-1}\}$

⁴If $f_i = G_{\mu_i}$ for $i = 1, \dots, T$, then the T -fold composition is exactly G_μ with $\mu = \sqrt{\sum_{i=1}^T \mu_i^2}$.

⁵For any trade-off function $f = T(P, Q)$, its inverse $f^{-1} = T(Q, P)$.

and is indeed a trade-off function. Taken together, all the pieces show that the subsampled algorithm $M \circ \text{Sample}_p$ 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 by Li et al. (2012).

3. ALGORITHMS AND THEIR PRIVACY ANALYSES

3.1. NoisySGD and NoisyAdam. SGD and Adam (Kingma & Ba, 2014) 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 $S = \{x_1, \dots, x_n\}$ denote the data set, 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 $\{1, 2, \dots, n\}$ 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, Bassily et al. (2014), Chaudhuri et al. (2011), Song et al. (2013), and Abadi et al. (2016) 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, $\|\cdot\|_2$ denotes the ℓ_2 norm, and we abuse notation by using T to denote the number of iterations. Formally, we present this in Algorithm 1, which we henceforth refer to as **NoisySGD**. As a contribution of the present paper, **NoisySGD** uses the Poisson subsampling, as opposed to the uniform sampling used in Dong, Roth, et al. (2019). For completeness, we remark that there is another possible subsampling method: shuffling (randomly permuting and dividing data into folds at each epoch). We emphasize that different subsampling mechanisms produce different privacy guarantees.

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 $\text{Sample}_p(S)$ for the mini-batch I_t (we drop the subscript t for simplicity). This allows us to use $M \circ \text{Sample}_p(S)$ 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 $f = (pG_{1/\sigma} + (1-p)\text{Id})^{\otimes T}$. As a remark, it has been empirically observed in Abadi et al. (2016) and Bagdasaryan et al. (2019) that the performance of the private neural

Algorithm 1 NoisySGD

Input: data set $S = \{x_1, \dots, x_n\}$, 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**

$$v_t^{(i)} \leftarrow \nabla_{\theta} \ell(\theta_t, x_i)$$

$$\bar{v}_t^{(i)} \leftarrow v_t^{(i)} / \max\{1, \|v_t^{(i)}\|_2 / R\}$$

▷ Clip gradient

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \frac{1}{|I_t|} \left(\sum_{i \in I_t} \bar{v}_t^{(i)} + \sigma R \cdot \mathcal{N}(0, I) \right)$$

▷ Apply Gaussian mechanism

Output θ_T

networks is not sensitive to the choice of the clipping norm bound R (see more discussion in Pichapati et al. (2019), Xu et al. (2019)).

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 $p\sqrt{T} \rightarrow \nu$ for a constant $\nu > 0$ as $T \rightarrow \infty$,

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

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

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

As such, the central limit theorem demonstrates that NoisySGD is approximately $p\sqrt{T(e^{1/\sigma^2} - 1)}$ -GDP. Denoting by $B = pn$ the mini-batch size, the privacy parameter $p\sqrt{T(e^{1/\sigma^2} - 1)}$ equals $\frac{B}{n}\sqrt{T(e^{1/\sigma^2} - 1)}$. 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 Dong, Roth, et al. (2019). Therein, the authors consider Algorithm 1 with uniform subsampling and obtain a privacy bound that is different from the one in the present article.

Next, we present a private version of Adam in Algorithm 2, which we refer to as NoisyAdam and can be found in Tensorflow (2019). 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 (Duchi et al., 2011) and therefore it is also asymptotically GDP in the same asymptotic regime.

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 Abadi et al. (2016), 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 by Dwork et al. (2010). In slightly more detail, the moments accountant uses the moment-generating function of the privacy loss random variable

Algorithm 2 NoisyAdam

Input: data set $S = \{x_1, \dots, x_n\}$, 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 , momentum parameters (β_1, β_2) , initial momentum m_0 , initial past squared gradient u_0 , and a small constant $\xi > 0$.

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

Output θ_T

to track the privacy loss under composition. As abuse of notation, this article uses functions $\delta_{\text{MA}} = \delta_{\text{MA}}(\varepsilon)$ and $\varepsilon_{\text{MA}} = \varepsilon_{\text{MA}}(\delta)$ to denote the mapping induced by the moments accountant in both directions⁶. For self-containedness, the appendix includes a formal description of the two functions.

