Journal of Privacy and Confidentiality
Vol. 10 (2) 2020 TPDP 2018
Submitted March 2019
Published June 2020

SUBSAMPLED RÉNYI DIFFERENTIAL PRIVACY AND ANALYTICAL MOMENTS ACCOUNTANT

YU-XIANG WANG, BORJA BALLE, AND SHIVA KASIVISWANATHAN

Computer Science Department, UC Santa Barbara, CA, USA

e-mail address: yuxiangw@cs.ucsb.edu

Google Deepmind, Cambridge, UK

e-mail address: borja.balle@gmail.com

Amazon AI, Sunnyvale, CA, USA

e-mail address: kasivisw@gmail.com

Abstract. We study the problem of subsampling in differential privacy (DP), a question that is the centerpiece behind many successful differentially private machine learning algorithms. Specifically, we provide a tight upper bound on the Rényi Differential Privacy (RDP) (Mironov, 2017) parameters for algorithms that: (1) subsample the dataset, and then (2) apply a randomized mechanism $\mathcal{M}$ to the subsample, in terms of the RDP parameters of $\mathcal{M}$ and the subsampling probability parameter. Our results generalize the moments accounting technique, developed by Abadi et al. (2016) for the Gaussian mechanism, to any subsampled RDP mechanism.

Key words and phrases: Rényi Differential Privacy, Amplification by Subsampling, Moments Accountant.

* A preliminary version of this work appeared at 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019 and at Theory and Practice of Differential Privacy (TPDP) 2018.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.724

1. Introduction

Differential privacy (DP) is a mathematical definition of privacy proposed by Dwork et al. (2006b). Ever since its introduction, DP has been widely adopted and as of today, it has become the *de facto* privacy definition in the academic world with also wide adoption in industry (Erlingsson et al., 2014; Apple, 2017; Uber Security, 2017). DP provides provable protection against adversaries with arbitrary side information and computational power, allows clear quantification of privacy losses, and satisfies graceful composition over multiple access to the same data. Over the past decade, a large body of work has been developed to design basic algorithms and tools for achieving differential privacy, understanding the privacy-utility trade-offs in different data access setups, and on integrating differential privacy with machine learning and statistical inference. We refer the reader to (Dwork and Roth, 2013) for a more comprehensive overview.

Rényi Differential Privacy (RDP, see Definition 4) (Mironov, 2017) is a recent refinement of differential privacy (Dwork et al., 2006b). It offers a unified view of the ϵ-differential privacy (pure DP), (ϵ, δ)-differential privacy (approximate DP), and the related notion of *Concentrated Differential Privacy* (Dwork and Rothblum, 2016; Bun and Steinke, 2016). The RDP point of view on differential privacy is particularly useful when the dataset is accessed by a sequence of randomized mechanisms, as in this case a *moments accountant* technique can be used to effectively keep track of the usual (ϵ, δ) DP parameters across the entire range $\{(\epsilon(\delta), \delta) | \forall \delta \in [0, 1]\}$ (Abadi et al., 2016).

A prime use case for the moments accountant technique is the *NoisySGD* algorithm (Song et al., 2013; Bassily et al., 2014) for differentially private learning, which iteratively executes:

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla f_i(\theta_t) + Z_t \right) \tag{1.1}$$

where θ_t is the model parameter at tth step, η_t is the learning rate, f_i is the loss function of data point i, ∇ is the standard gradient operator, $\mathcal{I}$ is an index set of size m drawn uniformly at random from $\{1, ..., n\}$, and $Z_t \sim \mathcal{N}(0, \sigma^2 I)$. Adding Gaussian noise (also known as the *Gaussian mechanism*) is a standard way of achieving (ϵ, δ)-differential privacy (Dwork et al., 2006a; Dwork and Roth, 2013; Balle and Wang, 2018). Since in the NoisySGD case the randomized algorithm first chooses (subsamples) the mini-batch $\mathcal{I}$ randomly before adding the Gaussian noise, the overall scheme can be viewed as a *subsampled Gaussian mechanism*. Therefore, with the right setting of σ, each iteration of NoisySGD can be thought of as a private release of a stochastic gradient.

