A Better Bound Gives a Hundred Rounds: Enhanced Privacy Guarantees via f-Divergences

Shahab Asoodeh[†], Jiachun Liao[*], Flavio P. Calmon[†], Oliver Kosut[*], Lalitha Sankar[*]
[†]Harvard University, {shahab, flavio}@seas.harvard.edu
[*]Arizona State University, {jiachun.liao, okosut, lalithasankar}@asu.edu

***Abstract*—We derive the optimal differential privacy (DP) parameters of a mechanism that satisfies a given level of Rényi differential privacy (RDP). Our result is based on the joint range of two f-divergences that underlie the approximate and the Rényi variations of differential privacy. We apply our result to the moments accountant framework for characterizing privacy guarantees of stochastic gradient descent. When compared to the state-of-the-art, our bounds may lead to about 100 more stochastic gradient descent iterations for training deep learning models for the same privacy budget.**

I. Introduction

Differential privacy (DP) [1] has become the *de facto* standard for privacy-preserving data analytics. Intuitively, a (potentially randomized) algorithm is said to be *differentially private* if its output does not vary significantly with small perturbations of the input. DP guarantees are usually cast in terms of properties of the *information density* [2] of the output of the algorithm conditioned on a given input—referred to as the *privacy loss variable* in the DP literature.

Several methods have recently been proposed to ensure differentially private training of machine learning (ML) models [3–8]. Here, the parameters of the model determined by a learning algorithm (e.g., weights of a neural network or coefficients of a regression) are sought to be differentially private with respect to the data used for fitting the model (i.e. the *training* data). When the model parameters are computed by applying stochastic gradient descent (SGD) to minimize a given loss function, DP can be ensured by directly adding noise to the gradient. The empirical and theoretical flexibility of this noise-adding procedure for ensuring DP was demonstrated, for example, in [3, 4]. This method is currently being used for privacy-preserving training of large-scale ML models in industry, see e.g., the implementation of [9] in the Google's open-source TensorFlow Privacy framework [10].

Not surprisingly, for a fixed training dataset, privacy deteriorates with each SGD iteration. In practice, the DP constraints are set *a priori*, and then mapped to a permissible number of SGD iterations for fitting the model parameters. Thus, a key question is: *given a DP constraint, how many iterations are allowed before the SGD algorithm is no longer private?* The main challenge in determining the DP guarantees provided by noise-added SGD is keeping track of the evolution of the privacy loss random variable during subsequent gradient descent iterations. This can be done, for example, by invoking advanced composition theorems for DP, such as [11, 12]. Such composition results, while theoretically significant, may be difficult to apply to the SGD setting due to their generality (e.g., they do not take into account the noise distribution used by the privacy mechanism).

Recently, Abadi et al. [3] circumvented the use of DP composition results by developing a method called *moments accountant* (MA). Instead of dealing with DP directly, the MA approach provides privacy guarantees in terms of *Rényi differential privacy* (RDP) [13] for which composition has a simple linear form. Once the privacy guarantees of the SGD execution are determined in terms of RDP, they are mapped back to DP guarantees via a conversion result between DP and RDP [3, Theorem 2]. This approach renders tighter DP guarantees than those obtained from advanced composition theorems (see [3, Figure 2]).

Our Contributions: We provide a framework which settles the *optimal* conversion from RDP to DP, and thus further enhances the privacy guarantee obtained by the MA approach. Our technique relies on the information-theoretic study of joint range of f-divergences: we first describe both DP and RDP using two certain types of the f-divergences, namely E_λ and χ^α divergences (see Section II). We then apply [14, Theorem 8] to characterize the joint range of these two f-divergences which, in turn, leads to the "optimal" conversion from RDP to DP (see Section III). Specifically, this optimal conversion allows us to derive bounds on the number of SGD iterations for a given DP constraint in the context of Gaussian perturbation of the gradient. Our result improves upon the state-of-the-art [3] by allowing more training iterations (often hundreds more) for the same privacy budget, and thus providing higher utility for free (see Section IV). In the interest of space, we delegate the proofs to the full version of this paper [15].

II. Preliminaries and Problem Setup

In this section, we give several definitions and basic results that will be used in the subsequent sections.

Let $\mathbb{D}$ be some universe of all possible datasets and $(\mathbb{X}, \mathcal{F})$ be a measurable space with Borel σ-algebra $\mathcal{F}$. A mechanism $\mathcal{M} : \mathbb{D} \to \mathcal{P}(\mathbb{X})$ assigns a probability distribution $\mathcal{M}_d$ to each dataset d where $\mathcal{P}(\mathbb{X})$ denotes the set of all probability measures on $\mathbb{X}$. Two datasets d and d' are said to be neighboring

This work was supported in part by NSF under grants CIF 1922971, 1815361, 1742836, 1900750, and CIF CAREER 1845852.

