Three Variants of Differential Privacy: Lossless Conversion and Applications

Shahab Asoodeh, Jiachun Liao, Flavio P. Calmon, *Member, IEEE*, Oliver Kosut, *Member, IEEE*, and Lalitha Sankar, *Senior Member, IEEE*

***Abstract*—We consider three different variants of differential privacy (DP), namely approximate DP, Rényi DP (RDP), and hypothesis test DP. In the first part, we develop a machinery for optimally relating approximate DP to RDP based on the joint range of two f-divergences that underlie the approximate DP and RDP. In particular, this enables us to derive the optimal approximate DP parameters of a mechanism that satisfies a given level of RDP. As an application, we apply our result to the moments accountant framework for characterizing privacy guarantees of noisy stochastic gradient descent (SGD). 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. In the second part, we establish a relationship between RDP and hypothesis test DP which allows us to translate the RDP constraint into a tradeoff between type I and type II error probabilities of a certain binary hypothesis test. We then demonstrate that for noisy SGD our result leads to tighter privacy guarantees compared to the recently proposed f-DP framework for some range of parameters.**

***Index Terms*—Differential privacy, Rényi divergence, binary hypothesis testing, f-divergences, moments accountant, stochastic gradient descent.**

I. Introduction

DIFFERENTIAL privacy (DP) [2] has become the *de facto* standard for privacy-preserving data analytics. Intuitively, a 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 difference of the *information density* [3] of the algorithm's output and two different inputs—referred to as the *privacy loss* random variable in the DP literature. In fact, several variants of DP has been proposed based on different properties of privacy loss random variable. Informally speaking, a mechanism is said to satisfy (ε, δ)-DP [2] if the privacy loss random variable is bounded by ε with probability $1-\delta$. A mechanism is said to be (α, γ)-*Rényi differential privacy* (RDP) [4] if the αth moment of the privacy loss random variable is upper bounded by γ; see Section II for more details.

Several methods have recently been proposed to ensure differentially private training of machine learning (ML) models [5]–[10]. 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 [5], [6]. This method is currently being used for privacy-preserving training of large-scale ML models in industry, see, e.g., the implementation of [11] in the Google's open-source TensorFlow Privacy framework [12].

Not surprisingly, for a fixed training dataset, privacy deteriorates with each SGD iteration. In practice, the DP constraints (i.e., ε and δ) 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 noisy 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 [13], [14]. Such composition results, while theoretically significant, may be loose due to their generality (e.g., they do not take into account the noise distribution used by the privacy mechanism).

Recently, Abadi *et al.* [5] 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 RDP for which composition has a simple linear form [4]. Once the privacy guarantees of the SGD execution are determined in terms of RDP, they are mapped back to DP guarantees in terms of ε and δ via a relationship between DP and RDP [5, Th. 2] allowing for converting from one to another. This approach renders tighter DP guarantees than those obtained from advanced composition theorems (see [5, Fig. 2]). Nevertheless, the existing conversion rules between RDP and DP are loose. In this article, we provide a framework which settles the *optimal*

Manuscript received August 14, 2020; revised November 25, 2020; accepted January 14, 2021. Date of publication January 26, 2021; date of current version March 16, 2021. This work was supported in part by NSF under Grant CIF-1922971, Grant CIF-1815361, Grant CIF-1742836, Grant CIF-1900750, and Grant CIF CAREER 1845852. Part of the results in this article was presented at the International Symposium on Information Theory 2020 [1]. *(Corresponding author: Shahab Asoodeh.)*

Shahab Asoodeh and Flavio P. Calmon are with the School of Engineering and Applied Science, Harvard University, Cambridge, MA 02138 USA (e-mail: shahab@seas.harvard.edu; flavio@seas.harvard.edu).

Jiachun Liao, Oliver Kosut, and Lalitha Sankar are with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: jiachun.liao@asu.edu; okosut@asu.edu; lalithasankar@asu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSAIT.2021.3054692, provided by the authors.

Digital Object Identifier 10.1109/JSAIT.2021.3054692

conversion between RDP and 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 [15, Th. 8] to characterize the joint range of these two f-divergences which, in turn, leads to the "optimal" conversion between RDP and DP (see Section III). Specifically, this optimal conversion allows us to derive bounds on the number of noisy SGD iterations for a given DP parameters ε and δ. Our result improves upon the state-of-the-art [5] 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 second part of this article, we revisit another variant of DP based on binary hypothesis testing. Consider an attacker who, given a mechanism's output, aims to determine if a certain individual (say Alice) has participated in the input dataset. This goal can be thought of as a hypothesis testing problem: rejecting the null hypothesis corresponds to the absence of Alice in the input dataset. It is well-known that (ε, δ)-DP is equivalent to enforcing that the type II error probability of *any* (possibly randomized) such test at significance level (or type I error probability) τ is lower bounded by $1 - \delta - e^\varepsilon \tau$ [14], [16]. Thus, for small ε and δ, any test is essentially powerless, i.e., it is impossible to have both small type I and type II error probabilities. This view of privacy (which we henceforth call *hypothesis test DP*) brings an operational interpretation for DP. This notion of privacy has recently been parameterized by a convex and decreasing function $f : [0, 1] \to [0, 1]$ that specifies the tradeoff between type I and type II error probabilities. A mechanism is said to be f-DP [17] if, given a mechanism's output, the type II error probability of any test for a given significance level τ is lower bounded by $f(\tau)$. Thus, if $f(\tau)$ is approximately $1 - \tau$, then any tests will be essentially powerless. This new definition is shown to provide easily interpretable privacy guarantees. This is in sharp contrast with RDP whose privacy guarantee does not enjoy a clear interpretation (see [18] for more details).

Our goal is to address the interpretability issue of RDP by relating RDP to f-DP. We first prove an explicit expression for the RDP guarantee of mechanism in terms of the type I and type II probabilities corresponding to the "optimal" test (given by Neyman-Pearson lemma). We remark that our expression is similar to an unproved formula that appeared first in [19, eq. (2.79)]. Conversely, we develop a machinery to implicitly relate RDP constraint to f-DP by constructing an achievable region of type I and type II error probabilities among all tests. This relationship is in particular interesting for the privacy analysis of iterative ML algorithm in that it converts the simple linear composition property of RDP to an interpretable privacy guarantee in terms of f-DP. Another approach for deriving an interpretable and tight privacy guarantee for ML algorithms is to resort to the general composition result of f-DP [17, Th. 3.2]. This approach is advocated in [20] for the privacy analysis of noisy SGD in training neural networks. We compare our results with [17], [20] in two different directions:

- The f-DP guarantee can be easily related to (ε, δ)-DP (see [17, Proposition 2.12]). It is argued in [20, Ths. 1 and 2] that f-DP guarantee of SGD *always* yields a strictly stronger (ε, δ)-DP guarantee than what would be obtained by moments accountant. We empirically show that this does not hold if one incorporates our optimal RDP-to-DP conversion rule into the moments accountant framework; i.e., the *improved* moments accountant might outperform f-DP, see Fig. 5.
- Rather then using the general composition results of f-DP, we propose to apply the linear composability of RDP and then convert the resulting guarantee to f-DP. Focusing on SGD with Gaussian noise, we demonstrate that there exists a threshold for variance below which our approach strictly outperforms f-DP, see Fig. 8 and Fig. 9.

A. Related Work

Since the introduction of the approximate DP in [2], it has been extensively studied especially for iterative ML algorithms, see [7], [10], [21]–[32] to name a few. Perhaps one of the most fundamental primitive in statistical privacy is the study of composition; how privacy degrades under as the algorithm iterates. There are still continued efforts to better understand the composition of DP. The advanced composition result for DP was derived [13]. In a pioneering work, [14] obtained an optimal homogeneous composition theorem for (ε, δ)-DP. It is, however, shown to be #P hard to compute the DP parameters under heterogeneous composition [33]. A substantial recent effort has been devoted to relaxing the DP constraints using divergences between probability distributions to address the weakness of (ε, δ)-DP in handling composition [4], [5], [34]–[37]. For instance, [4], [5], [34], [35] considered Rényi divergence and showed that the optimal privacy parameters under composition have simple linear forms. Once composition is handled, the resulting privacy parameters are converted to (ε, δ)-DP via some conversion rule, e.g., [5, Th. 2], [35, Proposition 1.3], and [4, Proposition 3]. This technique significantly improves on earlier privacy analysis of SGD. This technique has been extended by follow-up work [11]. More recently, a new relaxed version of DP (not divergence-based), termed f-DP was proposed in [17] and shown to enjoy a rather simple composition property. This new definition of DP was used in [20] for the privacy analysis in training deep neural networks.

B. Paper Organization

In Section II, we provide several preliminary definitions and results and also mathematically formulate our main goals. In Section III, we characterize the optimal relationship between RDP and DP and apply it to the moments accountant framework in Section IV. The content of Sections III and IV appeared in the conference version [1] without proofs. Section V concerns the second main goal of this article, that is, deriving a relationship between RDP and hypothesis test DP.

C. Notation

We denote by $\mathbb{D}$ the universe of all possible datasets and by $(\mathbb{X}, \mathcal{F})$ a measurable space with Borel σ-algebra $\mathcal{F}$. We also use $\mathcal{P}(\mathbb{X})$ to denote the set of all probability measures on $\mathbb{X}$.

We use capital letters, e.g., X to denote random variables. We write $X \sim P$ to describe the fact that X is distributed according to P. We also use the notation $\sim$ to indicate the neighborhood relationship between datasets, i.e., given two datasets d and d', we write $d \sim d'$ if their Hamming distance is equal to one. For a pair of distributions P and Q and constant $\alpha \geq 1$, we let

$$D_\alpha(P\|Q) \coloneqq \frac{1}{\alpha-1}\log \mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\alpha\right] \tag{1}$$

denote the Rényi divergence of order α. Also, given a real-valued convex function f satisfying $f(1) = 0$, the f-divergence [38], [39] between P and Q is defined as

$$D_f(P\|Q) \coloneqq \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right]. \tag{2}$$

For any real number a, we write $(a)_+$ for $\max\{a, 0\}$ and for $a \in [0, 1]$, we write $\bar{a}$ for $1 - a$.

II. Preliminaries and Problem Setup

In this section, we revisit several definitions and basic results that will be key for the discussion in the subsequent sections. A mechanism $\mathcal{M} : \mathbb{D} \to \mathcal{P}(\mathbb{X})$ assigns a probability distribution $\mathcal{M}_d$ to each dataset $d \in \mathbb{D}$. Given a pair of neighboring datasets $d \sim d'$, the privacy loss random variable is defined as $L_{d,d'} \coloneqq \log \frac{\mathrm{d}\mathcal{M}_d}{\mathrm{d}\mathcal{M}_{d'}}(X)$ where $X \sim \mathcal{M}_d$ and $\frac{\mathrm{d}\mathcal{M}_d}{\mathrm{d}\mathcal{M}_{d'}}$ represents the Radon-Nikodym derivative. Given an output of the mechanism $\mathcal{M}$ and a pair $d \sim d'$, consider the following problem of testing $\mathcal{M}_d$ against $\mathcal{M}_{d'}$:

$$H_0 : X \sim \mathcal{M}_d \quad \text{vs.} \quad H_1 : X \sim \mathcal{M}_{d'}. \tag{3}$$

Let $\beta_{\mathcal{M}}^{dd'} : [0, 1] \to [0, 1]$ denote the *optimal* tradeoff between type I error (i.e., the probability of declaring H_1 when the truth is H_0) and type II error (i.e., the probability of declaring H_0 when the truth is H_1). More specifically, $\beta_{\mathcal{M}}^{dd'}(\tau)$ is the smallest type II error when type I error equals τ. The mapping $\tau \mapsto \beta_{\mathcal{M}}^{dd'}(\tau)$ is sometimes called the *tradeoff function*.

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

- (ε, δ)-DP [2] for $\varepsilon \geq 0$ and $\delta \in [0, 1)$ if

$$\sup_{A\in\mathcal{F}, d\sim d'} \mathcal{M}_d(A) - e^\varepsilon \mathcal{M}_{d'}(A) \leq \delta. \tag{4}$$

- (α, γ)-RDP [4] for $\alpha > 1$ and $\gamma \geq 0$ if

$$\sup_{d\sim d'} D_\alpha(\mathcal{M}_d\|\mathcal{M}_{d'}) \leq \gamma. \tag{5}$$

- f-DP [17] for a convex and non-increasing function[1] $f : [0, 1] \to [0, 1]$ if for all $\tau \in [0, 1]$

$$\inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau) \geq f(\tau). \tag{6}$$

It can be shown that (4) is implied if the tail event $\{L_{d,d'} > \varepsilon\}$ occurs with probability at most δ for all $d \sim d'$, and (5) is implied if (and only if) the αth moment of $L_{d,d'}$ is upper bounded by γ. It is worth noting that the definition of RDP is closely related to *zero-concentrated* DP [34], [35]. Different properties of these two variants of DP have been extensively studied. One well-studied property of these two definitions is the composition (to be discussed in details in Section IV). As mentioned earlier, RDP tightly handles composition as opposed to the existing composition theorems for (ε, δ)-DP [13], [14] known to be either loose for many practical mechanisms or intractable to compute [33]. With this clear advantage comes a shortcoming: RDP suffers from the lack of operational interpretation, see [18]. To address this issue, the RDP guarantee is often translated into a DP guarantee via the following result.