More generally, a *subsampled mechanism* first takes a subsample of the input dataset through some subsampling procedure[1], and then applies a known randomized mechanism $\mathcal{M}$ on the subsampled data points. It is important to exploit the randomness in subsampling because if $\mathcal{M}$ is (ϵ, δ)-DP, then (informally) a subsampled mechanism obeys $(O(\gamma\epsilon), \gamma\delta)$-DP for some $\gamma < 1$ related to the sampling procedure. This is often referred to as the "privacy amplification" lemma[2] – a key property that enables NoisySGD and variants to achieve optimal rates in convex problems (Bassily et al., 2014), and to work competitively in Bayesian

[1]There are various different subsampling methods, such as Poisson subsampling, sampling without replacement, sampling with replacement, etc. We will formalize this later.

[2]Informally, this lemma states that, if a private algorithm is run on a random subset of a larger dataset (and the identity of that subset remains hidden), then this new algorithm provides better privacy protection

learning (Wang et al., 2015) and deep learning (Abadi et al., 2016) settings. A side note is that privacy amplification is also the key underlying technical tool for characterizing learnability in statistical learning (Wang et al., 2016) and achieving tight sample complexity bounds for simple function classes (Beimel et al., 2013; Bun et al., 2015).

While privacy amplification via subsampling is a very important tool for designing good private algorithms, computing the RDP parameters for a subsampled mechanism is a non-trivial task. A natural question, with wide ranging implications for designing successful differentially private algorithms is the following: Can we obtain good bounds for privacy parameters of a subsampled mechanism in terms of privacy parameters of the original mechanism? With the exception of the special case of the Gaussian mechanism under Poisson subsampling analyzed in (Abadi et al., 2016), there is no analytical formula available to generically convert the RDP parameters of a mechanism $\mathcal{M}$ to the RDP parameters of the subsampled mechanism.

In this paper, we tackle this central problem in private data analysis and provide the first general result in this area. Specifically, we analyze RDP amplification under a *sampling without replacement* procedure: $\mathsf{subsample}$, which takes a dataset of n points and outputs a sample from the uniform distribution over all subsets of size $m \leq n$. Our contributions can be summarized as follows:

(i) We provide an explicit bound (Theorem 9) on the RDP parameter ($\epsilon_{\mathcal{M}\circ\mathsf{subsample}}(\alpha)$) for a subsampled mechanism ($\mathcal{M}\circ\mathsf{subsample}$) in terms of the RDP parameter ($\epsilon_{\mathcal{M}}(\alpha)$) of the original mechanism ($\mathcal{M}$) itself and the subsampling ratio $\gamma := m/n$. Here, α is the order of the Rényi divergence in the RDP definition (see Definition 4 and the following discussion). This is the first general result in this area that can be applied to any RDP mechanism. For example, in addition to providing RDP parameter bounds for the subsampled Gaussian mechanism case, our result enables analytic calculation of similar bounds for many more commonly used privacy mechanisms including subsampled Laplace mechanisms, subsampled randomized response mechanisms, subsampled "posterior sampling" algorithms under exponential family models (Geumlek et al., 2017), etc. Even for the subsampled Gaussian mechanism our bounds are tighter than those provided by Abadi et al. (2016) (albeit the subsampling procedure and the dataset neighboring relation they use are slightly different from ours).

(ii) Consider a mechanism $\mathcal{M}$ with RDP parameter $\epsilon_{\mathcal{M}}(\alpha)$. Interestingly, our bound on the RDP parameter of the subsampled mechanism indicates that as the order of RDP α increases, there is a phase transition around the point α^* satisfying $\gamma\alpha^* e^{\epsilon_{\mathcal{M}}(\alpha^*)} = 1$. For $\alpha < \alpha^*$, the subsampled mechanism has an RDP parameter $\epsilon_{\mathcal{M}\circ\mathsf{subsample}}(\alpha) = O(\alpha\gamma^2(e^{\epsilon_{\mathcal{M}}(2)} - 1))$, while for $\alpha > \alpha^*$, the RDP parameter $\epsilon_{\mathcal{M}\circ\mathsf{subsample}}(\alpha)$ either quickly converges to $\epsilon_{\mathcal{M}}(\alpha)$ which does not depend on γ, or tapers off at $O(\gamma\epsilon_{\mathcal{M}}(\infty))$ which happens when $e^{\epsilon_{\mathcal{M}}(\infty)} - 1 \ll 1/\gamma$. The subsampled Gaussian mechanism falls into the first category, while the subsampled Laplace mechanism falls into the second.