978-1-7281-6432-8/20/$31.00 ©2020 IEEE

(denoted by $d \sim d'$) if their Hamming distance is one. For any pair of neighboring datasets d and d', the privacy loss random variable is defined as $L_{d,d'} \coloneqq \log \frac{\mathcal{M}_d(Y)}{\mathcal{M}_{d'}(Y)}$ where $Y \sim \mathcal{M}_d$.

Definition 1. A mechanism $\mathcal{M} : \mathbb{D} \to \mathcal{P}(\mathbb{X})$ is said to be

- (ε, δ)-DP for given $\delta \in [0, 1)$ and $\varepsilon \geq 0$, if
$$\sup_{A \in \mathcal{F}, d \sim d'} \mathcal{M}_d(A) - e^{\varepsilon}\mathcal{M}_{d'}(A) \leq \delta. \tag{1}$$
- (α, γ)-RDP for a given $\alpha > 1$, if
$$\sup_{d \sim d'} D_\alpha(\mathcal{M}_d \| \mathcal{M}_{d'}) \leq \gamma, \tag{2}$$
where $D_\alpha(P\|Q) \coloneqq \frac{1}{\alpha-1} \log \mathbb{E}_Q \left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right)^{\alpha} \right]$ denotes the Rényi divergence of order α between P and Q in $\mathcal{P}(\mathbb{X})$.

It can be shown that (1) is implied if the tail event $\{L_{d,d'} > \varepsilon\}$ occurs with probability at most δ for all $d \sim d'$, and (2) is implied if (and only if) the α-moment of $L_{d,d'}$ is upper bounded by γ. Built on this insight, the MA restricts the α-moment of $L_{d,d'}$ for *all* $\alpha > 1$.

As mentioned earlier, RDP (and hence MA) composes linearly, as opposed to the strong composition theorem for DP which is known to be loose for many practical mechanisms, including Gaussian. With this clear advantage comes a shortcoming: RDP suffers from the lack of operational interpretation, see e.g., [16]. To address this issue, the RDP guarantee is often translated into a DP guarantee via the following result.

Theorem 1. *([3, Thm 2], [13, Prop 3]) If the mechanism $\mathcal{M}$ is (α, γ)-RDP, then it satisfies (ε, δ)-DP for any $\varepsilon > \gamma$ and*
$$\delta = e^{-(\alpha-1)(\varepsilon-\gamma)}. \tag{3}$$

For MA, this constraint must hold for *all* $\alpha > 1$ and thus it leads to (ε, δ)-DP for
$$\delta = \inf_{\alpha > 1} e^{-(\alpha-1)(\varepsilon-\gamma(\alpha))}, \tag{4}$$
where $\gamma(\alpha) = \sup_{d \sim d'} D_\alpha(\mathcal{M}_d \| \mathcal{M}_{d'})$ and the dependence on α is made clear. Since $\alpha \mapsto (\alpha - 1)D_\alpha(P\|Q)$ is convex [17, Corollary 2] for any pair of probability measures P and Q, the above minimization is a log-convex problem and hence can be solved to an arbitrary accuracy. We will show in Section IV that this minimization has a simple form for Gaussian mechanisms and can be solved analytically.

Theorem 1 establishes a relationship for converting RDP to DP that is extensively used in several recent differentially private ML applications, e.g., [7, 18–23] to name a few. However, despite its extensive use, this relationship is loose. For instance, as we see later, for Gaussian mechanisms this relationship holds for $\varepsilon \to 0$ only when the variance of noise goes to infinity. In Section III, we present the *optimal* conversion from RDP to DP, thus improving the privacy guarantees of recent ML applications involving MA. Specifically, we investigate the following two closely-related questions:

Question One: *Given an (α, γ)-RDP mechanism $\mathcal{M}$, what are the smallest ε and δ such that $\mathcal{M}$ is (ε, δ)-DP?*

We show in Section III that such minimal ε and δ can be obtained via a simple one-variable optimization problem.

We then turn our attention to privacy guarantees in applications where the data may need to be accessed many times (say T times) such as with SGD. In such applications, each data access renders the application of a privacy mechanism, i.e., T privacy mechanisms are applied. An oft-used model, that we also adopt here, is one in which each mechanism adds Gaussian noise with pre-specified variance σ^2. This model is referred to as the T-fold homogeneous composition of Gaussian mechanisms each with variance σ^2.