Although NoisySGD and NoisyAdam are our primary focus, the 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 $t = 1, \dots, T$ and write M for their composition. On the one hand, the moments accountant technique ensures that M is $(\varepsilon, \delta_{\text{MA}}(\varepsilon))$ -DP for any ε or, put equivalently,⁷ is $(\varepsilon_{\text{MA}}, \delta)$ -DP. On the other hand, the composition algorithm is $f_1 \otimes \dots \otimes f_T$ -DP from the f -DP viewpoint and, following from the central limit theorem (2.2), 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 μ_{CLT} -GDP with privacy parameter

$$(3.1) \quad \mu_{\text{CLT}} = p \sqrt{T(e^{1/\sigma^2} - 1)}.$$

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 and, more importantly, the privacy central limit theorem is asymptotically exact. To formally state the result, note that the moments accountant asserts that the private optimizer is $(\varepsilon, \delta_{\text{MA}}(\varepsilon))$ -DP for all $\varepsilon \geq 0$, which is equivalent to $\sup_{\varepsilon \geq 0} f_{\varepsilon, \delta_{\text{MA}}(\varepsilon)}$ -DP by recognizing (2.1) (see also Proposition 2.11 in Dong, Roth, et al. (2019)).

⁶We omit the dependence of the functions on the specification of the composition algorithm such as p, σ, T as in NoisySGD and NoisyAdam.

⁷The 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 Dong, Roth, et al. (2019).

Roughly speaking, the following theorem says that $\sup_{\varepsilon \geq 0} f_{\varepsilon, \delta_{\text{MA}}(\varepsilon)}$ -DP (asymptotically) promises no more privacy guarantees than the bound of μ_{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 $T \rightarrow \infty$. Then, both **NoisySGD** and **NoisyAdam** satisfy

$$\limsup_{T \rightarrow \infty} \left(\sup_{\varepsilon \geq 0} f_{\varepsilon, \delta_{\text{MA}}(\varepsilon)}(\alpha) - G_{\mu_{\text{CLT}}}(\alpha) \right) \leq 0$$

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

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 as to which approach gives a sharper privacy analysis 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 $\varepsilon \geq 0$, where⁸

$$(3.2) \quad \delta(\varepsilon; \mu) = 1 + G_{\mu}^*(-e^{\varepsilon}) = \Phi\left(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}\right) - e^{\varepsilon} \Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\right).$$

Answering the question is, therefore, reduced to the comparison between $\delta_{\text{MA}}(\varepsilon)$ and $\delta_{\text{CLT}}(\varepsilon) := \delta(\varepsilon; \mu_{\text{CLT}})$ or, equivalently⁹, between $\varepsilon_{\text{MA}}(\delta)$ and $\varepsilon_{\text{CLT}}(\delta) := \varepsilon(\delta; \mu_{\text{CLT}})$.

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,

$$\limsup_{T \rightarrow \infty} (\delta_{\text{CLT}}(\varepsilon) - \delta_{\text{MA}}(\varepsilon)) < 0$$

for all $\varepsilon \geq 0$.

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 $\limsup_{T \rightarrow \infty} (\varepsilon_{\text{CLT}}(\delta) - \varepsilon_{\text{MA}}(\delta)) < 0$ for¹⁰ any δ . 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 .

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 $\delta(\varepsilon; \mu_{\text{CLT}}) \equiv \delta_{\text{CLT}}(\varepsilon) < \delta_{\text{MA}}(\varepsilon)$ (for sufficiently large T) and that $\delta(\varepsilon; \mu)$ increases as μ increases, one can find $\tilde{\mu}_{\text{CLT}} > \mu_{\text{CLT}}$ such that

$$(3.3) \quad \delta(\varepsilon; \tilde{\mu}_{\text{CLT}}) = \delta_{\text{MA}}(\varepsilon).$$

⁸See Section 2.4 of Dong, Roth, et al. (2019) for this result. See also Balle and Wang (2018), Dwork et al. (2014).

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

¹⁰Write $\delta_{\text{CLT}}^* = \delta_{\text{CLT}}(0)$ and set $\varepsilon_{\text{CLT}}(\delta) = 0$ for $\delta \geq \delta_{\text{CLT}}^*$. Apply the same adjustment for ε_{MA} .

Put differently, we can carefully adjust some parameters in Algorithm 1 and Algorithm 2 in order to let the algorithms be $\tilde{\mu}_{\text{CLT}}$ -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 (3.3) and

$$(3.4) \quad \tilde{\mu}_{\text{CLT}} = p\sqrt{T(e^{1/\tilde{\sigma}^2} - 1)}.$$

Note that this is adapted from (3.1).