(iii) Our analysis reveals a new theoretical quantity of interest that has not been investigated before — a *ternary* version of the Pearson-Vajda divergence (formally defined

(reflected through improved privacy parameters) to the entire dataset as a whole than the original algorithm did.

in Section 3.1). A privacy definition defined through this divergence seems naturally coupled with understanding the effects of subsampling, just like how Rényi differential privacy (RDP) (Mironov, 2017) seems naturally coupled with understanding the effects of composition.

(iv) From a computational efficiency perspective, we propose an efficient data structure to keep track of the Rényi differential privacy parameters in its symbolic form, and output the corresponding (ϵ, δ)-differential privacy as needed using efficient numerical methods. This avoids the need to specify a discrete list of moments ahead of time as required in the *moments accountant* method of Abadi et al. (2016) (see the discussion in Section 4). Finally, our experiments confirm the improvements in privacy parameters that can be obtained by applying our bounds.

We end this introduction with a methodological remark. The main result of this paper is the bound in Theorem 9, which at first glance looks cumbersome. Remarks following the statement of the theorem in Section 3 discuss some of the asymptotic implications of this bound, as well as its meaning in several special cases. These provide intuitive explanations justifying the optimal scaling the bound in those cases. In practice, however, asymptotic bounds are of limited interest: concrete bounds with explicit constants that can be efficiently computed are needed to provide the best possible privacy-utility trade-off in practical applications of differential privacy. Thus, our results should be interpreted under this point of view, which is summarized by the leitmotif *"in differential privacy, constants matter"*.

Organization. The rest of the paper is organized as follows. In Section 2, we present some background about differential privacy and Rényi differential privacy, and mention some related work in privacy amplification through subsampling. In Section 3, we present our main result of Rényi differential privacy amplification via subsampling. Based on these bounds, in Section 4, we discuss the analytical moments accountant data structure that can efficiently track privacy parameters under composition. In Section 5, we present numerical experiments that support our theoretical bounds. We conclude in Section 6. Certain supplemental details are collected in appendices.

2. Background and Related Work

In this section, we review some background about differential privacy, some related privacy notions, and the moments accountant technique (Abadi et al., 2016).

Differential privacy and Privacy Loss Random Variable. We start with the definition of (ϵ, δ)-differential privacy. We assume that $\mathcal{X}$ is the domain that the data points are drawn from. We call two datasets X and X' *neighboring* (adjacent) if they differ in at most one data point, meaning that we can obtain X' by *replacing* one data point from X by another arbitrary data point. We represent this as $d(X, X') \leq 1$.

Definition 1 Differential Privacy**.** *A randomized algorithm $\mathcal{M} : \mathcal{X}^n \to \Theta$ is (ϵ, δ)-DP (differentially private) if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ (i.e., that differs only by one datapoint), and every possible (measurable) output set $E \subseteq \Theta$ the following inequality holds:* $\Pr[\mathcal{M}(X) \in E] \leq e^\epsilon \Pr[\mathcal{M}(X') \in E] + \delta$.

The definition ensures that it is information-theoretically impossible for an adversary to infer whether the input dataset is X or X' beyond a certain confidence, hence offering a degree of *plausible deniability* to individuals in the dataset. Here, ϵ, δ are what we call privacy loss parameters and the smaller they are, the stronger the privacy guarantee is. A helpful way to work with differential privacy is in terms of tail bounds on the *privacy loss random variable*. Let $\mathcal{M}(X)$ and $\mathcal{M}(X')$ be the probability distribution induced by $\mathcal{M}$ on neighboring datasets X and X' respectively, the *the privacy loss random variable* is defined as: $\log(\mathcal{M}(X)(\theta)/\mathcal{M}(X')(\theta))$ where $\theta \sim \mathcal{M}(X)$. Up to constant factors, (ϵ, δ)-DP (Definition 1) is equivalent to requiring that the probability of the privacy loss random variable being greater than ϵ is at most δ for all neighboring datasets X, X'.[3] An important strength of differential privacy is the ability to reason about cumulative privacy loss under composition of multiple analyses on the same dataset.