Question Two: *Given $\varepsilon \geq 0$, $\delta \in [0, 1]$ and σ^2, what is the largest T such that the T-fold homogeneous composition of Gaussian mechanism with variance σ^2 is (ε, δ)-DP?*

The linearity of the RDP guarantee (in T) and the optimal conversion from RDP to DP (addressed in Question One) enable us to express the answer to this question as a minimization (over $\alpha > 1$) of the answer to Question One, analogous to (4). Although this additional minimization significantly complicates the analytic derivation, we nevertheless obtain tight bounds for the largest T provided that δ is sufficiently small. Details are deferred to Section IV.

To mathematically formulate these goals, we need the following definitions and basic results.

Definition 2. ([24, 25]) Given two probability distributions P and Q and a real-valued convex function f satisfying $f(1) = 0$, the f-divergence between P and Q is given by
$$D_f(P\|Q) \coloneqq E_Q \left[f \left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right) \right]. \tag{5}$$

We frequently use two particular instances of f-divergences. Given $\lambda \geq 1$, the f-divergence associated with $f(t) = (t - \lambda)_+ = \max\{t - \lambda, 0\}$, is called E_λ-divergence (also known as *hockey-stick* divergence [26]) and given by
$$\mathsf{E}_\lambda(P\|Q) = \int (\mathrm{d}P - \lambda \mathrm{d}Q)_+ = \sup_{A \in \mathcal{F}} \left[P(A) - \lambda Q(A) \right]. \tag{6}$$

Also, for any $\alpha > 1$, the f-divergence associated with $f(t) = \frac{1}{\alpha-1}(t^\alpha - 1)$ is denoted by[1] $\chi^\alpha(P\|Q)$. Note that $D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \left(1 + (\alpha - 1)\chi^\alpha(P\|Q)\right)$ for a pair of probability distributions P and Q.

It is shown in [28], [29] that
$$\mathcal{M} \text{ is } (\varepsilon, \delta)\text{-DP} \iff \sup_{d \sim d'} \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d \| \mathcal{M}_{d'}) \leq \delta. \tag{7}$$

Similarly, it can be verified that:
$$\mathcal{M} \text{ is } (\alpha, \gamma)\text{-RDP} \iff \sup_{d \sim d'} \chi^\alpha(\mathcal{M}_d \| \mathcal{M}_{d'}) \leq \chi(\gamma), \tag{8}$$
where
$$\chi(\gamma) \coloneqq \frac{e^{(\alpha-1)\gamma} - 1}{\alpha - 1}. \tag{9}$$

For any $\alpha > 1$ and non-negative γ, we let $\mathbb{M}_\alpha(\gamma)$ be the set of all (α, γ)-RDP mechanisms $\mathcal{M}$. This definition, together

[1] χ^α-divergence is also referred to as α-Hellinger divergence, see, e.g., [27].

with (7), enables us to precisely formulate Question One. If a mechanism is (α,γ)-RDP then the smallest δ, for a given ε, such that it is (ε,δ)-DP is upper bounded by

$$\delta_\alpha^\varepsilon(\gamma) := \sup_{\mathcal{M}\in\mathbb{M}_\alpha(\gamma)} \sup_{d\sim d'} \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'}) \tag{10}$$

Given α,γ and ε, this quantity corresponds to the smallest δ guaranteed by the worst mechanism in $\mathbb{M}_\alpha(\gamma)$, thus establishing an upper bound for the smallest δ such that a given (α,γ)-RDP mechanism is (ε,δ)-DP. In fact, we can write

$$\delta_\alpha^\varepsilon(\gamma) = \inf\left\{\delta\in[0,1] : \forall\mathcal{M}\in\mathbb{M}_\alpha(\gamma) \text{ is } (\varepsilon,\delta)\text{-DP}\right\}. \tag{11}$$

Such quantity is key for indicating the "optimality" of a conversion from RDP to DP. It may be equivalently identified by closely related quantities

$$\gamma_\alpha^\varepsilon(\delta) := \sup\left\{\gamma\geq 0 : \forall\mathcal{M}\in\mathbb{M}_\alpha(\gamma) \text{ is } (\varepsilon,\delta)\text{-DP}\right\}, \tag{12}$$

or

$$\varepsilon_\alpha^\delta(\gamma) := \inf\left\{\varepsilon\geq 0 : \forall\mathcal{M}\in\mathbb{M}_\alpha(\gamma) \text{ is } (\varepsilon,\delta)\text{-DP}\right\}. \tag{13}$$

In the next section, we exploit (7)–(8) to compute or bound these quantities.

III. Optimal Conversion from RDP to DP

In this section, we aim at computing the fundamental worst-case DP privacy parameter guaranteed by an (α,γ)-RDP mechanism; a quantity defined in (10). To this goal, we first show that this quantity is an upper boundary of a convex set defined by E_λ-divergence and χ^α-divergence and then invoke the well-known result of [14] about the joint range of f-divergences.