Theorem 1 [5, Th. 2]: If the mechanism $\mathcal{M}$ is (α, γ)-RDP, then it satisfies (ε, δ)-DP for any $\varepsilon > \gamma$ and

$$\delta = e^{-(\alpha-1)(\varepsilon-\gamma)}. \tag{7}$$

This theorem establishes a relationship between RDP and DP that is extensively used in several recent differentially private ML applications, e.g., [9], [37], [40]–[44] to name a few. A prime use case for this relationship is the *moments accountant* (MA) [5] which is the current state-of-the-art privacy analysis technique for ML algorithms. However, despite its extensive use, Theorem 1 is loose in general and does not hold for all range of $\varepsilon \geq 0$. 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. Given its widespread applications, it seems very natural to aim at determining the *optimal* relationship between (ε, δ)-DP and (α, γ)-RDP. More precisely, we seek to answer the following question.

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 settle this question in Section III by expressing the optimal relationship via a simple one-variable convex optimization program. Incorporating this relationship into MA, we introduce the *improved* MA and quantify the resulting improvement in terms of privacy and utility in two different settings: T-fold homogeneous composition of Gaussian mechanism and noisy SGD algorithm.

As we shall see later, RDP is remarkably efficient in handling composition, making it an appealing notion of privacy for iterative algorithms such as SGD. Nevertheless, it lacks interpretability, see [18] for more details. Following the success of hypothesis test (3) in providing interpretation of approximate DP, we seek to relate RDP constraint to the tradeoff function $\inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau)$. Such relationship enables us to provide interpretable privacy guarantees for several iterative machine learning algorithms. It can be verified that $1 - \tau - \inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau)$ quantifies the fundamental indistinguishability of neighboring datasets based on the mechanism's output. Therefore, one effective way to describe the above relationship is to construct an outer bound for the region encompassed between the curves $\tau \mapsto 1 - \tau$ and $\tau \mapsto \inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau)$; the so-called *privacy region* of $\mathcal{M}$. Now we can describe the above relationship as follows.

Question Two: Given an (α, γ)-RDP mechanism, what is the characterization of its privacy region?

We provide an outer bound for the solution of this question in Section V and then demonstrate it for both T-fold homogeneous composition of Gaussian mechanism and noisy

[1] Both f-DP and f-divergence are defined in terms of convex functions. In order to be consistent with their original notation, we use f to denote the function in both definitions. It will be clear from the context and as a result will not lead to confusion.

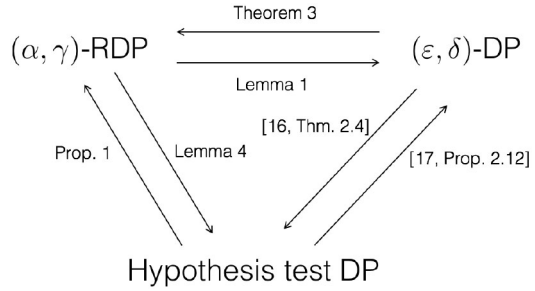


Fig. 1. The diagrammatic summary of the key relationships between three variants of DP studied in this article.

SGD algorithm. Interestingly, for the latter scenario, this outer bound is tighter than what would be obtained from applying results in [20].

We summarize our results on the relationship between (ε,δ)-DP, (α,γ)-RDP, and f-DP in Fig.1.

III. Optimal Relationship between RDP and DP

In this section, we aim at computing the fundamental worst-case DP privacy parameter guaranteed by an (α,γ)-RDP mechanism, thereby answering Question One. To this goal, we first express constraints in both (ε,δ)-DP and (α,γ)-RDP in terms of two f-divergences. Given $\lambda \geq 1$, the f-divergence associated with $f(t) = (t-\lambda)_+ = \max\{t-\lambda, 0\}$, is called E_λ-divergence [45] (aka *hockey-stick* divergence [46]) and given by

$$\mathsf{E}_\lambda(P\|Q) = \int (\mathrm{d}P - \lambda \mathrm{d}Q)_+ = \sup_{A\in\mathcal{F}} [P(A) - \lambda Q(A)]. \quad (8)$$

Also, for any $\alpha > 1$, the f-divergence associated with $f(t) = \frac{1}{\alpha-1}(t^\alpha - 1)$ is denoted by[2] $\chi^\alpha(P\|Q)$. Note that $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log(1+(\alpha-1)\chi^\alpha(P\|Q))$ for a pair of probability distributions P and Q.

It is shown in [48] that

$$\mathcal{M} \text{ is } (\varepsilon,\delta)\text{-DP} \iff \sup_{d\sim d'} \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'}) \leq \delta. \quad (9)$$

Similarly, it can be verified that:

$$\mathcal{M} \text{ is } (\alpha,\gamma)\text{-RDP} \iff \sup_{d\sim d'} \chi^\alpha(\mathcal{M}_d\|\mathcal{M}_{d'}) \leq \chi(\gamma), \quad (10)$$

where

$$\chi(\gamma) \coloneqq \frac{e^{(\alpha-1)\gamma}-1}{\alpha-1}. \quad (11)$$

Let the set of all (α,γ)-mechanisms be denoted by $\mathbb{M}_\alpha(\gamma)$, i.e.,

$$\mathbb{M}_\alpha(\gamma) \coloneqq \{\mathcal{M} : \mathbb{D} \to \mathcal{P}(\mathbb{X}) : \mathcal{M} \text{ is } (\alpha,\gamma)\text{-RDP}\}.$$

This definition, together with (9), enables us to precisely formulate Question One. In fact, Question One amounts to computing $\delta^\varepsilon_\alpha(\gamma)$

$$\delta^\varepsilon_\alpha(\gamma) \coloneqq \inf\{\delta\in(0,1) : \ \forall \mathcal{M}\in\mathbb{M}_\alpha(\gamma) \text{ is } (\varepsilon,\delta)\text{-DP}\} \quad (12)$$

$$= \sup_{\mathcal{M}\in\mathbb{M}_\alpha(\gamma)} \sup_{d\sim d'} \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'}), \quad (13)$$

[2] χ^α-divergence is also referred to as α-Hellinger divergence, see [47].

where the equality comes from (9) and (10). The map $\gamma \mapsto \delta^\varepsilon_\alpha(\gamma)$ in fact specifies the "optimal" conversion rule from RDP to DP for a given $\varepsilon \geq 0$. An equivalent way of describing such conversion is through the following quantity fixing $\delta \in (0,1)$

$$\varepsilon^\delta_\alpha(\gamma) \coloneqq \inf\{\varepsilon \geq 0 : \ \forall \mathcal{M}\in\mathbb{M}_\alpha(\gamma) \text{ is } (\varepsilon,\delta)\text{-DP}\}. \quad (14)$$

Similarly, the optimal conversion from DP to RDP is formulated by

$$\gamma^\varepsilon_\alpha(\delta) \coloneqq \sup\{\gamma \geq 0 : \forall \mathcal{M}\in\mathbb{M}_\alpha(\gamma) \text{ is } (\varepsilon,\delta)\text{-DP}\} \quad (15)$$

$$= \inf_{\mathcal{M}:\mathbb{D}\to\mathcal{P}(\mathbb{X})} \inf_{d\sim d'} \chi^{-1}\big(\chi^\alpha(\mathcal{M}_d\|\mathcal{M}_{d'})\big) \quad (16)$$

$$\text{s.t. } \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'}) \geq \delta \ \forall d\sim d', \quad (17)$$

where the $\chi^{-1}(\cdot)$ denotes the functional inverse of $\chi(\cdot)$ in (11), i.e., and is given by $\chi^{-1}(t) = \frac{1}{\alpha-1}\log(1+(\alpha-1)t)$. We seek to compute $\delta^\varepsilon_\alpha(\gamma)$ (or equivalently, $\varepsilon^\delta_\alpha(\gamma)$); however, it turns out that $\gamma^\varepsilon_\alpha(\delta)$ is simpler to compute. As a result, in the following we focus on the latter first.

Notice that, according to (10), 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 (11). Hence, $\gamma \mapsto \delta^\varepsilon_\alpha(\gamma)$ in fact constitutes the upper boundary of the convex set

$$\mathcal{R}_\alpha \coloneqq \Big\{\big(\chi^\alpha(\mathcal{M}_d\|\mathcal{M}_{d'}), \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'})\big)\Big| \forall \mathcal{M}, d\sim d'\Big\}. \quad (18)$$

This simple observation has some key implications. First, $\delta^\varepsilon_\alpha(\cdot)$ is non-decreasing and concave. Second, the upper boundary can be equivalently given by the map $\delta \mapsto \gamma^\varepsilon_\alpha(\delta)$. Furthermore, to compute $\gamma^\varepsilon_\alpha(\cdot)$ or $\delta^\varepsilon_\alpha(\cdot)$, it suffices to characterize $\mathcal{R}_\alpha$. This allows us to cast the problem of computing $\gamma^\varepsilon_\alpha(\cdot)$ as characterizing the joint range of E_λ and χ^α divergences. To tackle the latter problem, we refer to [15] whose main result is as follows.

Theorem 2 [15, Th. 8]: We have

$$\Big\{\big(D_f(P\|Q), D_g(P\|Q)\big)\Big| P, Q\in\mathcal{P}(\mathbb{X})\Big\} = \mathsf{conv}(\mathcal{B}) \quad (19)$$

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

$$\mathcal{B} \coloneqq \Big\{\big(D_f(P_\mathsf{b}\|Q_\mathsf{b}), D_g(P_\mathsf{b}\|Q_\mathsf{b})\big)\Big| P_\mathsf{b}, Q_\mathsf{b}\in\mathcal{P}(\{0,1\})\Big\}.$$

This theorem states that characterizing the joint range of any pair of f-divergences can be reduced without loss of generality to the binary case. For completeness, we give a more direct proof for the case of χ^α and E_λ divergences in Appendix A. We formalize this insight in Theorem 3 and establish a simple variational formula for $\gamma^\varepsilon_\alpha(\cdot)$ involving a one-parameter log-convex minimization program. Hence, the optimization (16), which can potentially be of significant complexity, turns into a simple tractable problem.

Theorem 3: For any $\alpha > 1$, $\varepsilon \geq 0$, and $\delta \in (0,1)$,

$$\gamma^\varepsilon_\alpha(\delta) = \varepsilon + \frac{1}{\alpha-1}\log M(\alpha,\varepsilon,\delta), \quad (20)$$

where $\bar{p} \coloneqq 1-p$ and

$$M(\alpha,\varepsilon,\delta) \coloneqq \min_{p\in(\delta,1)} \Big[p^\alpha(p-\delta)^{1-\alpha} + \bar{p}^\alpha\big(e^\varepsilon - p + \delta\big)^{1-\alpha}\Big].$$

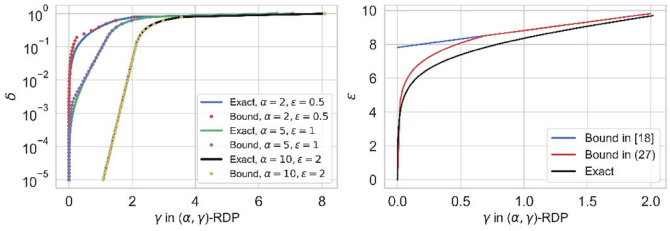


Fig. 2. Left: The exact values of the map $\delta \mapsto \gamma_\alpha^\varepsilon(\delta)$ obtained via numerically solving convex optimization problem (20). The dotted curves indicate the lower bound on $\gamma_\alpha^\varepsilon(\delta)$ according to Theorem 4 for three pairs of (α, ε). Right: Comparison of the exact values of the map $\varepsilon \mapsto \gamma_\alpha^\varepsilon(\delta)$ with the bounds obtained from (25) and [18, Th. 21] (i.e., considering only the first term in the minimization in (25)) with $\alpha = 2$ and $\delta = 10^{-4}$.

The proof of this theorem relies on Theorem 2 and is given in Appendix B. It can be shown that 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 lower bound in the following theorem that closely approximates $\gamma_\alpha^\varepsilon(\delta)$. We provide its proof in Appendix C.

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

$$\gamma_\alpha^\varepsilon(0) = 0, \tag{21}$$

$$\gamma_\alpha^\varepsilon(\delta) = \varepsilon - \log(1-\delta), \quad \text{if} \quad \alpha\delta \geq 1, \tag{22}$$

$$\gamma_\alpha^\varepsilon(\delta) \geq \max\{g(\alpha,\varepsilon,\delta), f(\alpha,\varepsilon,\delta)\}, \quad \text{if} \quad 0 < \alpha\delta < 1, \tag{23}$$

where

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

with $\zeta_\alpha \coloneqq \frac{1}{\alpha}(1-\frac{1}{\alpha})^{\alpha-1}$ and

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

In Fig. 2 (left panel), we numerically solve (20) 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 this figure, the lower bound on $\gamma_\alpha^\varepsilon(\cdot)$ in Theorem 4 is translated into an upper bound on $\delta_\alpha^\varepsilon(\cdot)$. 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 $\varepsilon_\alpha^\delta(\cdot)$.