Classical design of differentially private mechanisms takes these ϵ, δ privacy parameters as inputs and then the algorithm carefully introduces some randomness to satisfy the privacy constraint (Definition 1), while simultaneously trying to achieve good utility (performance) bounds. However, this paradigm has shifted a bit recently as it has come to our realization that a more fine-grained analysis tailored for specific mechanisms could yield more favorable privacy-utility trade-offs and better privacy loss parameters under composition (See, e.g., Dwork and Rothblum, 2016; Abadi et al., 2016; Balle and Wang, 2018).

A common technique for achieving differential privacy while working with a real-valued function $f : \mathcal{X}^n \to \mathbb{R}$ is via the addition of noise calibrated to f's sensitivity S_f, which is defined as the maximum of the absolute distance $|f(X) - f(X')|$ where X, X' are adjacent inputs.[4] In this paradigm, the Gaussian mechanism is defined as: $\mathcal{G}(X) := f(X) + \mathcal{N}(0, S_f^2\sigma^2)$. A single application of the Gaussian mechanism to a function f with sensitivity S_f satisfies (ϵ, δ)-differential privacy if[5] $\delta \geq 0.8 \cdot \exp(-(\sigma\epsilon)^2/2)$ and $\epsilon \leq 1$ (Dwork and Roth, 2013, Theorem 3.22).

Stochastic Gradient Descent and Subsampling Lemma. A popular way of training machine learning models under differential privacy is to use Stochastic Gradient Descent (SGD) with differentially private releases of (sometimes clipped) gradients evaluated on mini-batches of a dataset (Song et al., 2013; Bassily et al., 2014; Wang et al., 2015; Foulds et al., 2016; Abadi et al., 2016). Algorithmically, these methods are nearly the same and are all based on the NoisySGD idea presented in (1.1). They differ primarily in how they keep track of their privacy loss. Song et al. (2013) uses a sequence of disjoint mini-batches to ensure each data point is used only once in every data pass. The results in (Bassily et al., 2014; Wang et al., 2016; Foulds et al., 2016) make use of the privacy amplification lemma to take advantage of the randomness introduced by subsampling. The first privacy amplification lemma appeared in (Kasiviswanathan et al., 2011; Beimel et al., 2013), with many subsequent improvements in different settings. For the case of (ϵ, δ)-DP, Balle et al. (2018) provide a unified account of privacy amplification techniques for different types of subsampling and dataset neighboring relations. In this paper, we work in the subsampling

[3] For meaningful guarantees, δ is typically taken to be "cryptographically" small.

[4] The restriction to a scalar-valued function is intended to simplify this presentation, but is not essential.

[5] Balle and Wang (2018) show that a more complicated relation between ϵ and δ yields an if and only if statement.

without replacement setup, which satisfies the following privacy amplification lemma for (ϵ, δ)-DP.

Definition 2 Subsample**.** *Given a dataset X of n points, the procedure* subsample *selects a random sample from the uniform distribution over all subsets of X of size m. The ratio $\gamma := m/n$ is defined as the sampling parameter of the* subsample *procedure.*

Lemma 3 (Ullman, 2017)[6]**.** *If $\mathcal{M}$ is (ϵ, δ)-DP, then $\mathcal{M}'$ that applies $\mathcal{M} \circ$ subsample obeys (ϵ', δ')-DP with $\epsilon' = \log\left(1 + \gamma(e^\epsilon - 1)\right)$ and $\delta' = \gamma\delta$.*

Roughly, the lemma says that subsampling with probability $\gamma < 1$ amplifies an (ϵ, δ)-DP algorithm to an $(O(\gamma\epsilon), \gamma\delta)$-DP algorithm for a sufficiently small choice of ϵ. The overall differentially private guarantees in (Wang et al., 2015; Bassily et al., 2014; Foulds et al., 2016) were obtained by keeping track of the privacy loss over each iterative update of the model parameters using the *strong composition theorem* in differential privacy (Dwork et al., 2010), which gives roughly $(\tilde{O}(\sqrt{k}\epsilon), \tilde{O}(k\delta))$-DP[7] for k iterations of an arbitrary (ϵ, δ)-DP algorithm (see Appendix A for a discussion about various composition results in differential privacy).