First note that, according to (8), the set $\mathbb{M}_\alpha(\gamma)$ can be equivalently characterized by the constraint $\chi^\alpha(\mathcal{M}_d\|\mathcal{M}_{d'})\leq\chi(\gamma)$, where $\chi(\gamma)$ is defined in (9). Hence, the quantity in (10) in fact constitutes the upper boundary of the convex set

$$\mathcal{R}_\alpha := \left\{\left(\chi^\alpha(\mathcal{M}_d\|\mathcal{M}_{d'}), \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'})\right)\,\middle|\,\forall\mathcal{M}, d\sim d'\right\}. \tag{14}$$

This simple observation has two key implications. First, the convexity of this set implies that the map $\gamma\mapsto\delta_\alpha^\varepsilon(\gamma)$, defined in (10), can be alternatively expressed by $\delta\mapsto\gamma_\alpha^\varepsilon(\delta)$, defined in (12). Note also that $\gamma_\alpha^\varepsilon$ can be equivalently written as

$$\gamma_\alpha^\varepsilon(\delta) = \inf_{\mathcal{M}:\mathbb{D}\to\mathcal{P}(\mathbb{X})} \inf_{d\sim d'} \chi^{-1}(\chi^\alpha(\mathcal{M}_d\|\mathcal{M}_{d'})) \tag{15}$$
$$\text{s.t. } \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'})\geq\delta,\ \forall d\sim d',$$

where $\chi^{-1}(\cdot)$ is the inverse of $\chi(\cdot)$, defined in (9), and is given by $\chi^{-1}(t)=\frac{1}{\alpha-1}\log(1+(\alpha-1)t)$. Second, to derive the upper boundary of $\mathcal{R}_\alpha$ (and thus $\gamma_\alpha^\varepsilon(\delta)$) it suffices to characterize $\mathcal{R}_\alpha$. This allows us to cast the problem of converting from (α,γ)-RDP to (ε,δ)-DP as characterizing the joint range of E_λ and χ^α divergences. To tackle the latter problem, we refer to [14] whose main result is as follows.

Theorem 2. *([14, Theorem 8]) The joint range of divergences D_f and D_g satisfies*

$$\left\{(D_f(P\|Q), D_g(P\|Q))\,\middle|\,P,Q\in\mathcal{P}(\mathbb{X})\right\} = \mathsf{conv}(\mathcal{B})$$

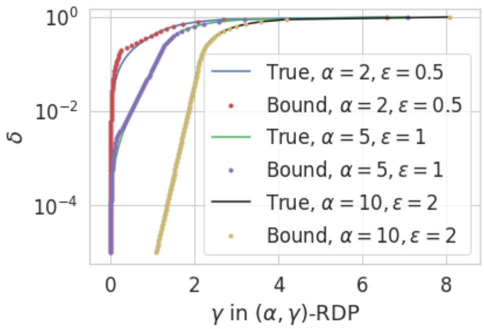

Fig. 1. True values (solid curves), obtained via numerically solving convex optimization problem (16), versus the bounds (dotted curves) obtained from Theorem 4 for three pairs of (α,ε).

where $\mathsf{conv}(\cdot)$ *denotes the convex hull operator and*

$$\mathcal{B} := \left\{(D_f(P_\mathsf{b}\|Q_\mathsf{b}), D_g(P_\mathsf{b}\|Q_\mathsf{b}))\,\middle|\,P_\mathsf{b},Q_\mathsf{b}\in\mathcal{P}(\{0,1\})\right\}.$$

This theorem provides an efficient method for characterizing the joint range of any pair of f-divergences. Specialized to χ^α and E_λ divergences, this theorem therefore enables us to characterize $\mathcal{R}_\alpha$ and thus derive $\gamma_\alpha^\varepsilon(\delta)$. This is formalized in Theorem 3 in which we establish a simple variational formula for $\gamma_\alpha^\varepsilon(\delta)$ involving a one-parameter log-convex minimization. Hence, the optimization (15), which can potentially be of significant complexity, turns into a simple tractable problem.

Theorem 3. *For any $\alpha>1$, $\varepsilon\geq 0$ and $\delta\in[0,1)$,*

$$\gamma_\alpha^\varepsilon(\delta) = \varepsilon + \min_{p\in(\delta,1)} \frac{1}{\alpha-1}\log\left(p^\alpha(p-\delta)^{1-\alpha} + \bar{p}^\alpha(e^\varepsilon-p+\delta)^{1-\alpha}\right), \tag{16}$$

where $\bar{p}:=1-p$.

It can be shown the term inside the logarithm is convex in p and hence this optimization problem can be numerically solved with an arbitrary accuracy. It seems, however, not simple to analytically derive $\gamma_\alpha^\varepsilon(\delta)$. Nevertheless, we obtain a tight lower bound in the following theorem.