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

$$\varepsilon_\alpha^\delta(\gamma) = (\gamma + \log(1-\delta))_+, \quad \text{if} \quad \alpha\delta \geq 1, \tag{24}$$

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

$$\varepsilon_\alpha^\delta(\gamma) \leq \frac{1}{\alpha-1}\min\left\{\left((\alpha-1)\gamma - \log\frac{\delta}{\zeta_\alpha}\right)_+, \times \log\left(\frac{e^{(\alpha-1)\gamma}-1}{\alpha\delta}+1\right)\right\}, \tag{25}$$

where ζ_α was defined in Theorem 4. Moreover, $\varepsilon_\alpha^\delta(0) = 0$. This lemma is obtained by solving equality (22) and inequality (23) for ε. Unlike $g(\alpha,\varepsilon,\delta)$, the map $\varepsilon \mapsto f(\alpha,\varepsilon,\delta)$ seems complicated to invert. To get around this difficulty, we use the first-order approximation of $f(\alpha,\varepsilon,\delta)$ around $\delta = 0$ to invert the inequality (23). The details are relegated to Appendix D. It is worth mentioning that Balle *et al.* [18, Th. 21] has recently proved $\varepsilon_\alpha^\delta(\gamma) \leq \gamma - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}$ via a fundamentally different approach. Their bound corresponds to the first term in (25), and thus weaker than Lemma 1. To emphasize on the advantage of (25) over [18, Th. 21], we plot these two bounds in Fig. 2 (right panel) for $\alpha = 2$ and $\delta = 10^{-4}$. As observed in this figure, considering only the first term in (25) would lead to non-trivial loss in ε especially when γ is sufficiently small. This observation is analytically justified by the fact that the first term in (25) does not tend to zero as $\gamma \to 0$ for reasonable values of α and δ whereas the second term does for any $\delta > 0$ and $\alpha > 1$.

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

IV. Improved Moments Accountant and Gaussian Mechanisms

Moments accountant (MA) was recently proposed by Abadi *et al.* [5] as a method to bypass advanced composition theorems [13], [14]. 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., $(\mathcal{M}^1, \ldots, \mathcal{M}^T)$ such that the input of $\mathcal{M}^i$ may depend on the outputs of $\mathcal{M}^1, \ldots, \mathcal{M}^{i-1}$. Determining the privacy parameters of $\mathcal{M}^{(T)}$ in terms of those of $\mathcal{M}$ is an important problem in practice and thus has been subject of an extensive body of research, see [5], [13], [14], [37].

Advanced composition theorems [13], [14] 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 (e.g., 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 $\mathcal{M}^1, \ldots, \mathcal{M}^T$ are each (α, γ)-RDP, then it is shown in [5, Th. 2] that $\mathcal{M}^{(T)}$ is $(\alpha, \gamma T)$-RDP. This result is then translated into DP privacy parameters via Theorem 1. In general, we assume this holds for *all* $\alpha > 1$ and hence one can obtain the *best* privacy parameters by optimizing over α. That is, $\mathcal{M}^{(T)}$ is (ε, δ)-DP for any $\varepsilon \geq 0$ and

$$\delta = \inf_{\alpha>1} e^{-(\alpha-1)(\varepsilon-\gamma(\alpha)T)}, \tag{26}$$

where $\gamma(\alpha) \coloneqq \max_{d\sim d'} D_\alpha(\mathcal{M}_d \| \mathcal{M}_{d'})$ is the RDP parameter of the constituent mechanism $\mathcal{M}$ and the dependence on α is made clear. Equivalently, $\mathcal{M}^{(T)}$ is (ε, δ)-DP for $\delta \in (0, 1)$ and

$$\varepsilon = \inf_{\alpha>1} \gamma(\alpha)T - \frac{1}{\alpha-1}\log\delta. \tag{27}$$

Since $\alpha \mapsto (\alpha-1)D_\alpha(P\|Q)$ is convex [49, Corollary 2] for any pair of probability measures P and Q, the above minimization

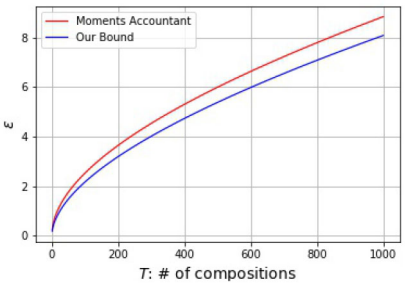


Fig. 3. The privacy parameter ε of the T-fold homogeneous composition of Gaussian mechanism each with $\sigma = 20$ according to MA (see (29)) and our bound in Lemma 2. We assume $\delta = 10^{-5}$.

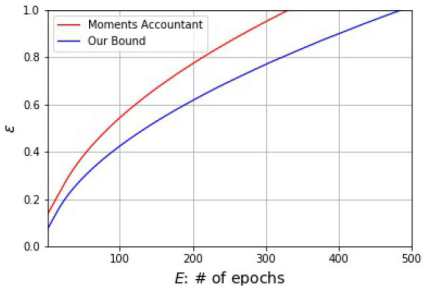


Fig. 4. Privacy parameter ε of noisy SGD algorithm according to MA (see (29)) and our bound (see Lemma 2) for $\delta = 10^{-5}$. The parameters of the algorithm are $\sigma = 4$ and the sub-sampling rate $q = 0.001$.

is a log-convex problem, and hence, can be solved within an arbitrary accuracy. Furthermore, we show in Section IV that this minimization has a simple form for Gaussian mechanisms and can be solved analytically. For the rest of this section, we assume $\mathcal{M}$ is a Gaussian mechanism and exploit Lemma 1 to derive tighter privacy parameters than (27).

A. Composition Results for Gaussian Mechanisms

Let $f : \mathbb{D} \to \mathbb{R}^n$ be a query function and $\mathcal{M}$ be a Gaussian mechanism with variance σ^2; more specifically, $\mathbb{X} = \mathbb{R}^n$ 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

$$\sup_{d\sim d'} D_\alpha(\mathcal{M}_d \| \mathcal{M}_{d'}) = \frac{\alpha}{2\sigma^2} \sup_{d\sim d'} \big\|f(d) - f(d')\big\|_2 = \frac{\alpha}{2\sigma^2}, \tag{28}$$

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. Hence, we deduce from (27) that $\mathcal{M}^{(T)}$ is (ε, δ)-DP for any $\delta \in (0, 1)$ and

$$\varepsilon = \inf_{\alpha>1} \gamma(\alpha)T - \frac{1}{\alpha-1}\log\delta = \rho T + \sqrt{4\rho T \log\frac{1}{\delta}}. \tag{29}$$

We next use the machinery developed in the previous section to obtain a tighter bound for the privacy parameter of $\mathcal{M}^{(T)}$ than (29). To do so, define

$$\varepsilon^\delta(\rho, T) \coloneqq \inf_{\alpha>1} \varepsilon_\alpha^\delta(\rho\alpha T). \tag{30}$$

Invoking Lemma 1, we can obtain an upper bound for $\varepsilon^\delta(\rho, T)$.

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

$$\varepsilon^\delta(\rho, T) \le \min\left\{\varepsilon_0(\rho, T),\ \varepsilon_1(\rho, T),\ \left(\frac{\rho T}{\delta} + \log(1-\delta)\right)_+\right\}, \tag{31}$$

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

$$\varepsilon_0(\rho, T) \coloneqq \inf_{\alpha\in\left(1,\frac{1}{\delta}\right]} \left(\rho\alpha T - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right)_+, \tag{32}$$

$$\varepsilon_1(\rho, T) \coloneqq \inf_{\alpha\in\left(1,\frac{1}{\delta}\right]} \frac{1}{\alpha-1}\log\left(1 + \frac{e^{\rho\alpha(\alpha-1)T} - 1}{\alpha\delta}\right), \tag{33}$$

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. To put our result in perspective, we first mention two previously-known bounds. Advanced composition theorems (see [13, Th. III.3]) require $\sigma^2 = \Omega(\frac{T\log(1/\delta)\log(T/\delta)}{\varepsilon^2})$. Abadi *et al.* [5, Th. 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. Plugging $\gamma(\alpha) = \frac{\alpha}{2\sigma^2}$ into (27) (or (26)), we can write

$$\frac{T}{2\sigma^2} \le \sup_{\alpha>1} \frac{\varepsilon}{\alpha} + \frac{1}{\alpha(\alpha-1)}\log\delta \tag{34}$$

$$= \varepsilon - 2\log\delta - 2\sqrt{(\varepsilon - \log\delta)\log\frac{1}{\delta}}, \tag{35}$$

and hence assuming δ is sufficiently small, we obtain

$$\sigma^2 \ge \frac{2T}{\varepsilon^2}\log\frac{1}{\delta} + \frac{T}{\varepsilon} + O\left(\frac{1}{\log\delta^{-1}}\right). \tag{36}$$

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

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

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 (36), Theorem 5 indicates that, providing δ is sufficiently small, the variance of each constituent Gaussian mechanism can be reduced by $\frac{2T}{\varepsilon^2}(\log(2\log\delta^{-1}) + 1 - \log\varepsilon)$ compared to what would be obtained from MA.

B. Illustration of Our Bounds

We now empirically compare Lemma 2 with the MA guarantee (29) that has been extensively used in the state-of-the-art

Algorithm 1 Noisy SGD

Input: Dataset $d = \{x_1, \ldots, x_n\}$, loss function $\ell(\theta, x)$, initial point θ_0, batch size m, noise variance σ^2, and clipping threshold C.
for $t = 1, \ldots, T$ **do**
  Select randomly a batch $\mathsf{I}_t \subset [n]$ of size m
  $g_t(x_i) = \nabla_\theta \ell(\theta_t, x_i)$, for $i \in \mathsf{I}_t$
  $\widetilde{g}_t(x_i) = g_t(x_i) \min\{1, \frac{C}{\|g_t(x_i)\|_2}\}$
  $\theta_{t+1} = \theta_t - \frac{\eta_t}{m}\big[\sum_{i\in\mathsf{I}_t} \widetilde{g}_t(x_i) + \sigma C Z\big], \quad Z \sim \mathcal{N}(0, \mathrm{I})$
end for
Output θ_T

differentially private algorithms, e.g., [9], [11], [37], [40]–[44]. We do so in two different settings: (1) vanilla T-fold composition of the Gaussian mechanisms with fixed variance, and (2) noisy SGD algorithm.

1) *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. 3, 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 (29). 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, e.g., 100 more iterations for any ε larger than 6.
2) *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., [5]–[8], [22], [40], [43]. The prime use of MA was to exploit the RDP's simple composition property in deriving the privacy parameters of the noisy SGD algorithm [5, Algorithm 1]. To have a fair comparison, we analyze this algorithm (see Algorithm 1) with the sub-sampling rate $q = 0.001$ and noise parameter $\sigma = 4$ and then compute its DP parameter via (29) with $\rho = \frac{q^2}{(1-q)\sigma^2}$ (see [5, Lemma 3]) and $\delta = 10^{-5}$. We then compare it in Fig. 4 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.

Since Lemma 1 is shown to improve on the composition results of MA, it is reasonable to construct the *improved* MA: First use the linear composability of RDP to take into account the composition and then use Lemma 1 to convert the resulting RDP guarantee to (ε, δ)-DP. We next show that improved MA might lead to tighter guarantee than hypothesis test privacy.

C. Comparison with f-DP

As mentioned earlier, f-DP (see Definition 6) leads to stronger DP guarantee than what is obtained by MA for noisy SGD algorithms. More precisely, Bu *et al.* [20, Th. 2] showed

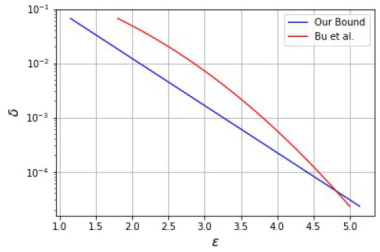


Fig. 5. Comparison of parameters of ε and δ in noisy SGD algorithm obtained from Lemma 2 and [20, Th. 2]. The parameters of the algorithm are as follows: $q = 0.003$, epoch $E = 30$ (hence $T = \frac{E}{q} = 10000$), and $\sigma = 0.6$.

that if one applies composition results of f-DP (i.e., [17, Th. 3.2]) to noisy SGD algorithms and then converts it to DP (via [17, Proposition 3.12]), then the resulting ε is asymptotically smaller than (29) for any $\delta \in (0, 1)$ provided that the sub-sampling rate q is scaled as $\frac{1}{\sqrt{T}}$ with T being the number of iteration. A natural question raised here is whether this result still holds if we replace MA with the improved MA.

In Fig. 5, we consider noisy SGD algorithm with Gaussian noise with $\sigma = 0.6$ and sub-sampling rate $q = 0.003$ (similar to [20, Fig. 2]) and compare Lemma 2 with [20, Th. 2]. As clearly illustrated by this figure, the improved MA may yield tighter privacy guarantees than what f-DP promises.