Figure 1. An illustration of the CLT approach in the f -DP framework and the moments accountant in the (ϵ, δ) -DP framework. $\text{NoisyOptimizer}(\sigma, \dots)$ using the moments accountant gives the same privacy guarantees in terms of (ϵ, δ) -DP as $\text{NoisyOptimizer}(\tilde{\sigma}, \dots)$ using the CLT approach (the ellipses denote omitted parameters). Note that the duality formula (3.2) is used in solving $\tilde{\mu}_{\text{CLT}}$ from (3.3).

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 data sets 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 data sets all contain sensitive information about individuals, and this fact necessitates privacy consideration in the training process. Code to reproduce the results using the `TensorFlow Privacy` library is available at https://github.com/tensorflow/privacy/tree/master/research/GDP_2019.

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. Throughout the experiments, the parameter δ we use always satisfies $\delta < 1/n$, where n is the number of training examples.

Figure 2. Tradeoffs between ϵ and δ for both CLT and MA, which henceforth denotes the moments accountant. The settings follow 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 (3.2) with $\mu = \mu_{\text{CLT}}$. The bottom right plot is consistent with the conclusion $\sigma > \bar{\sigma}$ shown in the cloud icon of Figure 1.

MNIST. The MNIST data set, created by LeCun and Cortes (2010), 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 Tensorflow (2019) and Abadi et al. (2016) on this data set. 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, the model from the third row preserves a decent amount of privacy since it is *not* always easy to distinguish $\mathcal{N}(0, 1)$ 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

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. *Note.* The accuracy is averaged over 10 independent runs. The hyperparameters in the first three rows are the same as in Tensorflow (2019). The μ in the 6th column is calculated using (3.1), which carries over to the 7th column via (3.2) with $\delta = 10^{-5}$. The number of epochs is equal to $T \times \text{mini-batch size}/n = pT$.

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 Dong, Roth, et al. (2019). The (ε, δ) -DP guarantees are plotted according to (2.1). 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.

of (ϵ, δ) -DP cannot be overcome by taking a larger δ , which, although giving 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. Note that the first row in Figure 3 corresponds to the first three rows in Table 1, and the third row in Figure 3 corresponds to the last three rows in Table 1. 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 data sets 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 in Dua and Graff (2017), the Adult income data set has been preprocessed into the LIBSVM format by Chang and Lin (2011). This data set 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 performance comparable to the baselines in the MLC++ library by Kohavi et al. (1996) in terms of test accuracy.

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	—	—	—
k NN (Cover & Hart, 1967)	—	79.7	—	—	—
naive Bayes	—	83.9	—	—	—
voted ID3 (Quinlan, 1986)	—	84.4	—	—	—
C4.5 (Quinlan, 2014)	—	84.5	—	—	—

Table 2. Results for NoisySGD on the Adult Income Data Set. *Note.* The ϵ parameters are with respect to $\delta = 10^{-5}$.

IMDb. We use the IMDb movie review data set from Maas et al. (2011) for binary sentiment classification (positive or negative reviews). The data set contains 25,000 training and 25,000 test examples. In our experiments, we preprocess the data set by including only 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 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 (https://www.tensorflow.org/tutorials/text/text_classification_rnn) as a baseline model. Table 3 reports the experimental results. Notably, the private neural networks perform comparably to the baseline model, at a 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	10	85.4	—	—	—
Hochreiter and Schmidhuber (1997)					

Table 3. Results for **NoisyAdam** on the IMDB Data Set, With $\delta = 10^{-5}$ Used in the Privacy Analyses

MovieLens. The MovieLens movie rating data set from Harper and Konstan (2016) is a benchmark data set for recommendation tasks. Our experiments consider the MovieLens 1M data set, 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 watched only 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 He et al. (2017). 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**.

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 (Sarwar et al., 2001)	—	0.923	—	—	—
global average	—	1.117	—	—	—

Table 4. Results for **NoisyAdam** on the MovieLens 1M Data Set, With $\delta = 10^{-6}$ Used in the Privacy Analyses. *Note.* CF stands for collaborative filtering.

Table 4 presents the numerical results of our neural networks as well as baseline models in the **Surprise** library from Hug (2017) in their default settings. The difference in RMSE between

the non-private networks and the private one is relatively large for the MovieLens 1M data set. Nevertheless, 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 $\sigma = 1.3, \bar{\sigma} = 1.06$, 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 a 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 $\epsilon = 4$ and the CLT approach achieves 96% under the same privacy budget.