Theorem 4. *For any $\varepsilon\geq 0$ and $\alpha>1$, we have*

$$\gamma_\alpha^\varepsilon(0) = 0,$$
$$\gamma_\alpha^\varepsilon(\delta) = \varepsilon-\log(1-\delta), \qquad \text{if } \alpha\delta\geq 1, \tag{17}$$
$$\gamma_\alpha^\varepsilon(\delta) \geq \max\{g(\alpha,\varepsilon,\delta), f(\alpha,\varepsilon,\delta)\}, \quad \text{if } 0<\alpha\delta<1, \tag{18}$$

where

$$g(\alpha,\varepsilon,\delta) := \varepsilon - \frac{1}{\alpha-1}\log\frac{\zeta_\alpha}{\delta},$$

with $\zeta_\alpha := \frac{1}{\alpha}\left(1-\frac{1}{\alpha}\right)^{\alpha-1}$ and

$$f(\alpha,\varepsilon,\delta) := \varepsilon + \frac{1}{\alpha-1}\log\left((e^\varepsilon-\alpha\delta)\left(\frac{\delta-1}{\delta-e^\varepsilon}\right)^\alpha + \alpha\delta\right).$$

In Fig. 1, we numerically solve (16) for three pairs of (α,ε) and compare them with their corresponding bounds obtained from Theorem 4, highlighting the tightness of the above lower bound.

As indicated earlier and illustrated in Fig. 1, the lower bound in $\gamma^{\varepsilon}_{\alpha}(\delta)$ in Theorem 4 is translated into an upper bound on $\delta^{\varepsilon}_{\alpha}(\gamma)$. In practice, it is often more appealing to design differentially private mechanisms with a hard-coded value of δ (as opposed to the fixed ε). To address this practical need, we convert the lower bound in Theorem 4 to an upper bound on $\varepsilon^{\delta}_{\alpha}(\gamma)$.

Lemma 1. *For $\alpha > 1$ and $\gamma \geq 0$, we have*

$$\varepsilon^{\delta}_{\alpha}(\gamma) = \big(\gamma + \log(1-\delta)\big)_{+}, \quad \textit{if} \quad \alpha\delta \geq 1, \tag{19}$$

and if $0 < \alpha\delta < 1$

$$\varepsilon^{\delta}_{\alpha}(\gamma) \leq \frac{1}{\alpha-1}\min\left\{\left((\alpha-1)\gamma - \log\frac{\delta}{\zeta_{\alpha}}\right)_{+}, \log\left(\frac{(\alpha-1)\chi(\gamma)}{\alpha\delta}+1\right)\right\}, \tag{20}$$

where $\chi(\gamma)$ is defined in (9). Moreover, $\varepsilon^{\delta}_{\alpha}(0) = 0$.

The proof of this lemma is based on writing the first-order approximation for f in terms of δ, thereby allowing us to invert the inequality (18). Note that g is a linear function of ε and hence invertible. It must be mentioned that Balle et al. [16, Theorem 21] has recently proved the bound $\varepsilon^{\delta}_{\alpha}(\gamma) \leq \gamma - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_{\alpha}}$, via a fundamentally different approach which is weaker than Lemma 1.

Remark 1. As an important special case, this lemma demonstrates that an (α,γ)-RDP mechanism provides $(0,\delta)$-DP guarantee if $1-e^{-\gamma} < \frac{1}{\alpha}$ and $\delta \in [\zeta_{\alpha}e^{(\alpha-1)\gamma}, \frac{1}{\alpha}]$, see [15] for the detailed derivation and also another sufficient condition for $(0,\delta)$-DP. Notice that this is significantly stronger than what would be obtained from Theorem 1: $\varepsilon^{\delta}_{\alpha}(\gamma) \leq \gamma - \frac{1}{\alpha-1}\log\delta$ from which $(0,\delta)$-DP cannot be achieved.

IV. Moments Accountant and Gaussian Mechanisms

Moments accountant (MA) was recently proposed by Abadi et al. [3] as a method to bypass advanced composition theorems [11, 12]. Given a mechanism $\mathcal{M}$, the T-fold adaptive homogeneous composition $\mathcal{M}^{(T)}$ is a mechanism that consists of T copies of $\mathcal{M}$, i.e., $(\mathcal{M}^1, \dots, \mathcal{M}^T)$ such that the input of $\mathcal{M}^i$ may depend on the outputs of $\mathcal{M}^1, \dots, \mathcal{M}^{i-1}$. Determining the privacy parameters of $\mathcal{M}^{(T)}$ in terms of those of $\mathcal{M}$ and T is an important problem in practice and thus has been the subject of an extensive body of research, see e.g., [3, 11, 12, 22].