V. Hypothesis Testing Privacy

In this section, we investigate the relationship between RDP and hypothesis test privacy, that is, we focus on Question Two in the introduction. Let X be the output of a mechanism $\mathcal{M}$. For any pair of neighboring dataset $d \sim d'$, we consider the hypothesis test (repeated from the introduction for convenience)

$$H_0 : X \sim \mathcal{M}_d \quad \text{vs.} \quad H_1 : X \sim \mathcal{M}_{d'}. \tag{37}$$

The fundamental efficiency of a randomized test between H_0 and H_1 is delineated by a *decision rule*, a random transformation $P_{Z|X} : \mathbb{X} \to \mathcal{P}(\{0, 1\})$ where 1 indicates that H_0 is rejected. Type I and type II error probabilities corresponding to the decision rule $P_{Z|X}$ are given by $\int P_{Z|X}(1|x)\mathcal{M}_d(\mathrm{d}x)$ and $\int P_{Z|X}(0|x)\mathcal{M}_{d'}(\mathrm{d}x)$, respectively. To capture the optimal tradeoff between type I and type II error probabilities, it is customary to define *tradeoff function* $\beta^{dd'}_{\mathcal{M}} : [0, 1] \to [0, 1]$ given by

$$\beta^{dd'}_{\mathcal{M}}(\tau) \coloneqq \inf \int P_{Z|X}(0|x)\mathcal{M}_{d'}(\mathrm{d}x) \tag{38}$$

where the infimum is taken over all decision rules $P_{Z|X}$ such that $\int P_{Z|X}(1|x)\mathcal{M}_d(\mathrm{d}x) \le \tau$.

Note that we can always assume, without loss of generality, that $\tau + \beta^{dd'}_{\mathcal{M}}(\tau) \le 1$, since for any decision rule one can take its negation. The line $\tau + \beta^{dd'}_{\mathcal{M}}(\tau) = 1$ indicates the complete indistinguishability between d and d' on the basis of a mechanism's output. It follows from the definition that the map $\tau \mapsto \beta^{dd'}_{\mathcal{M}}(\tau)$ is non-increasing and convex. Recall that the mechanism $\mathcal{M}$ is said to be f-DP for a convex and non-increasing function f that is majorized by $\inf_{d\sim d'} \beta^{dd'}_{\mathcal{M}}$,

that is if $f(\tau) \leq \inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau)$ for any $\tau \in [0,1]$. Hence, the problem of determining the relationship between RDP and f-DP reduces to characterizing the set (τ, β) such that $\beta \geq \inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau)$ for all mechanisms $\mathcal{M}$ with a certain level of RDP guarantee. To this goal, we define the *privacy region* of mechanism $\mathcal{M}$ as

$$\mathcal{C}_{\mathcal{M}} \coloneqq \bigcup_{d\sim d'} \Big\{(\tau,\beta) \in [0,1]^2 : \beta_{\mathcal{M}}^{dd'}(\tau) \leq \beta \leq 1-\tau\Big\}.$$

It was shown by [14], [16] that a mechanism $\mathcal{M}$ is (ε, δ)-DP if and only if

$$\mathcal{C}_{\mathcal{M}} \subseteq \mathcal{C}(\varepsilon,\delta) \coloneqq \Big\{(\tau,\beta) \in [0,1]^2 : \tau + e^{\varepsilon}\beta \geq 1-\delta, \beta + e^{\varepsilon}\tau \geq \bar{\delta}\Big\}. \tag{39}$$

Remark 2: Recall from the definition of E_λ-divergence (8) that, for any pair of distributions (P, Q) and positive λ, we have $\mathsf{E}_\lambda(P\|Q) = P(\frac{\mathrm{d}P}{\mathrm{d}Q} \geq \lambda) - \lambda Q(\frac{\mathrm{d}P}{\mathrm{d}Q} \geq \lambda)$. Since according to Neyman-Pearson lemma $\beta_{\mathcal{M}}^{dd'}(\tau) = \mathcal{M}_{d'}(\log \frac{\mathrm{d}\mathcal{M}_d}{\mathrm{d}\mathcal{M}_{d'}} \geq \varepsilon)$ where $\tau = \mathcal{M}_d(\log \frac{\mathrm{d}\mathcal{M}_d}{\mathrm{d}\mathcal{M}_{d'}} \leq \varepsilon)$, it follows that the line $\beta = e^{-\varepsilon}(1-\tau-\delta_\varepsilon)$, with $\delta_\varepsilon \coloneqq \min_{d\sim d'} \mathsf{E}_{e^\varepsilon}(\mathcal{M}_d\|\mathcal{M}_{d'})$, supports $\mathcal{M}$ from below. Swapping d and d', we deduce that the line $\beta = 1-\delta_\varepsilon - e^{\varepsilon}\tau$ is another supporting line of $\mathcal{C}_{\mathcal{M}}$ with slope e^{ε}. Due to the convexity of $\mathcal{C}_{\mathcal{M}}$, the collection of all supporting lines losslessly constructs $\mathcal{C}_{\mathcal{M}}$; thus, $\bigcap_{\varepsilon\geq 0} \mathcal{C}(\varepsilon,\delta_\varepsilon) = \mathcal{C}_{\mathcal{M}}$. In other words, the collection of $\{(\varepsilon,\delta_\varepsilon)\}_{\varepsilon\geq 0}$ and the mapping $\tau \mapsto \inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau)$ capture the same privacy guarantee. This provides a new lens to explore, delineate and interpret privacy guarantee achieved by differential privacy. This new perspective has recently been adopted by Dong *et al.* [17]. To illustrate this observation, consider the Gaussian mechanism. It is easy to see that for Gaussian mechanisms (assuming unit L_2-sensitivity)

$$\delta_\varepsilon = \Phi\Big(-\varepsilon\sigma + \frac{1}{2\sigma}\Big) - e^{\varepsilon}\Phi\Big(-\varepsilon\sigma - \frac{1}{2\sigma}\Big), \tag{40}$$

where Φ is the standard normal CDF. On the other hand, for a Gaussian mechanism $\mathcal{M}$ with variance σ^2, the Neyman-Pearson lemma implies that the tradeoff function $\inf_{d\sim d'} \beta_{\mathcal{M}}^{dd'}(\tau) = G_{\frac{1}{\sigma}}(\tau)$, where

$$G_\mu(\tau) = \Phi\Big(\Phi^{-1}(1-\tau) - \mu\Big), \tag{41}$$

and Φ^{-1} is the inverse of Φ. It is worth mentioning that $G_\mu(\tau)$ in fact corresponds to the smallest type II error probability of testing $\mathcal{N}(0,1)$ against $\mathcal{N}(\mu,1)$ with type I error probability being τ. In Fig. 6, we identify the region $\mathcal{C}_{\mathcal{M}}$ by its lower boundary (red curve) given by the above tradeoff function and its upper boundary $\beta = 1-\tau$. The blue curve is the lower boundary of $\mathcal{C}(\varepsilon,\delta_\varepsilon)$ for $\varepsilon = 1$.

While the DP constraint can be operationally interpreted via (39), it is not clear how to obtain a similar interpretation for RDP constraint. Nevertheless, we wish to obtain some implications of a mechanism's RDP constraints on its privacy regions. We begin by giving an explicit formula for the RDP guarantee of a mechanism in terms of the derivative of the map $\tau \mapsto \beta_{\mathcal{M}}^{dd'}(\tau)$ for $d \sim d'$.

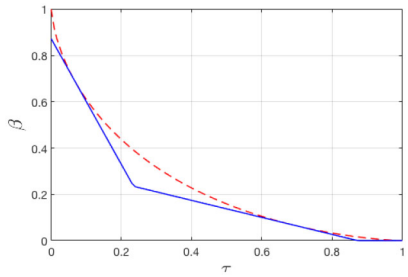


Fig. 6. Two outer bounds for the Gaussian mechanism with $\sigma^2 = 1$: The red curve is the map $\tau \mapsto \Phi(\Phi^{-1}(\bar{\tau}) - 1/\sigma)$ and the blue curve specifies the region $\mathcal{C}(\varepsilon,\delta_\varepsilon)$ for $\varepsilon = 1$ and δ_ε given in (40).

Proposition 1: Given $\alpha > 1$, a mechanism $\mathcal{M}$ is (α,γ)-RDP for

$$\gamma = \sup_{d\sim d'} \frac{1}{\alpha-1} \log\Bigg(1 - \beta_{\mathcal{M}}^{dd'}(0) + \int_0^1 |\Gamma_{dd'}(\tau)|^{1-\alpha}\mathrm{d}\tau\Bigg), \tag{42}$$

where $\Gamma_{dd'}(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\beta_{\mathcal{M}}^{dd'}(\tau)$.

The proof of this result relies on a general fact: all f-divergences between $\mathcal{M}_d$ and $\mathcal{M}_{d'}$ can be explicitly expressed in terms of the derivative of $\beta_{\mathcal{M}}^{dd'}$. This was mentioned, without a proof, in [19, eq. (2.79)] in a completely different context and was recently proved in [17, Proposition B.4]. We give a more direct proof in Appendix H.

Proposition 1 provides an explicit RDP guarantee for a mechanism with a given hypothesis test privacy constraint. The other direction seems more practical: Given an (α,γ)-RDP mechanism, what can we say about its privacy region $\mathcal{C}_{\mathcal{M}}$? There are two approaches to address this question. First, one can use the machinery developed in Section III to relate (α,γ)-RDP constraint to $(\varepsilon,\delta_\alpha^\varepsilon(\gamma))$-DP and then declare $\mathcal{C}(\varepsilon,\delta_\alpha^\varepsilon(\gamma))$ as an outer bound for the privacy region for any $\varepsilon \geq 0$. Alternatively, one can use information theoretic results (such as data processing inequality) to *directly* relate Rényi divergence to type I and type II error probabilities in hypothesis testing (37) (see [50]). In the following, we delineate these two approaches.

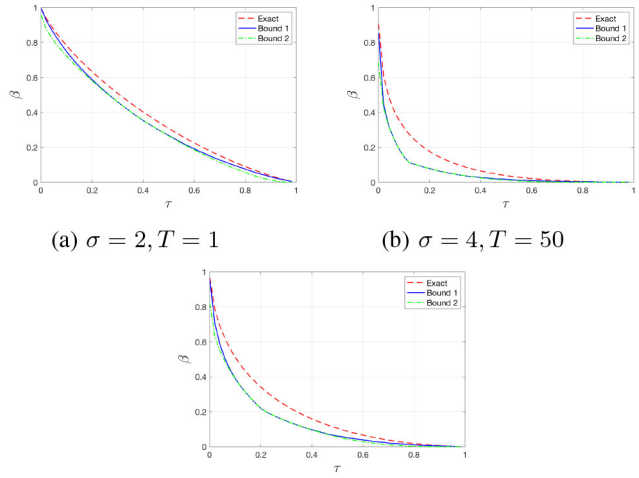
Since all (α,γ)-RDP mechanisms are $(\varepsilon,\delta_\alpha^\varepsilon(\gamma))$-DP, we immediately obtain the following result from (39).

Lemma 3: Let $\mathcal{M}$ be an (α,γ)-RDP mechanism. Then, we have

$$\mathcal{C}_{\mathcal{M}} \subseteq \bigcap_{\varepsilon\geq 0} \mathcal{C}\big(\varepsilon,\delta_\alpha^\varepsilon(\gamma)\big). \tag{43}$$

Note that since $\varepsilon \mapsto \delta_\alpha^\varepsilon(\gamma)$ characterizes the DP parameters of the worst mechanism in $\mathbb{M}_\alpha(\gamma)$, it follows that the privacy regions of *all* (α,γ)-RDP mechanisms are contained in $\bigcap_{\varepsilon\geq 0}\mathcal{C}(\varepsilon,\delta_\alpha^\varepsilon(\gamma))$, or equivalently, $\bigcup_{\mathcal{M}\in\mathbb{M}_\alpha(\gamma)}\mathcal{C}_{\mathcal{M}} \subseteq \bigcap_{\varepsilon\geq 0}\mathcal{C}(\varepsilon,\delta_\alpha^\varepsilon(\gamma))$.

Instead of dealing with the infinite collection of $(\varepsilon,\delta_\alpha^\varepsilon(\gamma))$ and taking the intersection of $\mathcal{C}(\varepsilon,\delta_\alpha^\varepsilon(\gamma))$, we can alternatively focus on the tradeoff function (see Remark 2). That is, we wish to study the privacy regions of RDP mechanisms by *directly* computing bounds on the tradeoff function rather than converting RDP into (ε,δ)-DP. Adopting this viewpoint, we establish



(a) $\sigma = 2, T = 1$ (b) $\sigma = 4, T = 50$

(c) $\sigma = 8, T = 100$

Fig. 7. The outer bounds for the privacy region of the T-fold homogeneous Gaussian mechanism. The regions marked as Bound 1 and Bound 2 correspond to (52) and (53), respectively and the region marked as Exact corresponds to (51). Recall that the privacy "regions" are to be interpreted as the region between depicted curves and the diagonal line $\tau + \beta = 1$.

two outer bounds for the privacy region of an (α, γ)-RDP mechanism in the following lemma.

Lemma 4: Let $\mathcal{M}$ be an (α, γ)-RDP mechanism. Then, the privacy region of $\mathcal{M}$ satisfies

$$\mathcal{C}_{\mathcal{M}} \subseteq \left\{(\tau, \beta) \in (0,1)^2 : d_\alpha(\bar{\tau}\|\beta) \le \gamma, d_\alpha\big(\bar{\beta}\|\tau\big) \le \gamma\right\} \quad (44)$$

$$\subseteq \left\{(\tau, \beta) \in (0,1)^2 : d(\bar{\tau}\|\beta) \le \gamma, d\big(\bar{\beta}\|\tau\big) \le \gamma\right\}, \quad (45)$$

where $d(a\|b) = a\log\frac{a}{b} + \bar{a}\log\frac{\bar{a}}{\bar{b}}$ and $d_\alpha(a\|b) := \frac{1}{\alpha-1}\log(a^\alpha b^{1-\alpha} + \bar{a}^\alpha \bar{b}^{1-\alpha})$ for $a, b \in (0,1)$.