5. DISCUSSION

In this article, 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. Employing numerical experiments, we show that the trained neural networks can be quite private from the f -DP viewpoint (for instance, 1.13-GDP¹¹) but are *not* in the (ϵ, δ) -DP sense due to overly 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 article 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 Lee and Kifer (2018) made considerable progress using concentrated differential privacy along this line. More generally, 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

¹¹This 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 $\mu + \mathcal{N}(0, 1)$.

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 ($\mu_{\text{CLT}} = 0.35$). 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 ($\mu_{\text{CLT}} = 1.78$). The right plots show the privacy loss during the training process in terms of the ϵ spending with respect to $\delta = 10^{-5}$.

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 (see Gupta et al. (2010)). Inspired by Lecuyer et al. (2019), another interesting direction is to explore the possible relationship between f -DP guarantees and adversarial robustness of neural networks. 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.

Disclosure Statement. The authors have no conflicts of interest to declare.

Acknowledgments. We are grateful to David Durfee, Ryan Rogers, Aaron Roth, and Qinqing Zheng for stimulating discussions in the early stages of this work. We would also like to thank two anonymous referees for their constructive comments that improved the presentation of the paper. 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 and RF1AG063481.

REFERENCES

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy, In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/2976749.2978318>
- Abowd, J. M. (2018). The US Census Bureau adopts differential privacy, In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/3219819.3226070>
- Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy, In *Advances in Neural Information Processing Systems*. <http://papers.nips.cc/paper/9681-differential-privacy-has-disparate-impact-on-model-accuracy>
- Balle, B., Barthe, G., & Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences, In *Advances in Neural Information Processing Systems*. <http://papers.nips.cc/paper/7865-privacy-amplification-by-subsampling-tight-analyses-via-couplings-and-divergences>
- Balle, B., Barthe, G., Gaboardi, M., Hsu, J., & Sato, T. (2019). Hypothesis testing interpretations and Renyi differential privacy. *arXiv preprint arXiv:1905.09982*.
- Balle, B., & Wang, Y.-X. (2018). Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising, In *International Conference on Machine Learning*. <http://proceedings.mlr.press/v80/balle18a.html>
- Bassily, R., Smith, A., & Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds, In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE. <https://ieeexplore.ieee.org/abstract/document/6979031>
- Bun, M., Dwork, C., Rothblum, G. N., & Steinke, T. (2018). Composable and versatile privacy via truncated CDP, In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/3188745.3188946>
- Bun, M., & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds, In *Theory of Cryptography Conference*. Springer. <https://dl.acm.org/citation.cfm?id=3081298>
- Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 1–27. <https://dl.acm.org/doi/abs/10.1145/1961189.1961199>
- Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar), 1069–1109. <http://www.jmlr.org/papers/volume12/chaudhuri11a/chaudhuri11a.pdf>

- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. https://ieeexplore.ieee.org/abstract/document/1053964/?casa_token=WyLTsD8whAQAAAAA:RHP-SXY--IPWvpEaGfHcQjugplhzAE9JDgFgPTbOMGX0lXj2Ieg23sr
- Differential-Privacy-Team. (2017). *Learning with privacy at scale* (tech. rep.). Apple. <https://machinelearning.apple.com/research/learning-with-privacy-at-scale>
- Ding, B., Kulkarni, J., & Yekhanin, S. (2017). Collecting telemetry data privately, In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*. <http://papers.nips.cc/paper/6948-collecting-telemetry-data-privately>
- Dong, J., Durfee, D., & Rogers, R. (2019). Optimal differential privacy composition for exponential mechanisms and the cost of adaptivity. *arXiv preprint arXiv:1909.13830*.
- Dong, J., Roth, A., & Su, W. J. (2019). Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*.
- Dua, D., & Graff, C. (2017). UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul), 2121–2159. <http://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf>
- Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., & Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation, In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer. https://link.springer.com/chapter/10.1007/11761679_29
- Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis, In *Theory of Cryptography Conference*. Springer. https://link.springer.com/chapter/10.1007/11681878_14
- Dwork, C., & Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.
- Dwork, C., Rothblum, G. N., & Vadhan, S. (2010). Boosting and differential privacy, In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. <https://ieeexplore.ieee.org/abstract/document/5670947/>
- Dwork, C., Talwar, K., Thakurta, A., & Zhang, L. (2014). Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis, In *Proceedings of the forty-sixth annual ACM Symposium on Theory of Computing*. Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/2591796.2591883>
- Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response, In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery; arXiv preprint arXiv:1407.6981. <https://dl.acm.org/doi/abs/10.1145/2660267.2660348>
- Gupta, A., Ligett, K., McSherry, F., Roth, A., & Talwar, K. (2010). Differentially private combinatorial optimization, In *Proceedings of the twenty-first annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. <https://epubs.siam.org/doi/abs/10.1137/1.9781611973075.90>
- Harper, F. M., & Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 5(4), 19. <https://dl.acm.org/doi/abs/10.1145/2827872>
- He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering, In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. <https://dl.acm.org/doi/abs/10.1145/3038912.3052569>