The work of Abadi et al. (2016) was the first to take advantage of the fact that $\mathcal{M}$ is a subsampled Gaussian mechanism and used a mechanism-specific way of doing the strong composition. Their technique, referred to as *moments accountant*, is described below.

Cumulant Generating Functions, Moments Accountant, and Rényi Differential Privacy. The moments accountant technique of Abadi et al. (2016) centers around the cumulant generating function (CGF, or the log of the moment generating function) of the privacy loss random variable:

$$K_{\mathcal{M}}(X, X', \lambda) := \log \mathbb{E}_{\theta\sim\mathcal{M}(X)}\left[e^{\lambda \log \frac{\mathcal{M}(X)(\theta)}{\mathcal{M}(X')(\theta)}}\right] = \log \mathbb{E}_{\theta\sim\mathcal{M}(X)}\left[\left(\frac{\mathcal{M}(X)(\theta)}{\mathcal{M}(X')(\theta)}\right)^{\lambda}\right]. \quad (2.1)$$

After a change of measure, this is equivalent to:

$$K_{\mathcal{M}}(X, X', \lambda) := \log \mathbb{E}_{\theta\sim\mathcal{M}(X')}\left[\left(\frac{\mathcal{M}(X)(\theta)}{\mathcal{M}(X')(\theta)}\right)^{\lambda+1}\right].$$

Recall that if two random variables have identical CGFs, then they are identically distributed (almost everywhere). Therefore, this function characterizes the entire distribution of the privacy loss random variable.

Before explaining the details behind the moments accountant technique, we introduce the notion of Rényi differential privacy (RDP) (Mironov, 2017) as a generalization of differential privacy that uses the α-Rényi divergences between $\mathcal{M}(X)$ and $\mathcal{M}(X')$.

[6]This result follows from Ullman's proof, though the notes state a weaker result. See also (Balle et al., 2018)

[7]The $\tilde{O}(\cdot)$ notation hides various logarithmic factors.

Definition 4 Rényi Differential Privacy. *We say that a mechanism $\mathcal{M}$ is (α, ϵ)-RDP with order $\alpha \in (1, \infty)$ if for all neighboring datasets X, X'*

$$D_\alpha(\mathcal{M}(X)\|\mathcal{M}(X')) := \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta \sim \mathcal{M}(X')} \left[\left(\frac{\mathcal{M}(X)(\theta)}{\mathcal{M}(X')(\theta)} \right)^\alpha \right] \leq \epsilon.$$

As $\alpha \to \infty$, RDP reduces to $(\epsilon, 0)$-DP (pure DP), i.e., a randomized mechanism $\mathcal{M}$ is $(\epsilon, 0)$-DP if and only if for any two adjacent inputs X and X' it satisfies $D_\infty(\mathcal{M}(X)\|\mathcal{M}(X')) \leq \epsilon$. For $\alpha \to 1$, the RDP notion reduces to the Kullback-Leibler based privacy notion, which is equivalent to a bound on the expectation of the privacy loss random variable. For a detailed exposition of the guarantee and properties of Rényi differential privacy that mirror those of differential privacy, see Mironov (2017, Section III). Here, we highlight two key properties that are relevant for this paper.