Proof: Let $P_{Z|X}$ be an optimal randomized test mapping the mechanism's output X to a binary variable Z corresponding to H_0 and H_1, i.e., $\int P_{Z|X}(0|x)\mathcal{M}_d(\mathrm{d}x) = 1 - \tau$ and $\int P_{Z|X}(0|x)\mathcal{M}_{d'}(\mathrm{d}x) = \beta^{dd'}_{\mathcal{M}}(\tau)$. (The existence of such an optimal randomized test is guaranteed by Neyman-Pearson lemma.) Due to the data processing inequality, we have

$$D_\alpha(\mathcal{M}_d\|\mathcal{M}_{d'}) \ge D_\alpha\Big(\mathsf{Bernoulli}(\tau)\|\mathsf{Bernoulli}\Big(1 - \beta^{dd'}_{\mathcal{M}}(\tau)\Big)\Big)$$
$$= d_\alpha\Big(1 - \tau\|\beta^{dd'}_{\mathcal{M}}(\tau)\Big), \quad (46)$$

This in turn implies that for all $d \sim d'$

$$\max\Big\{d_\alpha\Big(1 - \tau\|\beta^{dd'}_{\mathcal{M}}(\tau)\Big), d_\alpha\Big(\beta^{dd'}_{\mathcal{M}}(\tau)\|1 - \tau\Big)\Big\} \le \gamma, \quad (47)$$

which in turn implies (44) by noticing that $a \mapsto d_\alpha(a\|b)$ is decreasing for $a < b$ and similarly $b \mapsto d_\alpha(a\|b)$ is decreasing for $b < a$. Since $\alpha \mapsto D_\alpha(P\|Q)$ is non-decreasing [49, Th. 3], the inclusion (45) follows immediately. ∎

It is worth mentioning that $d_\alpha(a\|b)$ is closely related to [18, Definition 9]. Note that although the set in (45) strictly contains the one in (44), it enables us to derive a simple outer bound for the privacy region of mechanisms when optimizing over α. This is formalized in the following result which is an immediate corollary of Lemma 4.

Corollary 1: If mechanism $\mathcal{M}$ is $(\alpha, \gamma(\alpha))$-RDP for all $\alpha > 1$. Then its privacy region satisfies

$$\mathcal{C}_{\mathcal{M}} \subseteq \bigcap_{\alpha>1}\left\{(\tau, \beta) : d_\alpha(\bar{\tau}\|\beta) \le \gamma(\alpha), d_\alpha\big(\bar{\beta}\|\tau\big) \le \gamma(\alpha)\right\} \quad (48)$$

$$\subset \bigcap_{\alpha>1}\left\{(\tau, \beta) : d(\bar{\tau}\|\beta) \le \gamma(\alpha), d\big(\bar{\beta}\|\tau\big) \le \gamma(\alpha)\right\}. \quad (49)$$

To demonstrate the accuracy of Corollary 1, we consider Gaussian mechanisms for the remainder of this section. Recall that the Gaussian mechanism with variance σ^2 is (α, γ)-RDP for $\gamma = \rho\alpha$ with $\rho = \frac{1}{2\sigma^2}$. Recall that the T-fold composition of such mechanism is $(\alpha, \rho\alpha T)$-RDP, implying that $\mathcal{M}^{(T)}$ is a Gaussian mechanism with variance $\frac{\sigma^2}{T}$. Hence, according to (41), we have

$$\inf_{d\sim d'} \beta^{dd'}_{\mathcal{M}^{(T)}}(\tau) = G_{\sqrt{2\rho T}}(\tau) \quad (50)$$

This, in turn, implies that $\mathcal{C}_{\mathcal{M}^{(T)}}$ the privacy region of $\mathcal{M}^{(T)}$ is given by

$$\mathcal{C}_{\mathcal{M}^{(T)}} = \left\{(\tau, \beta) \in (0,1)^2 : G_{\sqrt{2\rho T}}(\tau) \le \beta \le 1 - \tau\right\}. \quad (51)$$

Specializing Corollary 1 to $\mathcal{M}^{(T)}$, we can express outer bounds given in (48) and (49) as

$$\mathcal{C}_{\mathcal{M}^{(T)}} \subseteq \bigcap_{\alpha>1}\left\{(\tau, \beta) : d_\alpha(\bar{\tau}\|\beta) \le \rho\alpha T, d_\alpha\big(\bar{\beta}\|\tau\big) \le \rho\alpha T\right\} \quad (52)$$

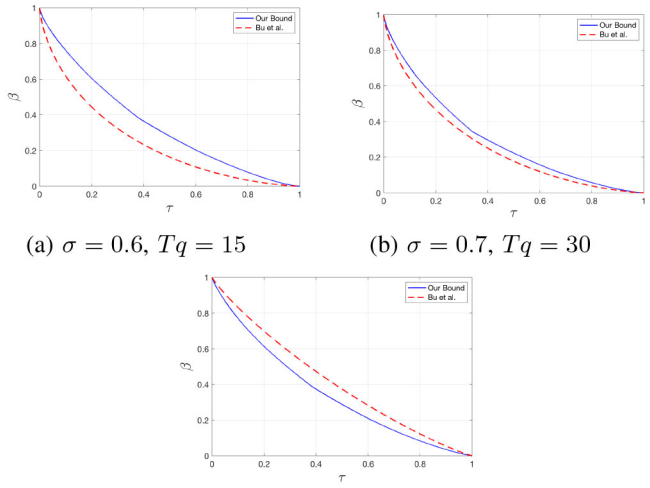
$$\subset \left\{(\tau, \beta) \in [0,1]^2 : d(\bar{\tau}\|\beta) \le \rho T, d\big(\bar{\beta}\|\tau\big) \le \rho T\right\}. \quad (53)$$

In Fig. 7, we compare these outer bounds with the exact privacy region given in (51). Note that the region (49), while being weaker than the region in (48), can be explicitly characterized for Gaussian mechanisms. For a more realistic application, we apply Corollary 1 to noisy SGD algorithm (i.e., Algorithm (1)). This algorithm can be thought of as a T-fold composition of Gaussian mechanism with an additional feature of subsampling (line 3 in Algorithm 1) with rate $q = \frac{m}{n}$. As before, we invoke [5, Lemma 3] to obtain that each iteration of this algorithm is approximately $(\alpha, \alpha\rho_q)$-RDP where $\rho_q = \frac{q^2}{(1-q)\sigma^2}$ for positive integer $\alpha \le 1 + \sigma^2\log\frac{1}{q\sigma}$ and $q < \frac{1}{16\sigma}$. Thus, after T iterations the algorithm is $(\alpha, \alpha\rho_q T)$-RDP. Corollary 1 therefore gives

$$\mathcal{C}_{\mathsf{SGD}}(T) \subseteq \bigcap_{\alpha\in\mathcal{A}}\left\{(\tau, \beta) : d_\alpha(\bar{\tau}\|\beta) \le \alpha\rho_q T, d_\alpha\big(\bar{\beta}\|\tau\big) \le \alpha\rho_q T\right\}, \quad (54)$$

where $\mathcal{A}$ is the set of admissible α indicated above. On the other hand, subsampling and composition results of f-DP ([17, Th. 4.2] and [17, Th. 3.2], respectively) can be exploited to approximate (asymptotically in T) the tradeoff function for the Algorithm 1 and thus to construct an outer bound for the privacy region [20]:

$$\mathcal{C}_{\mathsf{SGD}}(T) \subseteq \left\{(\tau, \beta) \in (0,1)^2 : G_\mu(\tau) \le \beta \le 1 - \tau\right\}, \quad (55)$$



(a) $\sigma = 0.6$, $Tq = 15$ (b) $\sigma = 0.7$, $Tq = 30$

(c) $\sigma = 1.3$, $Tq = 30$

Fig. 8. The outer bounds for the privacy region of SGD algorithm according to our RDP-based bound (54) (blue solid curve) and f-DP [20] (red dashed curve) with the subsampling rate $q = 256/60000$. As the blue curve lies above the red curve for $\sigma \leq 0.7$, our bound yields tighter privacy region. Since the intersection in (54) is over only integer α, the blue curve may not be smooth for large T.

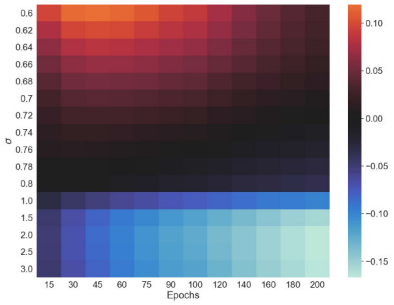


Fig. 9. The difference between the area of the region in the right-hand side of (55) and (54) with the subsampling rate $q = 256/60000$. A positive value indicates that the outer bound in (54) is smaller than that in (55), or equivalently, RDP leads to a tighter privacy guarantee that f-DP.

where $\mu = q\sqrt{T(e^{1/\sigma^2} - 1)}$ and $G_\mu(\cdot)$ was defined in (41). In Fig. 8, we illustrate this bound together with (54) for different number of iterations and σ. The numerical findings indicate that there always exists a σ_0 for any sub-sampling rate q such that our RDP-based outer bound (54) is tighter than f-DP bound (55) for all $\sigma \leq \sigma_0$ irrespective of the number of iterations. For instance, $\sigma_0 \approx 0.7$ in Fig. 8, that is, (54) is tighter than (55) for all $\sigma \leq 0.7$ and any number of iterations. To better support this claim, we compute the area of the regions on the right-hand sides of (54) and (55) and report the differences in Fig. 9 for different values of σ and T. Positive numbers indicate that the former is a smaller region, or equivalently, the outer bound in (54) is tighter than (55); thus supporting our claim.

Conclusion

In this article, we investigated the relationship between three variants of differential privacy, namely approximate DP, Rényi DP, and hypothesis test DP. First, we established the *optimal* relationship between Rényi DP and approximate DP that enables us to derive the optimal approximate DP parameters of a mechanism that satisfies a given level of Rényi DP. In order to show its practicality, we applied this result to the moments accountant framework for characterizing privacy guarantees of noisy stochastic gradient descent. When compared to the state-of-the-art, our result was shown to lead to about 100 more stochastic gradient descent iterations for training deep learning models for the same privacy budget, and thus provide better accuracy without any privacy degradation. In the second part, we analyzed the implications of Rényi DP constraint in terms of the tradeoff between type I and type II error probabilities of a certain binary hypothesis test which formalizes the hypothesis test DP. More specifically, we derived an outer bound for the region of type I and type II error probabilities (also known as the *privacy region*) achievable by a mechanism that satisfies a given level of Rényi DP. We then used this result to characterize the privacy region of noisy stochastic gradient descent algorithm. Compared to the existing results (obtained via sub-sampling and composition results of recently proposed f-DP framework), our outer bound was empirically shown to be tighter for a practical range of the noise variance.

Appendix A
Sufficiency of Binary Distributions for Characterizing $\mathcal{R}_\alpha$

We provide a direct proof for the fact that it suffices to consider the Bernoulli distributions for characterizing $\mathcal{R}_\alpha$. The following argument is a natural extension of the proof of [51, Lemma 2]. Let P and Q be two general distributions on $\mathbb{X}$. We wish to show that the for any $\lambda \geq 1$ and $\alpha > 1$ the optimization

$$\begin{aligned} \inf_{P,Q\in\mathcal{P}(\mathbb{X})} \quad & D_\alpha(P\|Q) \\ \text{s.t.} \quad & \mathsf{E}_\lambda(P\|Q) \geq \delta, \end{aligned} \tag{56}$$

is achieved by Bernoulli distributions. Let $\phi : \mathbb{X} \to \{1, 2\}$ be defined as

$$\phi(x) = \begin{cases} 1, & \text{if } \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \geq \lambda \\ 2, & \text{if } \frac{\mathrm{d}P}{\mathrm{d}Q}(x) < \lambda. \end{cases} \tag{57}$$

Also, define Bernoulli distributions P_b and Q_b on $\{1, 2\}$ as follows

$$P_\mathsf{b}(j) = \int_{x:\phi(x)=j} P(\mathrm{d}x), \tag{58}$$

and

$$Q_\mathsf{b}(j) = \int_{x:\phi(x)=j} Q(\mathrm{d}x), \tag{59}$$

for $j \in \{1, 2\}$. Note that in this case, we can write

$$\|P - \lambda Q\| = \int_{\mathbb{X}} |P(\mathrm{d}x) - \lambda Q(\mathrm{d}x)| \tag{60}$$

$$= \int_{\phi(x)=1} (P(\mathrm{d}x) - \lambda Q(\mathrm{d}x)) + \int_{\phi(x)=2} (\lambda Q(\mathrm{d}x) - P(\mathrm{d}x)) \tag{61}$$

$$= P_{\mathsf{b}}(1) - \lambda Q_{\mathsf{b}}(1) + \lambda Q_{\mathsf{b}}(2) - P_{\mathsf{b}}(2) \tag{62}$$
$$= |P_{\mathsf{b}}(1) - \lambda Q_{\mathsf{b}}(1)| + |P_{\mathsf{b}}(2) - \lambda Q_{\mathsf{b}}(2)| \tag{63}$$
$$= \|P_{\mathsf{b}} - \lambda Q_{\mathsf{b}}\|. \tag{64}$$

Notice that $\mathsf{E}_\lambda(P\|Q) = \frac{1}{2}\|P - \lambda Q\| + \frac{1}{2}(1-\lambda)$ and hence the above implies that $\mathsf{E}_\lambda(P\|Q) = \mathsf{E}_\lambda(P_{\mathsf{b}}\|Q_{\mathsf{b}})$. On the other hand, the data processing inequality for Rényi divergence implies that $D_\alpha(P\|Q) \geq D_\alpha(P_{\mathsf{b}}\|Q_{\mathsf{b}})$. These two observations demonstrate that the minimum of $D_\alpha(P\|Q)$ subject to $\mathsf{E}_\lambda(P\|Q) \geq \delta$ is achieved by Bernoulli distributions.