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735>
- Hug, N. (2017). Surprise, a Python library for recommender systems. <http://surpriselib.com/>
- Kairouz, P., Oh, S., & Viswanath, P. (2017). The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6), 4037–4049. <http://proceedings.mlr.press/v37/kairouz15.html>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kohavi, R., Sommerfield, D., & Dougherty, J. (1996). 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*. IEEE. <https://ieeexplore.ieee.org/abstract/document/560457/>
- LeCun, Y., & Cortes, C. (2010). MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>
- Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). Certified robustness to adversarial examples with differential privacy, In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. <https://ieeexplore.ieee.org/abstract/document/8835364/>
- Lee, J., & Kifer, D. (2018). 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*. Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/3219819.3220076>
- Li, N., Qardaji, W., & Su, D. (2012). 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*. Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/2414456.2414474>
- Liu, C., He, X., Chanyaswad, T., Wang, S., & Mittal, P. (2019). Investigating statistical privacy frameworks from the perspective of hypothesis testing. *Proceedings on Privacy Enhancing Technologies*, 2019(3), 233–254. <https://content.sciendo.com/view/journals/popets/2019/3/article-p233.xml>
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics. <https://www.aclweb.org/anthology/P11-1015.pdf>
- McMahan, B., Andrew, G., Mironov, I., Papernot, N., Kairouz, P., Chien, S., & Erlingsson, Ú. (2018). A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210*.
- McMahan, B., Ramage, D., Talwar, K., & Zhang, L. (2017). Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.
- Mironov, I. (2017). Rényi differential privacy, In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE. <https://ieeexplore.ieee.org/abstract/document/8049725>
- Murtagh, J., & Vadhan, S. (2016). The complexity of computing the optimal composition of differential privacy, In *Theory of Cryptography Conference*. Springer. https://link.springer.com/chapter/10.1007/978-3-662-49096-9_7

- Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets, In *2008 IEEE Symposium on Security and Privacy*. IEEE. <https://ieeexplore.ieee.org/abstract/document/4531148/>
- Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., & Talwar, K. (2016). Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.
- Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., & Erlingsson, Ú. (2018). Scalable private learning with PATE. *arXiv preprint arXiv:1802.08908*.
- Pichapati, V., Suresh, A. T., Yu, F. X., Reddi, S. J., & Kumar, S. (2019). AdaCliP: Adaptive clipping for private SGD. *arXiv preprint arXiv:1908.07643*.
- Quinlan, R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81–106. <https://link.springer.com/article/10.1007/BF00116251>
- Quinlan, R. (2014). *C4. 5: Programs for machine learning*. Elsevier.
- Rockafellar, R. T. (1970). *Convex analysis* (Vol. 28). Princeton University Press.
- Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *WWW*, 1, 285–295. <https://dl.acm.org/doi/pdf/10.1145/371920.372071>
- Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning, In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/2810103.2813687>
- Sommer, D., Meiser, S., & Mohammadi, E. (2018). Privacy loss classes: The Central Limit Theorem in differential privacy. *Proceedings on Privacy Enhancing Technologies 2019 (2)*, 245–269. <https://content.sciendo.com/view/journals/popets/2019/2/article-p245.xml>
- Song, S., Chaudhuri, K., & Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates, In *2013 IEEE Global Conference on Signal and Information Processing*. IEEE. <https://ieeexplore.ieee.org/abstract/document/6736861/>
- Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3), 98–110. <https://journals.sagepub.com/doi/abs/10.1111/j.1748-720x.1997.tb01885.x>
- Tensorflow. (2019). Tensorflow Privacy library. github.com/tensorflow/privacy
- Wang, Y.-X., Balle, B., & Kasiviswanathan, S. P. (2019). Subsampled Renyi differential privacy and analytical moments accountant, In *The 22nd International Conference on Artificial Intelligence and Statistics*. <http://proceedings.mlr.press/v89/wang19b.html>
- Wasserman, L., & Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489), 375–389. <https://amstat.tandfonline.com/doi/abs/10.1198/jasa.2009.tm08651>
- Xiang, L., Yang, J., & Li, B. (2019). Differentially-private deep learning from an optimization perspective, In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE. <https://ieeexplore.ieee.org/abstract/document/8737494/>
- Xu, Z., Shi, S., Liu, A. X., Zhao, J., & Chen, L. (2019). An adaptive and fast convergent approach to differentially private deep learning. *arXiv preprint arXiv:1912.09150*.
- Zhang, J., Zheng, K., Mou, W., & Wang, L. (2017). Efficient private ERM for smooth objectives, In *Proceedings of the 26th international joint conference on artificial intelligence*. IJCAI. <https://www.ijcai.org/Proceedings/2017/0548.pdf>

APPENDIX A. OMITTED DETAILS IN SECTION 2

We present the inequality (2.3) as the following proposition, which is given in Section 2 but not in the foundational work of Dong, Roth, et al. (2019).