Advanced composition theorems [11, 12] are well-known results that provide the DP parameters of $\mathcal{M}^{(T)}$ for general mechanisms. However, they can be loose and do not take into account the particular noise distribution under consideration (i.e., Gaussian noise). MA was shown to significantly improve upon advanced composition theorems in specific applications such as SGD. The cornerstone of MA is the linear composability of RDP: If $\mathcal{M}^1, \dots, \mathcal{M}^T$ are (α,γ)-RDP, then it is shown [3, Theorem 2] that $\mathcal{M}^{(T)}$ is $(\alpha,\gamma T)$-RDP. This result is then translated into DP privacy parameters via Theorem 1. Since the above composability and conversion hold for all $\alpha > 1$, one can obtain the *best* privacy parameters by optimizing over α according to (4). More precisely, $\mathcal{M}^{(T)}$ is (ε,δ)-DP with

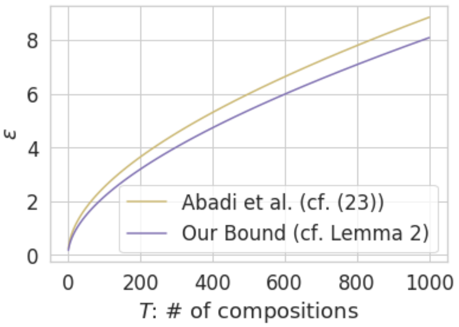

Fig. 2. The comparison of our bound in Lemma 2 on $\varepsilon^{\delta}(\rho,T)$ with (23) for $\sigma = 20$ and $\delta = 10^{-5}$.

$$\delta = \inf_{\alpha>1} e^{-(\alpha-1)(\varepsilon-\gamma(\alpha)T)}, \tag{21}$$

for a given ε or equivalently,

$$\varepsilon = \inf_{\alpha>1} \gamma(\alpha)T - \frac{1}{\alpha-1}\log\delta, \tag{22}$$

for a given δ, where $\gamma(\alpha) = \sup_{d\sim d'} D_{\alpha}(\mathcal{M}_d\|\mathcal{M}_{d'})$ is the RDP parameter of the constituent mechanism $\mathcal{M}$.

For the rest of this section, we assume $\mathcal{M}$ is a Gaussian mechanism and apply Theorem 4 and Lemma 1 in place of (21) and (22) respectively, in order to improve the DP privacy parameters obtained by MA.

A. Bounds on Privacy Parameters of Gaussian Composition

Let $f : \mathbb{D} \to \mathbb{R}^n$ be a query function and $\mathcal{M}$ be a Gaussian mechanism with variance σ^2, i.e., $\mathbb{X} = \mathbb{R}^n$ and $\mathcal{M}_d = \mathcal{N}(f(d), \sigma^2\mathrm{I}_n)$ for each $d \in \mathbb{D}$. For simplicity, we assume that f has unit L_2-sensitivity, i.e., $\sup_{d\sim d'}\|f(d)-f(d')\|_2 = 1$. Since

$$\sup_{d\sim d'} D_{\alpha}(\mathcal{M}_d\|\mathcal{M}_{d'}) = \frac{\alpha}{2\sigma^2}\sup_{d\sim d'}\|f(d)-f(d')\|_2 = \frac{\alpha}{2\sigma^2},$$

it follows that $\mathcal{M}$ is $(\alpha,\gamma(\alpha))$-RDP for all $\alpha > 1$ where $\gamma(\alpha) = \rho\alpha$ and $\rho = \frac{1}{2\sigma^2}$. In light of the linear composability of RDP, we obtain that $\mathcal{M}^{(T)}$ the T-fold adaptive composition of $\mathcal{M}$ is $(\alpha,\gamma(\alpha)T)$-RDP. In this setting, the optimization problem given in (22) can be solved analytically. Consequently, MA implies that $\mathcal{M}^{(T)}$ is (ε,δ)-DP for any $\delta \in (0,1)$ and

$$\varepsilon = \inf_{\alpha>1} \gamma(\alpha)T - \frac{1}{\alpha-1}\log\delta = \rho T + \sqrt{4\rho T\log\frac{1}{\delta}} \tag{23}$$

We next use the machinery developed in the previous section to improve (23) the DP parameter of $\mathcal{M}^{(T)}$ implied by MA. To do so, define

$$\varepsilon^{\delta}(\rho,T) := \inf_{\alpha>1}\varepsilon^{\delta}_{\alpha}(\rho\alpha T). \tag{24}$$

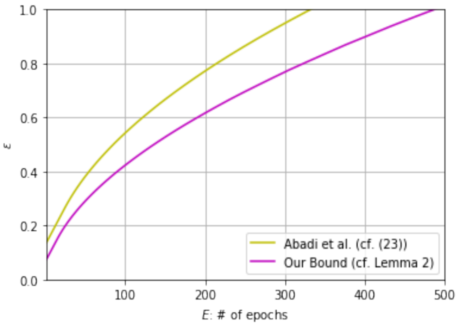

Fig. 3. Privacy parameter ε of noisy SGD where the Gaussian noise with $\sigma = 4$ is added to the gradient of mini-batches with size rate (or sub-sampling rate) 0.001. Also, δ is assumed to be 10^{-5}.