Lemma 5 Adaptive Composition of RDP, Proposition 1 of (Mironov, 2017). *If $\mathcal{M}_1$ takes a dataset as input and obeys (α, ϵ_1)-RDP, and $\mathcal{M}_2$ takes a dataset and the output of $\mathcal{M}_1$ as its input and obeys (α, ϵ_2)-RDP, then their composition obeys $(\alpha, \epsilon_1 + \epsilon_2)$-RDP.*

Lemma 6 RDP to DP conversion, Proposition 3 of (Mironov, 2017). *If $\mathcal{M}$ obeys (α, ϵ)-RDP, then $\mathcal{M}$ obeys $(\epsilon + \log(1/\delta)/(\alpha - 1), \delta)$-DP for all $0 < \delta < 1$.*

RDP Functional View. While RDP for each fixed α can be used as a standalone privacy measure, we emphasize its *functional view* in which ϵ is a function of α for $1 \leq \alpha \leq \infty$, and this function is completely determined by $\mathcal{M}$. This is denoted by $\epsilon_\mathcal{M}(\alpha)$, and with this notation, mechanism $\mathcal{M}$ satisfies $(\alpha, \epsilon_\mathcal{M}(\alpha))$-RDP in Definition 4. In other words,

$$\sup_{X, X' : d(X, X') \leq 1} D_\alpha(\mathcal{M}(X)\|\mathcal{M}(X')) \leq \epsilon_\mathcal{M}(\alpha).$$

Here $\epsilon_\mathcal{M}(\alpha)$ is referred to as the RDP parameter. We drop the subscript from $\epsilon_\mathcal{M}$ when $\mathcal{M}$ is clear from the context. We use $\epsilon_\mathcal{M}(\infty)$ (or $\epsilon(\infty)$) to denote the case where $\alpha = \infty$, which indicates that the mechanism $\mathcal{M}$ is $(\epsilon, 0)$-DP (pure DP) with $\epsilon = \epsilon(\infty)$.

Our goal is, given a mechanism $\mathcal{M}$ that satisfies $(\alpha, \epsilon(\alpha))$-RDP, to investigate the RDP parameter of the subsampled mechanism $\mathcal{M} \circ \mathsf{subsample}$, i.e., to get a bound on $\epsilon_{\mathcal{M} \circ \mathsf{subsample}}(\alpha)$ such that the mechanism $\mathcal{M} \circ \mathsf{subsample}$ satisfies $(\alpha, \epsilon_{\mathcal{M} \circ \mathsf{subsample}}(\alpha))$-RDP.

Consider a data-independent upper bound of the CGF defined as

$$K_\mathcal{M}(\lambda) := \sup_{X, X' : d(X, X') \leq 1} K_\mathcal{M}(X, X', \lambda).$$

Then the following remark follows immediately.

Remark 7 RDP $\Leftrightarrow$ CGF. *A randomized mechanism $\mathcal{M}$ obeys $(\lambda + 1, K_\mathcal{M}(\lambda)/\lambda)$-RDP for all λ.*

The idea behind moments accountant (Abadi et al., 2016) is to essentially keep track of the evaluations of CGF at a list of fixed values of λ through Lemma 5. Then Lemma 6 allows

one to find the smallest ϵ given a desired δ or vice versa using:

$$\delta \Rightarrow \epsilon : \qquad \epsilon(\delta) = \min_{\lambda} \frac{\log(1/\delta) + K_{\mathcal{M}}(\lambda)}{\lambda}, \tag{2.2}$$

$$\epsilon \Rightarrow \delta : \qquad \delta(\epsilon) = \min_{\lambda} e^{K_{\mathcal{M}}(\lambda) - \lambda\epsilon}. \tag{2.3}$$

Using the convexity of CGF $K_{\mathcal{M}}(\lambda)$ and monotonicity of $K_{\mathcal{M}}(\lambda)/\lambda$ in λ (Van Erven and Harremos, 2014, Corollary 2, Theorem 3), we observe that the optimization problem in (2.3) is log-convex and the optimization problem (2.2) is unimodal/quasi-convex. Therefore, the optimization problem in (2.2) (similarly, in (2.3)) can be solved to an arbitrary accuracy τ in time $\log(\lambda^*/\tau)$ using the bisection method, where λ^* is the optimal value for λ from (2.2) (similarly, (2.3)). The same result holds even if all we have is (possibly noisy) blackbox access to $K_{\mathcal{M}}(\cdot)$ or its derivative (see more details in Section 4).

For other useful properties of the CGF and an elementary proof of its convexity and how it implies the monotonicity of the Rényi divergence, see Appendix E.