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

$$T(M \circ \text{Sample}_p(S), M \circ \text{Sample}_p(S')) \geq pf + (1-p)\text{Id}.$$

Proof of Proposition A.1. We first write the two distributions $M \circ \text{Sample}_p(S)$ and $M \circ \text{Sample}_p(S')$ 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 Sample_p when applied to S is a bit string $\vec{b} = (b_1, \dots, b_n) \in \{0, 1\}^n$. Bit b_i depends 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 $\text{Sample}_p(S)$. 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, $M \circ \text{Sample}_p(S)$ can be written as the following mixture:

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

Similarly, $M \circ \text{Sample}_p(S')$ 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,

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

Note that $S_{\vec{b}} \cup \{x_0\}$ and $S_{\vec{b}}$ are neighbors, i.e., $M \circ \text{Sample}_p(S')$ 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 $T(P_i, Q_i) \geq f$ for all i , then

$$T\left(\sum \theta_i P_i, (1-p) \sum \theta_i P_i + p \sum \theta_i Q_i\right) \geq pf + (1-p)\text{Id}.$$

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 $T(P_i, Q_i) \geq f$ is a consequence of M being f -DP. The conclusion simply translates to

$$T(M \circ \text{Sample}_p(S), M \circ \text{Sample}_p(S')) \geq pf + (1-p)\text{Id},$$

which is what we want. The proof is complete. $\square$

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 $T(P_i, Q_i) \geq f$. Hence $\mathbb{E}_{Q_i} \phi \leq 1 - f(\mathbb{E}_{P_i} \phi)$ and so

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

Since f is convex, Jensen's inequality implies

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

□

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

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

This function converges to G_μ with $\mu = \nu\sqrt{e^{1/\sigma^2} - 1}$ as $T \rightarrow \infty$ provided $p\sqrt{T} \rightarrow \nu$. 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, $\approx 14\text{k}$ iterations. To best illustrate that convergence appears in an 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 A.1. $(pG_{1/\sigma} + (1-p)\text{Id})^{\otimes T}$ (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 remark that when σ is small, $\mu = \nu\sqrt{e^{1/\sigma^2} - 1}$ becomes large and yields challenges in the numerical computation of $(pG_{1/\sigma} + (1-p)\text{Id})^{\otimes T}$. We leave rigorous and complete study to future work.

APPENDIX 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

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

Proof. The proof is mostly done in the main text, except for the composition step. Let V be the vector space that all θ_t live in and $\widetilde{M} = M \circ \text{Sample}_p : X^n \times V \rightarrow V$ 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

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

Note that we cannot say M is f_p -DP because $T(M(S'), M(S))$ is not necessarily lower bounded by f_p . So we need a more specific composition theorem than that stated in Dong, Roth, et al. (2019).

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

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

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

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

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

The proof of this theorem is identical to that of Theorem 3.2 in Dong, Roth, et al. (2019).

Taking Algorithm 1 as an example, since

$$\begin{aligned} \text{NoisySGD} : X^n &\rightarrow V \times V \times \dots \times V \\ S &\mapsto (\theta_1, \theta_2, \dots, \theta_T) \end{aligned}$$

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

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

Moreover, $T(\text{NoisySGD}(S), \text{NoisySGD}(S')) \geq f^{-1}$. The two inequalities allow us to 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. $\square$

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 $p\sqrt{T} \rightarrow \nu$. Then we have the following uniform convergence as $T \rightarrow \infty$:

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

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

This theorem is a 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) $\int_0^1 (f'(x) + 1)^4 dx < +\infty$. Let $f_p = pf + (1-p)\text{Id}$ as usual. Furthermore, assume $p\sqrt{T} \rightarrow \nu$ as $T \rightarrow \infty$ for some constant $\nu > 0$. Then we have the uniform convergence

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

as $T \rightarrow \infty$, where $\chi^2(f) = \int_0^1 (f'(x))^2 dx - 1$.