Thus, $\mathcal{M}^{(T)}$ is $(\varepsilon^\delta(\rho,T),\delta)$-DP for any $\delta\in(0,1)$. Invoking Lemma 1, we can obtain a bound $\varepsilon^\delta(\rho,T)$.

Lemma 2. *The T-fold adaptive homogeneous composition of the Gaussian mechanism with variance σ^2 is $(\varepsilon^\delta(\rho,T),\delta)$-DP with $\delta\in(0,1)$ and*

$$\varepsilon^\delta(\rho,T)\le\min\left\{\varepsilon_0(\rho,T),\ \varepsilon_1(\rho,T),\ \left(\frac{\rho T}{\delta}+\log(1-\delta)\right)_+\right\}, \tag{25}$$

where $\rho=\frac{1}{2\sigma^2}$ and

$$\varepsilon_0(\rho,T):=\inf_{\alpha\in(1,\frac{1}{\delta}]}\left(\rho\alpha T-\frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right)_+, \tag{26}$$

$$\varepsilon_1(\rho,T):=\inf_{\alpha\in(1,\frac{1}{\delta}]}\frac{1}{\alpha-1}\log\left(1+\frac{e^{\rho\alpha(\alpha-1)T}-1}{\alpha\delta}\right), \tag{27}$$

and ζ_α is as defined in Theorem 4.

The bound given in this lemma can shed light on the optimal variance of the Gaussian mechanism $\mathcal{M}$ required to ensure that $\mathcal{M}^{(T)}$ is (ε,δ)-DP (cf. Question Two in Section II). To put our result about the variance in perspective, we first mention two previously-known bounds on σ^2. Advanced composition theorems (see, e.g., [11, Theorem III.3]) require $\sigma^2=\Omega(\frac{T\log(1/\delta)\log(T/\delta)}{\varepsilon^2})$. Abadi et al. [3, Theorem 1] improved this result by showing that σ^2 suffices to be linear in T; more precisely, $\sigma^2=\Omega(\frac{T\log(1/\delta)}{\varepsilon^2})$. To have a better comparison with our final result, we write this result more explicitly. It follows from (21) or (22) that

$$\begin{aligned}\frac{T}{2\sigma^2}&\le\sup_{\alpha>1}\frac{\varepsilon}{\alpha}+\frac{1}{\alpha(\alpha-1)}\log\delta\\&=\varepsilon-2\log\delta-2\sqrt{(\varepsilon-\log\delta)\log\frac{1}{\delta}},\end{aligned}$$

and hence assuming δ is sufficiently small, we obtain

$$\sigma^2\ge\frac{2T}{\varepsilon^2}\log\frac{1}{\delta}+\frac{T}{\varepsilon}+O\left(\frac{1}{\log\delta^{-1}}\right). \tag{28}$$

We are now in order to state our result.

Theorem 5. *The T-fold adaptive homogeneous composition of a Gaussian mechanism with variance σ^2 is (ε,δ)-DP, for $\varepsilon>2\delta\log\frac{1}{\delta}$, if*

$$\begin{aligned}\sigma^2\ge\frac{2T}{\varepsilon^2}\log\frac{1}{\delta}+\frac{T}{\varepsilon}-\frac{2T}{\varepsilon^2}\left(\log(2\log\delta^{-1})+1-\log\varepsilon\right)\\+O\left(\frac{\log^2(\log\delta^{-1})}{\log\delta^{-1}}\right).\end{aligned}$$

The proof of this theorem is based on a relaxation of Theorem 4 obtained by ignoring $f(\alpha,\varepsilon,\delta)$. Considering both f and g will result in a stronger result at the expense of more involved analysis. Comparing with (28), Theorem 5 indicates that, providing δ is sufficiently small, the variance of each constituent Gaussian mechanism can be reduced by $\frac{2T}{\varepsilon^2}\left(\log(2\log\delta^{-1})+1-\log\varepsilon\right)$ compared to what would be obtained from MA.

B. Illustration of Our Bounds

In this section, we empirically compare our bound on $\varepsilon^\delta(\rho,T)$ given in Lemma 2 with the privacy parameter (23) obtained via MA and has been extensively used in the state-of-the-art differentially private machine learning algorithms, e.g., [7, 9, 18–23]. We do so in two different settings: (1) vanilla T-fold composition of the Gaussian mechanism with fixed variance, and (2) noisy SGD algorithm.