Appendix B
Proof of Theorem 3

First notice that, in light of Theorem 2, the convex set $\mathcal{R}_\alpha$ defined in (18) is equal to the convex hull of the set $\mathcal{B}_{\alpha,\varepsilon}$ given by

$$\mathcal{B}_{\alpha,\varepsilon} = \left\{\left(\chi^\alpha(P_{\mathsf{b}}\|Q_{\mathsf{b}}), \mathsf{E}_{e^\varepsilon}(P_{\mathsf{b}}\|Q_{\mathsf{b}})\right) \middle| P_{\mathsf{b}}, Q_{\mathsf{b}} \in \mathcal{P}(\{0,1\})\right\} \tag{65}$$

where $P_{\mathsf{b}} = \mathsf{Bernoulli}(p)$ and $Q_{\mathsf{b}} = \mathsf{Bernoulli}(q)$ with parameters $p, q \in (0,1)$. For any pair of such distributions, define $\widetilde{\gamma} \coloneqq \chi^\alpha(P_{\mathsf{b}}\|Q_{\mathsf{b}})$ and $\delta \coloneqq \mathsf{E}_{e^\varepsilon}(P_{\mathsf{b}}\|Q_{\mathsf{b}})$. We first show that the convex hull of $\mathcal{B}_{\alpha,\varepsilon}$ is given by

$$\bar{\mathcal{B}}_{\alpha,\varepsilon} = \left\{(\widetilde{\gamma}, \delta) \middle| \delta \in [0,1), \widetilde{\gamma} \geq \widetilde{\gamma}(\delta)\right\} \tag{66}$$

with $\widetilde{\gamma}(\delta)$ given by

$$\widetilde{\gamma}(\delta) = \inf_{0<p,q<1} \ \chi^\alpha(P_{\mathsf{b}}\|Q_{\mathsf{b}}) \quad \text{s.t.} \quad \mathsf{E}_{e^\varepsilon}(P_{\mathsf{b}}\|Q_{\mathsf{b}}) \geq \delta. \tag{67}$$

To this goal, we need to demonstrate that for any $\lambda \in [0,1]$ and pairs of points $(\widetilde{\gamma}_1, \delta_1), (\widetilde{\gamma}_2, \delta_2) \in \mathcal{B}_{\alpha,\varepsilon}$, we have $(\lambda\widetilde{\gamma}_1 + \bar{\lambda}\widetilde{\gamma}_2, \lambda\delta_1 + \bar{\lambda}\delta_2) \in \bar{\mathcal{B}}_{\alpha,\varepsilon}$, where $\bar{\lambda} = 1 - \lambda$, or equivalently $\lambda\delta_1 + \bar{\lambda}\delta_2 \in [0,1)$ and $\lambda\widetilde{\gamma}_1 + \bar{\lambda}\widetilde{\gamma}_2 \geq \widetilde{\gamma}(\lambda\delta_1 + \bar{\lambda}\delta_2)$. Hence, it suffices to show that $\delta \mapsto \widetilde{\gamma}(\delta)$ is convex.

Let $p_i, q_i \in (0,1)$ with $p_i \geq q_i$ be the optimal solution of (67) for δ_i, $i = 1, 2$, and $P_{\mathsf{b},i}, Q_{\mathsf{b},i}$ be the corresponding Bernoulli distributions. For any $\lambda \in [0,1]$, we construct two Bernoulli distribution $P_{\mathsf{b},\lambda}$ and $Q_{\mathsf{b},\lambda}$ with parameters $p_\lambda = \lambda p_1 + \bar{\lambda} p_2$ and $q_\lambda = \lambda q_1 + \bar{\lambda} q_2$, respectively. It can be verified that

$$\mathsf{E}_{e^\varepsilon}\left(P_{\mathsf{b},\lambda}\|Q_{\mathsf{b},\lambda}\right) = p_\lambda - e^\varepsilon q_\lambda \tag{68}$$
$$= \lambda p_1 + \bar{\lambda} p_2 - e^\varepsilon\left(\lambda q_1 + \bar{\lambda} q_2\right) \tag{69}$$
$$\geq \lambda\delta_1 + \bar{\lambda}\delta_2, \tag{70}$$

i.e., (p_λ, q_λ) is feasible for $\lambda\delta_1 + \bar{\lambda}\delta_2$. In addition, from the convexity of χ^α, we have that

$$\lambda\widetilde{\gamma}(\delta_1) + \bar{\lambda}\widetilde{\gamma}(\delta_2) = \lambda\chi^\alpha\left(P_{\mathsf{b},1}\|Q_{\mathsf{b},1}\right) + \bar{\lambda}\chi^\alpha\left(P_{\mathsf{b},2}\|Q_{\mathsf{b},2}\right) \tag{71}$$
$$\geq \chi^\alpha\left(P_{\mathsf{b},\lambda}\|Q_{\mathsf{b},\lambda}\right) \tag{72}$$
$$\geq \widetilde{\gamma}\left(\lambda\delta_1 + \bar{\lambda}\delta_2\right). \tag{73}$$

Therefore, the function $\widetilde{\gamma}(\delta)$ is convex in δ and hence $\bar{\mathcal{B}}_{\alpha,\varepsilon}$ is the convex hull of $\mathcal{B}_{\alpha,\varepsilon}$. In light of Theorem 2, this in turn implies that $\mathcal{R}_\alpha = \bar{\mathcal{B}}_{\alpha,\varepsilon}$.

The above analysis shows that $\delta \mapsto \widetilde{\gamma}(\delta)$ in fact constitutes the upper boundary of $\mathcal{B}_{\alpha,\varepsilon}$ and thus $\mathcal{R}_\alpha$. Since $\chi(\cdot)$ is a bijection, this allows us to deduce

$$\gamma_\alpha^\varepsilon(\delta) = \inf_{0<p,q<1} \ \chi^{-1}\left(\chi^\alpha(P_{\mathsf{b}}\|Q_{\mathsf{b}})\right) \quad \text{s.t.} \quad \mathsf{E}_{e^\varepsilon}(P_{\mathsf{b}}\|Q_{\mathsf{b}}) \geq \delta, \tag{74}$$

and hence the optimization problem (16) can be converted to the above two-parameter optimization problem.

Expanding both χ^α and $\mathsf{E}_{e^\varepsilon}$, we can explicitly write (74) as

$$\gamma_\alpha^\varepsilon(\delta) = \inf_{0<q<p<1} \ \frac{1}{\alpha - 1}\log\left(p^\alpha q^{1-\alpha} + \bar{p}^\alpha \bar{q}^{1-\alpha}\right) \quad \text{s.t.} \quad p - qe^\varepsilon \geq \delta, \tag{75}$$

where $\delta < 1$ and $\gamma < \infty$. Let $h(p, q; \alpha)$ indicate the objective function of the optimization problem in (75). For any given $\alpha > 1$ and $p \in (0,1)$, the partial derivative of $h(p, q; \alpha)$ with respect to q is given by

$$\frac{\partial\, h(p,q;\alpha)}{\partial q} = \frac{p^\alpha q^{-\alpha} - (1-p)^\alpha(1-q)^{-\alpha}}{p^\alpha q^{1-\alpha} + (1-p)^\alpha(1-q)^{1-\alpha}}, \tag{76}$$

which is negative for all $0 < q < p < 1$, and therefore, $h(p, q; \alpha)$ is decreasing in q. In addition, for $\varepsilon \geq 0$ and $\delta \in [0,1)$, the two constraints $0 < q < p < 1$ and $p - qe^\varepsilon \geq \delta$ in (75) can be equivalently rewritten as

$$\begin{cases} \delta < p < 1 \\ 0 < q < \frac{p-\delta}{e^\varepsilon}. \end{cases} \tag{77}$$

Thus, the infimum in (75) is attained at $q = \frac{p-\delta}{e^\varepsilon}$, and therefore, for $\alpha > 1$, $\delta \in [0,1)$ and $\varepsilon \geq 0$, the optimization problem in (75) is simplified as

$$e^{(\alpha-1)\left(\gamma_\alpha^\varepsilon(\delta)-\varepsilon\right)} = \inf_{p\in(\delta,1)} \ p^\alpha(p-\delta)^{1-\alpha} + \bar{p}^\alpha\left(e^\varepsilon - p + \delta\right)^{1-\alpha}, \tag{78}$$

which is the desired result. ∎

Appendix C
Proof of Theorem 4

Recall that the optimization problem in Theorem 3 is equivalent to (78). Let $h_1(p; \alpha, \delta, \varepsilon)$ indicate the objective function in (78). One can verify that for $\alpha > 1$, $\delta \in [0,1)$ and $\varepsilon > 0$, the mapping $p \mapsto h_1(p; \alpha, \delta, \varepsilon)$ is convex. Therefore, the numerical result of $\gamma_\alpha^\varepsilon(\delta)$ can be easily obtained for any given α, δ and ε.

To get closed-form expressions, we explore lower bounds of (78) as follows.

Lower bound 1: Ignoring the second term in $h_1(p; \alpha, \delta, \varepsilon)$, we obtain

$$e^{(\alpha-1)\left(\gamma_\alpha^\varepsilon(\delta)-\varepsilon\right)} \geq \inf_{p\in(\delta,1)} \ p^\alpha(p-\delta)^{1-\alpha} \tag{79}$$

We note that the objective function in (79) is convex in p, as it can be verified that $\frac{\partial^2}{\partial p^2} p^\alpha(p-\delta)^{1-\alpha}$ equals

$$(\alpha-1)\alpha\left(p^{\frac{\alpha}{2}}(p-\delta)^{\frac{-1-\alpha}{2}} - p^{\frac{\alpha-2}{2}}(p-\delta)^{\frac{1-\alpha}{2}}\right)^2 \geq 0,$$

and therefore, by setting the first derivative to be 0, we obtain the optimal solution for the corresponding unconstrained problem as $p^* = \alpha\delta$. Since $\alpha > 1$, it follows that the optimal solution of (79) is given by $p^* = \min\{\alpha\delta, 1\}$, and therefore

$$e^{(\alpha-1)\left(\gamma_\alpha^\varepsilon(\delta)-\varepsilon\right)} \geq \left(\delta\alpha^\alpha(\alpha-1)^{1-\alpha}\right)\mathbf{1}\{\alpha\delta < 1\} + \left((1-\delta)^{1-\alpha}\right)\mathbf{1}\{\alpha\delta \geq 1\} \quad (80)$$

with equality holds if and only if $\alpha\delta \geq 1$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Thus, if $\alpha\delta \geq 1$, we have $\gamma_\alpha^\varepsilon(\delta) = \varepsilon - \log(1-\delta)$, and if $\alpha\delta < 1$, we have the lower bound

$$\gamma_\alpha^\varepsilon(\delta) \geq \varepsilon - \frac{1}{\alpha-1}\log\left(\frac{1}{\delta\alpha}\left(1-\frac{1}{\alpha}\right)^{\alpha-1}\right) \quad (81)$$

$$= \varepsilon - \frac{1}{\alpha-1}\log\frac{\zeta_\alpha}{\delta}. \quad (82)$$

Lower bound 2: To obtain the second lower bound, we note that the function $h_1(p;\alpha,\delta,\varepsilon)$ is convex in δ. This enables us to bound $h_1(p;\alpha,\delta,\varepsilon)$ from below by using its linear approximation at $\delta = 0$. Hence we can write

$$h_1(p;\alpha,\delta,\varepsilon) \geq h_1(p;\alpha,\delta=0,\varepsilon) + \frac{\partial h_1(p;\alpha,\delta=0,\varepsilon)}{\partial\delta}\delta$$
$$= p + (\alpha-1)\delta + \left(\frac{1-p}{e^\varepsilon-p}\right)^\alpha \times \left(e^\varepsilon - p - (\alpha-1)\delta\right),$$

with equality if and only if $\delta = 0$. Therefore, we have

$$e^{(\alpha-1)\left(\gamma_\alpha^\varepsilon(\delta)-\varepsilon\right)} \geq \inf_{p\in(\delta,1)}\left(1-\left(\frac{1-p}{e^\varepsilon-p}\right)^\alpha\right)p + \left(\frac{1-p}{e^\varepsilon-p}\right)^\alpha \times \left(e^\varepsilon-(\alpha-1)\delta\right) + (\alpha-1)\delta. \quad (83)$$