Lemma B.1. We have

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

In order to prove Theorem 6, we need an even more general CLT. The first privacy CLT was introduced in Dong, Roth, et al. (2019). However, that version is valid only when each component trade-off function is symmetric, which is not true for $pG_{1/\sigma} + (1-p)\text{Id}$. In order to state the general CLT that applies to asymmetric trade-off functions, we must introduce the following functionals:

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

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

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

Then, we have

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

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

Proof of this theorem exactly mimics that of Theorem 3.5 in Dong, Roth, et al. (2019), which we omit here due to its length and tedium.

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

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

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

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

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

Since $T \sim p^{-2}$, we can consider $p^{-2}(\text{kl}(f_p) + \tilde{\text{kl}}(f_p))$ 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

$$\int_0^1 g(x)^4 \, dx < \infty.$$

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

$$\chi^2(f) = \int_0^1 (f'(x))^2 \, dx - 1 = \int_0^1 (f'(x) + 1)^2 \, dx = \int_0^1 g(x)^2 \, dx.$$

For the functional kl , by Lemma B.2,

$$\begin{aligned}\text{(B.1)} \quad \lim_{p \rightarrow 0^+} \frac{1}{p^2} (\text{kl}(f_p) + \tilde{\text{kl}}(f_p)) &= \lim_{p \rightarrow 0^+} \int_0^1 g(x) \cdot \frac{1}{p} \log(1 + pg(x)) \, dx \\ &= \int_0^1 g(x) \cdot \lim_{p \rightarrow 0^+} \frac{1}{p} \log(1 + pg(x)) \, dx \\ &= \int_0^1 g(x)^2 \, dx = \chi^2(f).\end{aligned}$$

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

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

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

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

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

For the functional κ_2 we have

$$\frac{1}{p^2} \kappa_2(f_p) = \int_0^1 \left[\frac{1}{p} \log(1 + pg(x)) \right]^2 dx.$$

By a similar dominating function argument,

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

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

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

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

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

Hence all the limits in Theorem 7 are verified 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. $\square$

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

$$\chi^2(P\|Q) := \mathbb{E}_P[(\frac{dQ}{dP} - 1)^2]$$

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 Dong, Roth, et al. (2019), which gives expressions for all F -divergence¹². 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:

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

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

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

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

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

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

□

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 notation used in the main text.

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

Let $f_T(\alpha) = \sup_{\varepsilon \geq 0} f_{\varepsilon, \delta_{\text{MA}}(\varepsilon)}(\alpha)$. It is supported by $f_{\varepsilon_T, \delta_{\text{MA}}(\varepsilon_T)}$ at α . Theorem 2 says this supporting function is smaller than that of $G_{\mu_{\text{CLT}}}$ at α by a strict gap. Taking the limit, $\limsup_{T \rightarrow \infty} f_T(\alpha)$ possesses at least that much gap from $G_{\mu_{\text{CLT}}}(\alpha)$, 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 $\limsup$ into $\liminf$.

Proposition B.4.

$$\liminf_{T \rightarrow \infty} \delta_{\text{MA}}(\varepsilon; \sigma, \frac{\nu}{\sqrt{T}}, T) - \delta_{\text{CLT}}(\varepsilon; \sigma, \nu) \geq e^\varepsilon \cdot \Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\right)$$

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

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

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

where $\alpha_{\text{GM}}(\lambda; \sigma, p)$ 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 $\text{GM}_{p, \sigma}$. Then

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

In Abadi et al. (2016), 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 $\varepsilon \geq 0$ if $\delta = \delta_{\text{MA}}(\varepsilon; \sigma, p, T)$. To evaluate the infimum, the domain is discretized¹³. This results in the numerical moment accountant method that is actually implemented. Since $\delta_{\text{nMA}}(\varepsilon; \sigma, p, T) \geq \delta_{\text{MA}}(\varepsilon; \sigma, p, T)$, Algorithm 1 is also (ε, δ) -DP with $\delta = \delta_{\text{nMA}}(\varepsilon; \sigma, p, T)$.

On the other hand, $\delta_{\text{CLT}}(\varepsilon; \sigma, \nu)$ is obtained by first observing Algorithm 1 is asymptotically μ_{CLT} -GDP with $\mu_{\text{CLT}} = \nu \cdot \sqrt{e^{1/\sigma^2} - 1}$ and then convert GDP to (ε, δ) -DP via formula (3.2), i.e., Algorithm 1 asymptotically satisfies (ε, δ) -DP where

$$\delta = \delta_{\text{CLT}}(\varepsilon; \sigma, \nu) = 1 + G_{\mu_{\text{CLT}}}^*(-e^\varepsilon).$$

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

¹³Code in tensorflow/privacy discretizes at [1.25, 1.5, 1.75, 2., 2.25, 2.5, 3., 3.5, ..., 63, 64, 128, 256, 512].