Vanilla Gaussian Composition: Here, we wish to obtain bounds on the privacy parameter ε of $\mathcal{M}^{(T)}$ where $\mathcal{M}$ is a Gaussian mechanism with $\sigma=20$. In Fig. 2, we compare Lemma 2 with MA when $\delta=10^{-5}$. According to this plot, our result enables us to achieve a smaller privacy parameter by up to 0.75, i.e., $\max_{T\in[1000]}\varepsilon^\delta_{\mathsf{MA}}(\rho,T)-\varepsilon^\delta(\rho,T)=0.75$ where $\varepsilon^\delta_{\mathsf{MA}}(\rho,t)$ is the ε given in (23). This privacy amplification may have important impacts on recent private deep leaning algorithms. Alternatively, one can observe that our result allows for more iteration for the same ε, for instance 100 more iterations for any ε larger than 6.

Noisy SGD: SGD is the standard algorithm for training many machine learning models. In order to fit a model without compromising privacy, a standard practice is to add Gaussian noise to the gradient of each mini-batch, see e.g., [3–6, 18, 21, 30]. The prime use of MA was to exploit the RDP's simple composition property in deriving the privacy parameters of the noisy SGD algorithm [3, Algorithm 1]. To have a fair comparison, we implement this algorithm with the sub-sampling rate $q=0.001$ and noise parameter $\sigma=4$ and then compute its DP parameter via (23) with $\rho=q^2/((1-q)\sigma^2)$ (see [3, Lemma 3]) and $\delta=10^{-5}$. We then compare it in Fig. 3 with Lemma 2 with the same ρ and σ. As demonstrated in this figure, our result allows remarkably more epochs (often over a hundred) within the same privacy budget and thus providing higher utility.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Conference on Theory of Cryptography*, ser. TCC'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 265–284.

[2] M. A. Pinsker, *Information and information stability of random variables and processes*. Holden-Day, 1964.

[3] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. of CCS*, 2016, pp. 308–318.

[4] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. of CCS*, 2015, pp. 1310–1321.

[5] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[6] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proceedings of FOCS*. Washington, DC, USA: IEEE Computer Society, 2014, pp. 464–473. [Online]. Available: https://doi.org/10.1109/FOCS.2014.56

[7] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Proc. 32Nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, 2018, pp. 6280–6290.

[8] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton, "Bolt-on differential privacy for scalable stochastic gradient descent-based analytics," in *SIGMOD*, 2017, pp. 1307–1322.

[9] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, and P. Kairouz, "A general approach to adding differential privacy to iterative training procedures," 2018. [Online]. Available: http://arxiv.org/abs/1812.06210

[10] Google, "Tensorflow privacy," 2018. [Online]. Available: https://github.com/tensorflow/privacy

[11] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, ser. FOCS '10, 2010, pp. 51–60.

[12] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 4037–4049, June 2017.

[13] I. Mironov, "Rényi differential privacy," in *Proceedings of 30th IEEE Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.

[14] P. Harremoës and I. Vajda, "On pairs of f-divergences and their joint range," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3230–3235, June 2011.

[15] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, "A better bound gives a hundred rounds:enhanced privacy guarantees via f-divergence." [Online]. Available: https://scholar.harvard.edu/files/shahab/files/isit2020_dp.pdf

[16] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato, "Hypothesis testing interpretations and Rényi differential privacy," *ArXiv*, vol. abs/1905.09982, 2019.

[17] T. van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[18] B. Balle, G. Barthe, M. Gaboardi, and J. Geumlek, "Privacy amplification by mixing and diffusion mechanisms," *ArXiv*, vol. abs/1905.12264, 2019.

[19] N. Papernot, M. Abadi, lfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *Proceedings of the International Conference on Learning Representations*, 2017.

[20] J. Geumlek, S. Song, and K. Chaudhuri, "Rényi differential privacy mechanisms for posterior sampling," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5289–5298.

[21] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532, 2018.

[22] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled Rényi differential privacy and analytical moments accountant," in *AISTATS*, 2018.

[23] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018.

[24] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[25] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of Royal Statistics*, vol. 28, pp. 131–142, 1966.

[26] N. Sharma and N. A. Warsi, "Fundamental bound on the reliability of quantum information transmission," *CoRR*, vol. abs/1302.5281, 2013. [Online]. Available: http://arxiv.org/abs/1302.5281

[27] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[28] G. Barthe and F. Olmedo, "Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs," in *ICALP*, 2013, pp. 49–60.

[29] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 10–15 July 2018, pp. 394–403.

[30] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization, revisited," in *ICML 2014 Workshop on Learning, Security and Privacy*, Beijing, China, 25 Jun 2014. [Online]. Available: http://arxiv.org/abs/1405.7085