Let $h_2(p;\alpha,\delta,\varepsilon)$ indicate the objective function of (83). In the following, we prove the monotonicity of $h_2(p;\alpha,\delta,\varepsilon)$ in p for $\alpha > 1$, $1 > \delta \geq 0$ and $\varepsilon \geq 0$. Taking the first derivative of $h_2(p;\alpha,\delta,\varepsilon)$ with respect to p, we have

$$\frac{\partial h_2(p;\alpha,\delta,\varepsilon)}{\partial p} = 1 + \left(\frac{1-p}{e^\varepsilon-p}\right)^\alpha \times \left(\frac{\alpha(e^\varepsilon-1)(p+(\alpha-1)\delta-e^\varepsilon)}{(e^\varepsilon-p)(1-p)} - 1\right)$$
$$=: h_3(p;\alpha,\delta,\varepsilon)$$
$$\geq 1 + \left(\frac{1-p}{e^\varepsilon-p}\right)^\alpha\left(-\frac{\alpha(e^\varepsilon-1)}{1-p} - 1\right) \quad (84)$$
$$=: h_4(p;\alpha,\varepsilon)$$
$$> h_4(p=\delta;\alpha,\varepsilon) \quad (85)$$
$$= \frac{(e^\varepsilon-\delta)^\alpha - \bar{\delta}^\alpha - \alpha(e^\varepsilon-1)\bar{\delta}^{\alpha-1}}{(e^\varepsilon-\delta)^\alpha} \quad (86)$$
$$=: \frac{h_5(\delta,\alpha,\varepsilon)}{(e^\varepsilon-\delta)^\alpha} \quad (87)$$
$$\geq \frac{h_5(\delta,\alpha,\varepsilon=0)}{(e^\varepsilon-\delta)^\alpha} = 0, \quad (88)$$

where

- the inequality in (84) follows from the fact that the function $h_3(p;\alpha,\delta,\varepsilon)$ is increasing in δ, and therefore, for $1 > \delta \geq 0$, $h_3(p;\alpha,\delta,\varepsilon) \geq h_3(p;\alpha,\delta=0,\varepsilon) = h_4(p;\alpha,\varepsilon)$
- the inequality in (85) is due to the fact that the function $h_4(p;\alpha,\varepsilon)$ is increasing in p as shown below

$$\frac{\partial h_4(p;\alpha,\varepsilon)}{\partial p} = \alpha(\alpha-1)\left(e^\varepsilon-1\right)^2\bar{p}^{\alpha-2}\left(e^\varepsilon-p\right)^{-\alpha-1} > 0$$

and therefore, for $p \in (\delta,1)$, $h_4(p;\alpha,\varepsilon) > h_4(p=\delta;\alpha,\varepsilon)$.

- the inequality in (88) is from the monotonicity of the function $h_5(\delta,\alpha,\varepsilon)$ in ε. Specifically,

$$\frac{\partial h_5(\delta,\alpha,\varepsilon)}{\partial\varepsilon} = \alpha e^\varepsilon\left(\left(e^\varepsilon-\delta\right)^{\alpha-1} - (1-\delta)^{\alpha-1}\right) \geq 0$$

and thus, for $\varepsilon \geq 0$, $h_5(\delta,\alpha,\varepsilon) \geq h_5(\delta,\alpha,\varepsilon=0) = 0$.

Therefore, the objective function $h_2(p;\alpha,\delta,\varepsilon)$ in (83) is increasing in p, and therefore, we have

$$e^{(\alpha-1)\left(\gamma_\alpha^\varepsilon(\delta)-\varepsilon\right)} \geq h_2(p=\delta;\alpha,\delta,\varepsilon) \quad (89)$$
$$= \alpha\delta + \left(\frac{1-\delta}{e^\varepsilon-\delta}\right)^\alpha\left(e^\varepsilon-\alpha\delta\right) \quad (90)$$

with equality if and only if $\delta = 0$. Thus, we have

$$\gamma_\alpha^\varepsilon(\delta) \geq \varepsilon + \frac{1}{\alpha-1}\log\left(\alpha\delta + \left(\frac{1-\delta}{e^\varepsilon-\delta}\right)^\alpha\left(e^\varepsilon-\alpha\delta\right)\right) \quad (91)$$

where the equality holds if and only if $\delta = 0$ which leads to $\gamma_\alpha^\varepsilon(\delta=0) = 0$. The lower bounds (82) and (91) give the desired result. ■

APPENDIX D
PROOF OF LEMMA 1

From the first part of the proof of Theorem 4, we have

$$\varepsilon_\alpha^\delta(\gamma)\begin{cases} \leq \left(\gamma - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right)_+, & \text{if } \alpha\delta \leq 1 \\ = (\gamma + \log(1-\delta))_+ & \text{otherwise.}\end{cases} \quad (92)$$

Next, we obtain a closed-form upper bound on $\varepsilon_\alpha^\delta(\gamma)$ from the function $f(\alpha,\varepsilon,\delta)$ in Theorem 4. To do so, let $f_1(\alpha,\varepsilon,\delta)$ be the expression inside the logarithm in $f(\alpha,\varepsilon,\delta)$, i.e., $f_1(\alpha,\varepsilon,\delta) \coloneqq (e^\varepsilon-\alpha\delta)(\frac{\delta-1}{\delta-e^\varepsilon})^\alpha + \alpha\delta$. The second partial derivative of $f_1(\delta,\alpha,\varepsilon)$ with respect to δ is given by

$$(\alpha-1)\alpha\left(e^\varepsilon-1\right)\bar{\delta}^\alpha\frac{(e^\varepsilon(1-2\delta+e^\varepsilon)-\alpha\delta(e^\varepsilon-1))}{\bar{\delta}^2(e^\varepsilon-\delta)^\alpha(\delta-e^\varepsilon)^2}.$$

Therefore, for $\alpha > 1$, $\varepsilon \geq 0$ and $\delta \in (0,1)$, the convexity of $f_1(\delta,\alpha,\varepsilon)$ in δ is guaranteed by

$$\delta - \frac{e^\varepsilon(e^\varepsilon+1)}{2e^\varepsilon+\alpha(e^\varepsilon-1)} \leq 0. \quad (93)$$

Let $f_2(\alpha,\varepsilon) \coloneqq \frac{e^\varepsilon(e^\varepsilon+1)}{2e^\varepsilon+\alpha(e^\varepsilon-1)}$, and therefore, if $\delta - f_2(\alpha,\varepsilon) \leq 0$, we have

$$\gamma_\alpha^\varepsilon(\delta) \geq f(\alpha,\varepsilon,\delta) = \varepsilon + \frac{1}{\alpha-1}\log(f_1(\alpha,\varepsilon,\delta)) \quad (94)$$
$$\geq \varepsilon + \frac{1}{\alpha-1}\log\left(f_1(\alpha,\varepsilon,\delta=0) + \frac{\partial f_1(\delta=0,\alpha,\varepsilon)}{\partial\delta}\delta\right)$$
$$= \varepsilon + \frac{1}{\alpha-1}\log\left(e^{-\varepsilon(\alpha-1)} + \alpha\delta - \alpha\delta e^{-\varepsilon(\alpha-1)}\right), \quad (95)$$

with equality if and only if $\delta = 0$. In the following, we prove that $\delta \leq \frac{1}{\alpha}$ is a sufficient condition for $\delta - f_2(\alpha,\varepsilon) \leq 0$ by

showing that $f_2(\alpha,\varepsilon) > 1/\alpha$ for any $\alpha > 1$. Taking the first partial derivative of $f_2(\alpha,\varepsilon)$ with respect to ε, we have

$$\frac{\partial f_2(\alpha,\varepsilon)}{\partial \varepsilon} = \frac{e^{\varepsilon}\big((2+\alpha)e^{2\varepsilon} - 2\alpha e^{2\varepsilon} - \alpha\big)}{(2e^{\varepsilon} + \alpha(e^{\varepsilon}-1))^2} \tag{96}$$

$$\begin{cases} \le 0, & 1 \le e^{\varepsilon} \le \frac{\alpha+\sqrt{2\alpha(\alpha+1)}}{2+\alpha} \\ > 0, & \text{otherwise,} \end{cases} \tag{97}$$

and therefore,

$$\begin{aligned} f_2(\alpha,\varepsilon) - \frac{1}{\alpha} &\ge f_2\left(\alpha, \varepsilon = \log\frac{\alpha+\sqrt{2\alpha(\alpha+1)}}{2+\alpha}\right) - \frac{1}{\alpha} \\ &= \frac{2\big(\alpha^2 + \alpha\big(\sqrt{2\alpha(\alpha+1)}-1\big)-2\big)}{\alpha(2+\alpha)^2} \qquad (98) \\ &=: \frac{f_3(\alpha)}{\alpha(2+\alpha)^2} \qquad (99) \\ &> \frac{f_3(1)}{\alpha(2+\alpha)^2} = 0, \qquad (100) \end{aligned}$$

where the inequality in (100) follows from the fact that $f_3(\alpha)$ is monotonically increasing in $\alpha > 1$ as shown below:

$$\frac{\mathrm{d}f_3(\alpha)}{\mathrm{d}\alpha} = \frac{\sqrt{2}\alpha(1+2\alpha)}{\sqrt{\alpha(1+\alpha)}} + 2\sqrt{2\alpha(1+\alpha)} + 4\alpha - 2 > 0. \tag{101}$$

Therefore, from the inequality in (95), we have that for $\delta \le 1/\alpha$,

$$\varepsilon_{\alpha}^{\delta}(\gamma) \le \frac{1}{\alpha-1}\log\left(\frac{e^{(\alpha-1)\gamma}-1}{\alpha\delta} + 1\right) \tag{102}$$

and equality holds if and only if $\gamma = 0$, i.e., $\varepsilon_{\alpha}^{\delta}(\gamma=0) = 0$. ■

Appendix E
Derivation of Remark 1

Note that it can be verified that $\gamma - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha} < 0$ for $\delta > \zeta_\alpha e^{(\alpha-1)\gamma}$. Combined with $\alpha\delta \le 1$, we therefore have $\varepsilon_{\alpha}^{\delta}(\gamma) = 0$ for $\delta \in [\zeta_\alpha e^{(\alpha-1)\gamma}, \frac{1}{\alpha}]$. To have a valid non-empty interval, we must have the condition $\zeta_\alpha e^{(\alpha-1)\gamma} < \frac{1}{\alpha}$ that is simplified to $1-e^{-\gamma} \le \frac{1}{\alpha}$. A similar holds for the case $\alpha\delta > 1$: we have $\gamma + \log(1-\delta) < 0$ if $\delta > 1-e^{-\gamma}$. Hence, $\varepsilon_{\alpha}^{\delta}(\gamma) = 0$ if $\delta > \max\{1-e^{-\gamma}, 1/\alpha\}$.

Appendix F
Proof of Lemma 2

Recall that for the T-fold composition of Gaussian mechanism with variance σ^2, we have $\gamma(\alpha) = \alpha\rho T$ where $\rho = 1/\sigma^2$. From Lemma 1, we have that for $\alpha\delta \ge 1$ and $0 < \delta < 1$,

$$\varepsilon_{\alpha}^{\delta}(\rho\alpha T) = (\rho\alpha T + \log(1-\delta))_+ \tag{103}$$

and therefore,

$$\begin{aligned} \varepsilon^{\delta}(\rho,T) &= \inf_{\alpha>1} \varepsilon_{\alpha}^{\delta}(\rho\alpha T) \qquad (104) \\ &\le \inf_{\alpha \ge \frac{1}{\delta}} (\rho\alpha T + \log(1-\delta))_+ \qquad (105) \\ &= \left(\frac{\rho T}{\delta} + \log(1-\delta)\right)_+. \qquad (106) \end{aligned}$$

In addition, from Lemma 1, we have that for $0 < \alpha\delta < 1$,

$$\begin{aligned} \varepsilon_{\alpha}^{\delta}(\alpha\rho T) \le \min\Bigg\{&\left(\alpha\rho T - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right)_+, \\ &\times \frac{1}{\alpha-1}\log\left(\frac{(\alpha-1)\chi(\alpha\rho T)}{\alpha\delta} + 1\right)\Bigg\}, \end{aligned}$$

where $\chi(\alpha\rho T) = \frac{e^{\rho\alpha(\alpha-1)T}-1}{\alpha-1}$, and therefore,

$$\begin{aligned} \varepsilon^{\delta}(\rho,T) &= \inf_{\alpha>1} \varepsilon_{\alpha}^{\delta}(\rho\alpha T) \\ &\le \inf_{1<\alpha<\frac{1}{\delta}} \min\Bigg\{\left(\alpha\rho T - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}\right)_+, \\ &\qquad \frac{1}{\alpha-1}\log\left(1 + \frac{e^{\rho\alpha(\alpha-1)T}-1}{\alpha\delta}\right)\Bigg\}. \end{aligned} \tag{107}$$

Combining the two inequalities in (106) and (107), we obtain the upper bound of $\varepsilon^{\delta}(\rho,T)$ in Lemma 2. ■

Appendix G
Proof of Theorem 5

Lemma 2 illustrates that the T-fold adaptive homogeneous composition of the Gaussian mechanism with variance σ^2 is (ε,δ)-DP where

$$\varepsilon = \inf_{1<\alpha\le\frac{1}{\delta}} \frac{\alpha T}{2\sigma^2} - \frac{1}{\alpha-1}\log\frac{\delta}{\zeta_\alpha}. \tag{108}$$