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

Lemma B.5.

$$T \cdot \alpha_{\text{GM}}(\lambda; \sigma, p) \geq \log \int_0^1 |f'_T(x)|^{\lambda+1} dx.$$

Let $x_T \in (0, 1)$ be the point for which $f'_T(x_T) = -e^\varepsilon$ (or $x_T \in \partial f_T^*(-e^\varepsilon)$ if readers worry about differentiability). We have

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

It is clear that $|f'_T(x)| \geq e^\varepsilon$ for $0 \leq x \leq x_T$.

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

$$\begin{aligned} \delta_{\text{MA}}(\varepsilon; \sigma, p, T) &= \inf_{\lambda > 0} \exp(T \cdot \alpha_{\text{GM}}(\lambda; \sigma, p) - \lambda \varepsilon) \\ &\geq \inf_{\lambda > 0} e^{-\lambda \varepsilon} \int_0^1 |f'_T(x)|^{\lambda+1} dx \\ &> \inf_{\lambda > 0} \int_0^{x_T} |f'_T(x)|^{\lambda+1} e^{-\lambda \varepsilon} dx \\ &= \inf_{\lambda > 0} \int_0^{x_T} |f'_T(x)| \cdot |f'_T(x)|^\lambda e^{-\lambda \varepsilon} dx \\ &\geq \inf_{\lambda > 0} \int_0^{x_T} |f'_T(x)| \cdot (e^\varepsilon)^\lambda e^{-\lambda \varepsilon} dx \\ &= f_T(0) - f_T(x_T) \\ &= 1 - f_T(x_T) \\ &= (1 - f_T(x_T) - e^\varepsilon x_T) + e^\varepsilon x_T \\ &= 1 + f_T^*(-e^\varepsilon) + e^\varepsilon x_T. \end{aligned}$$

In summary, we have

$$(B.2) \quad \delta_{\text{MA}}(\varepsilon; \sigma, p, T) > 1 + f_T^*(-e^\varepsilon) + e^\varepsilon x_T.$$

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

Lemma B.6.

$$\lim_{T \rightarrow \infty} x_T = x^* := \Phi\left(-\frac{\varepsilon}{\mu_{\text{CLT}}} - \frac{\mu_{\text{CLT}}}{2}\right).$$

Combining these results, we can take limits on both sides of (B.2):

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

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

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

Alternatively,

$$(B.3) \quad \lambda D_{\lambda+1}(Q\|P) = \log \int_0^1 |f'(x)|^\lambda dx.$$

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

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

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

Consequently,

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

The last step is the tensorization identity of Rényi divergence.

On the other hand, notice that $pG_\mu + (1-p)\text{Id} = T(P, \text{GM}_{p,\sigma})$, where we continue the use of notation $P = \mathcal{N}(0, 1)$, $Q = \mathcal{N}(\frac{1}{\sigma}, 1)$ and $\text{GM}_{p,\sigma} = pQ + (1-p)P$. We have¹⁴

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

Using (B.3), we have

$$\begin{aligned} T \cdot \alpha_{\text{GM}}(\lambda; \sigma, p) &\geq \lambda D_{\lambda+1}((pQ + (1-p)P)^T\|P^T) \\ &= \log \int_0^1 |f'_T(x)|^\lambda dx. \end{aligned}$$

□

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

Since f_T converges uniformly to $G_{\mu_{\text{CLT}}}$ over $[0, 1]$, we have uniform convergence $f_T^* \rightarrow G_{\mu_{\text{CLT}}}^*$. By convexity of these functions, the convergence also implies the convergence of derivatives (see Theorem 25.7 of Rockafellar (1970)), namely,

$$\nabla f_T^* \rightarrow \nabla G_{\mu_{\text{CLT}}}^*.$$

Therefore,

$$x_T = \nabla f_T^*(-e^\varepsilon) \rightarrow \nabla G_{\mu_{\text{CLT}}}^*(-e^\varepsilon).$$

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

$$-e^\varepsilon = G'_{\mu_{\text{CLT}}}(x^*).$$

We can solve for x^* using the expression of G_μ (3.2). After some algebra, we have

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

The proof is complete. □

¹⁴Here we abuse the symbol T to denote two distinct objects, both the trade-off function and the number of iterations.