Assuming that $\frac{2\log\delta^{-1}}{\varepsilon} \le \frac{1}{\delta}$, or equivalently $\varepsilon \ge 2\delta\log\delta^{-1}$, we can plug $\alpha = \frac{2\log\delta^{-1}}{\varepsilon}$ in the above expression to derive the following lower bound for σ^2

$$\frac{\left(\varepsilon - 2\log\frac{1}{\delta}\right)T\log\frac{1}{\delta}}{\varepsilon^2\left(\varepsilon - \log\frac{1}{\delta} + \frac{-\varepsilon+2\log\frac{1}{\delta}}{\varepsilon}\log\left(\frac{-\varepsilon+2\log\frac{1}{\delta}}{2\log\frac{1}{\delta}}\right) - \log\left(\frac{2\log\frac{1}{\delta}}{\varepsilon}\right)\right)} \tag{109}$$

$$\begin{aligned} &= \frac{2T\log\frac{1}{\delta}}{\varepsilon^2} + \frac{T}{\varepsilon} - \frac{2T\left(\log\left(2\log\frac{1}{\delta}\right) + 1 - \log\varepsilon\right)}{\varepsilon^2} + \frac{T}{2\varepsilon^2\log\frac{1}{\delta}} \\ &\quad \times \left[4\log^2 A + (8-6\varepsilon)\log A + 2\varepsilon^2 - 5\varepsilon + 4\right] + O\left(\frac{1}{\log^2\frac{1}{\delta}}\right) \qquad (110) \\ &= \frac{2T}{\varepsilon^2}\log\frac{1}{\delta} + \frac{T}{\varepsilon} - \frac{2T}{\varepsilon^2}\left(\log\left(2\log\delta^{-1}\right) + 1 - \log\varepsilon\right) \\ &\quad + O\left(\frac{\log^2\left(\log\delta^{-1}\right)}{\log\delta^{-1}}\right). \qquad (111) \end{aligned}$$

where

- the expression in (110) is the Taylor expansion of (109) at $\delta = 0$ and $A := \frac{\log\frac{1}{\delta^2}}{\varepsilon}$,
- in (110) as $\delta \to 0$, we have $\log\delta^{-1} \to \infty$, therefore, for any fixed finite ε and T, the fourth term is of order $O(\frac{\log^2(\log\delta^{-1})}{\log\delta^{-1}})$ and dominates $O(\frac{1}{\log^2\delta^{-1}})$.

It is worth mentioning that this choice of α has already appeared in literature, see [43, Discussion after Thm 35]. ■

APPENDIX H
PROOF OF PROPOSITION 1

Recall that $\mathcal{M}_d$ and $\mathcal{M}_{d'}$ are the output distributions of mechanism $\mathcal{M}$ when running on two neighboring datasets d and d', respectively. For notational simplicity, let P and Q denote $\mathcal{M}_d$ and $\mathcal{M}_{d'}$, respectively and also $\beta(\tau)$ denote $\beta_{\mathcal{M}}^{dd'}(\tau)$. We wish to prove a more general result than Proposition 1: For any convex real-valued function f with $f(1) = 0$, we show

$$D_f(P\|Q) = \int_0^1 |\beta'(\tau)| f\left(\frac{1}{|\beta'(\tau)|}\right) \mathrm{d}\tau, \tag{112}$$

where $\beta'(\tau) \coloneqq \frac{\mathrm{d}}{\mathrm{d}\tau}\beta(\tau)$. For a given $\lambda \geq 0$, define

$$\tau_\lambda \coloneqq P\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \leq \lambda\right) = \int \mathbf{1}\left\{\frac{\mathrm{d}P}{\mathrm{d}Q} \leq \lambda\right\} \mathrm{d}P, \tag{113}$$

where, as before, $\mathbf{1}\{\cdot\}$ denotes the indicator function. Then, since $\tau \mapsto \beta(\tau)$ specifies the optimal tradeoff between type I and type II error probabilities of testing P against Q, we have from Neyman-Pearson lemma that

$$\beta(\tau_\lambda) = Q\left(\frac{\mathrm{d}P}{\mathrm{d}Q} > \lambda\right). \tag{114}$$

Before we begin the proof of (112), we need the following fact that will be needed later.

Fact: We have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\tau_\lambda = -\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda}\beta(\tau_\lambda). \tag{115}$$

We prove this fact as follows:

$$\bar{\tau}_\lambda - \lambda\beta(\tau_\lambda) = 1 - \int \left[\frac{\mathrm{d}P}{\mathrm{d}Q}\mathbf{1}\left\{\frac{\mathrm{d}P}{\mathrm{d}Q} \leq \lambda\right\} + \lambda \mathbf{1}\left\{\frac{\mathrm{d}P}{\mathrm{d}Q} > \lambda\right\}\right] \mathrm{d}Q \tag{116}$$

$$= 1 - \int \min\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}, \lambda\right\} \mathrm{d}Q \tag{117}$$

$$= 1 - \int_0^1 P\left[\frac{\mathrm{d}Q}{\mathrm{d}P} \geq \frac{t}{\lambda}\right] \mathrm{d}t \tag{118}$$

$$= \lambda \int_\lambda^\infty \frac{1}{t^2} P\left[\frac{\mathrm{d}P}{\mathrm{d}Q} > t\right] \mathrm{d}t \tag{119}$$

where equality in (118) comes from the formula that $\mathbb{E}[U] = \int \Pr(U \geq t)\mathrm{d}t$ for any non-negative random variable U. We can hence write

$$\frac{1 - \tau_\lambda - \lambda\beta(\tau_\lambda)}{\lambda} = \int_\lambda^\infty \frac{1}{t^2} P\left[\frac{\mathrm{d}P}{\mathrm{d}Q} > t\right] \mathrm{d}t. \tag{120}$$

Taking a derivative, with respect to λ, of both sides of this identity, we obtain the desired result (115). It is worth noting that if we consider E_λ-divergence for any non-negative λ (rather than $\lambda \geq 1$), then the left-hand side of (116) is in fact equal to $\mathsf{E}_\lambda(P\|Q)$, because it can be easily verified that

$$\mathsf{E}_\lambda(P\|Q) = \sup_{A \subset \mathbb{X}} P(A) - \lambda Q(A) = \bar{\tau}_\lambda - \lambda\beta(\tau_\lambda).$$

Hence, (119) gives an equivalent formula for E_λ-divergence for $\lambda \geq 0$.

Proof of (112): We have

$$D_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\frac{\mathrm{d}Q}{\mathrm{d}P}\mathrm{d}P = \int_0^\infty f(t)\frac{1}{t}\mathrm{d}\tau_t \tag{121}$$

$$= -\int_0^\infty f\left(-\frac{\mathrm{d}\tau_t/\mathrm{d}t}{\mathrm{d}\beta(\tau_t)/\mathrm{d}t}\right)\frac{\mathrm{d}\beta(\tau_t)/\mathrm{d}t}{\mathrm{d}\tau_t/\mathrm{d}t}\mathrm{d}\tau_t \tag{122}$$

$$= -\int_0^1 f\left(-\frac{1}{\beta'(\tau)}\right)\beta'(\tau)\mathrm{d}\tau \tag{123}$$

where the equality in (122) follows from (115). The desired result follows by noticing that $\tau \mapsto \beta(\tau)$ is decreasing and hence $\beta'(\tau) \leq 0$ implying that $-\beta'(\tau) = |\beta'(\tau)|$. ∎

Plugging $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$ for some $\alpha > 1$ into (112), we obtain

$$\chi^\alpha(P\|Q) = \frac{1}{\alpha - 1}\left[-\beta(0) + \int_0^1 \left|\beta'(\tau)\right|^{1-\alpha}\mathrm{d}\tau\right], \tag{124}$$

implying

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\left(1 - \beta(0) + \int_0^1 \left(\left|\beta'(\tau)\right|\right)^{1-\alpha}\mathrm{d}\tau\right).$$

REFERENCES

[1] 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," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, CA, USA, 2020, pp. 920–925.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Conf. Theory Cryptogr. (TCC)*, 2006, pp. 265–284.

[3] M. A. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA, USA: Holden-Day, 1964.

[4] I. Mironov, "Rényi differential privacy," in *Proc. IEEE Comput. Security Found. Symp. (CSF)*, Santa Barbara, CA, USA, 2017, pp. 263–275.

[5] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. Conf. Comput. Commun. Security (CCS)*, 2016, pp. 308–318.

[6] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 53rd Annu. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, 2015, pp. 1310–1321.

[7] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Res.*, vol. 12, pp. 1069–1109, Mar. 2011.

[8] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proc. IEEE 55th Symp. Found. Comput. Sci. (FOCS)*, 2014, pp. 464–473.

[9] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 6280–6290.

[10] 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 *Proc. ACM Int. Conf. Manag. Data*, 2017, pp. 1307–1322.

[11] 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

[12] *Tensorflow Privacy*, Google, Mountain View, CA, USA, 2018. [Online]. Available: https://github.com/tensorflow/privacy

[13] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Proc. IEEE 51st Annu. Symp. Found. Comput. Sci. (FOCS)*, Las Vegas, NV, USA, 2010, pp. 51–60.

[14] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 4037–4049, Jun. 2017.

[15] P. Harremoës and I. Vajda, "On pairs of f-divergences and their joint range," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3230–3235, Jun. 2011.

[16] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *J. Amer. Stat. Assoc.*, vol. 105, no. 489, pp. 375–389, 2010.

[17] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," 2019. [Online]. Available: http://arxiv.org/abs/1905.02383

[18] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato, "Hypothesis testing interpretations and Rényi differential privacy," in *Proc. 33rd Int. Conf. Artif. Intell. Stat. (AISTAT)*, 2020, pp. 2496–2506.

[19] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Dept. Elect. Eng., Princeton Univ., Princeton, NJ, USA, Sep. 2010.

[20] Z. Bu, J. Dong, Q. Long, and W. J. Su, "Deep learning with Gaussian differential privacy," 2019. [Online]. Available: arXiv 1911.11607.
[21] K. Chaudhuri and N. Mishra, "When random sampling preserves privacy," in *Proc. Int. Conf. on Advances in Cryptology*, Santa Barbara, CA, USA, 2006, pp. 198–213.
[22] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization, revisited," in *Proc. ICML Workshop Learn. Security Privacy*, Jun. 2014, pp. 1–35. [Online]. Available: http://arxiv.org/abs/1405.7085
[23] R. Bassily, V. Feldman, K. Talwar, and A. G. Thakurta, "Private stochastic convex optimization with optimal rates," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 11282–11291.
[24] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 289–296.
[25] P. Jain, P. Kothari, and A. Thakurta, "Differentially private online learning," in *Proc. 25th Annu. Conf. Learn. Theory (COLT)*, vol. 23, 2012, pp. 1–34.
[26] A. G. Thakurta and A. Smith, "Differentially private feature selection via stability arguments, and the robustness of the lasso," in *Proc. 26th Annu. Conf. Learn. Theory (COLT)*, 2013, pp. 819–850.
[27] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, 2013, pp. 245–248.
[28] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci. (FOCS)*, Berkeley, CA, USA, 2013, pp. 429–438.
[29] P. Jain and A. G. Thakurta, "(Near) dimension independent risk bounds for differentially private learning," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 476–484.
[30] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *Proc. IEEE Symp. Security Privacy (SP)*, San Jose, CA, USA, 2017, pp. 58–77.
[31] K. Talwar, A. Thakurta, and L. Zhang, "Nearly-optimal private lasso," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2015, pp. 3025–3033.
[32] D. Wang, M. Ye, and J. Xu, "Differentially private empirical risk minimization revisited: Faster and more general," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 2719–2728.
[33] J. Murtagh and S. Vadhan, "The complexity of computing the optimal composition of differential privacy," in *Proc. Int. Conf. Theory Cryptogr.*, 2016, pp. 157–175.
[34] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," 2016. [Online]. Available: https://arxiv.org/abs/1603.01887
[35] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Proc. Theory Cryptogr. Conf.*, 2016, pp. 635–658.
[36] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke, "Composable and versatile privacy via truncated CDP," in *Proc. 50th Annu. ACM SIGACT Symp. Theory Comput. (STOC)*, 2018, pp. 74–86.
[37] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled Rényi differential privacy and analytical moments accountant," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2018, pp. 1226–1235.
[38] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar*, vol. 2, nos. 3–4, pp. 299–318, 1967.
[39] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Stat.*, vol. 28, no. 1, pp. 131–142, 1966.
[40] B. Balle, G. Barthe, M. Gaboardi, and J. Geumlek, "Privacy amplification by mixing and diffusion mechanisms," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 13298–13308.
[41] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.
[42] J. Geumlek, S. Song, and K. Chaudhuri, "Rényi differential privacy mechanisms for posterior sampling," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5289–5298.
[43] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," in *Proc. IEEE 59th Annu. Symp. Found. Comput. Sci. (FOCS)*, Paris, France, 2018, pp. 521–532.
[44] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018. [Online]. Available: arXiv:1812.00984
[45] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[46] N. Sharma and N. A. Warsi, "Fundamental bound on the reliability of quantum information transmission," 2013. [Online]. Available: http://arxiv.org/abs/1302.5281
[47] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[48] G. Barthe and F. Olmedo, "Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs," in *Proc. Int. Colloq. Automata Lang. Program. (ICALP)*, 2013, pp. 49–60.
[49] T. van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
[50] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. Allerton Conf. Commun. Control Comput. (Allerton)*, 2010, pp. 1327–1333.
[51] I. Sason, "On the rényi divergence, joint range of relative entropies, and a channel coding theorem," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 23–34, Jan. 2016.