Other Related Work. A closely related notion to RDP is that of *zero-concentrated differential privacy* (zCDP) introduced in (Bun and Steinke, 2016) (see also (Dwork and Rothblum, 2016)). zCDP is related to CGF of the privacy loss random variable as we note here.

Remark 8 Relation between CGF and Zero-concentrated Differential Privacy**.** *If randomized mechanism $\mathcal{M}$ obeys (ξ, ρ)-zCDP for some parameters ξ, ρ, then the CGF $K_{\mathcal{M}}(\lambda) \leq \lambda\xi + \lambda(\lambda+1)\rho$. On the other hand, if $\mathcal{M}$'s privacy loss r.v. has CGF $K_{\mathcal{M}}(\lambda)$, then $\mathcal{M}$ is also (ξ, ρ)-zCDP for all (ξ, ρ) such that the quadratic function $\lambda\xi + \lambda(\lambda+1)\rho \geq K_{\mathcal{M}}(\lambda)$.*

For CDP, subsampling does not improve the privacy parameters (Bun et al., 2018). A truncated variant of the zCDP has been very recently proposed by Bun et al. (2018) and they studied the effect of subsampling in tCDP. In particular, (Bun et al., 2018) show that this truncated concentrated differential privacy (tCDP) definition satisfies a privacy amplification property. Since tCDP provides an upper bound of the RDP function, the amplification bounds of tCDP implies an RDP bound up to some restricted order α. The focus on our paper is on a closely related problem of understanding the effects of subsampling on RDP directly.

3. Privacy Amplification for RDP

In this section, we present first our main result, an amplification theorem for Rényi differential privacy via subsampling. We first provide the upper bound, and then discuss the optimality of this bound in Section 3.3.

We start with our main theorem that bounds $\epsilon_{\mathcal{M}\circ\mathsf{subsample}}(\alpha)$ for the mechanism $\mathcal{M}\circ\mathsf{subsample}$ in terms of $\epsilon_{\mathcal{M}}(\alpha)$ of the mechanism $\mathcal{M}$ and sampling parameter γ used in the $\mathsf{subsample}$ procedure.

Theorem 9 RDP for Subsampled Mechanisms**.** *Given a dataset of n points drawn from a domain $\mathcal{X}$ and a (randomized) mechanism $\mathcal{M}$ that takes an input from $\mathcal{X}^m$ for $m \leq n$, let the randomized algorithm $\mathcal{M} \circ \mathsf{subsample}$ be defined as: (1) subsample without replacement m*

data points of the dataset (sampling parameter $\gamma = m/n$), and (2) apply $\mathcal{M}$ to the subsampled dataset. If $\mathcal{M}$ obeys $(\alpha, \epsilon(\alpha))$-RDP for all $\alpha \geq 2$, then for all integers $\alpha \geq 2$ the randomized algorithm $\mathcal{M} \circ \mathsf{subsample}$ obeys $(\alpha, \epsilon'(\alpha))$-RDP where,

$$\epsilon'(\alpha) \leq \frac{1}{\alpha-1}\log\left(1+\gamma^2\binom{\alpha}{2}\min\left\{4(e^{\epsilon(2)}-1), e^{\epsilon(2)}\min\{2,(e^{\epsilon(\infty)}-1)^2\}\right\}\right.$$
$$\left.+\sum_{j=3}^{\alpha}\gamma^j\binom{\alpha}{j}e^{(j-1)\epsilon(j)}\min\{2,(e^{\epsilon(\infty)}-1)^j\}\right).$$

The bound in the above theorem might appear complicated, and this is partly because of our efforts to get a precise non-asymptotic bound (and not just a $O(\cdot)$ bound) that can be implemented in practice to keep track of the privacy loss of a real system. Some additional practical considerations related to evaluating the bound in this theorem such as computational resources needed, numerical stability issues, etc., are discussed in Section 4. The phase transition behavior of this bound, noted in the introduction, is probably most easily observed through Figure 3 (Section 5), where we empirically illustrate the behavior of our bound for some commonly used subsampled mechanisms. Before describing the proof, we make a few remarks about this result.

Generality. Our results cover any Rényi differentially private mechanism, including those based on any exponential family distribution (see Geumlek et al., 2017, and our exposition in Appendix F). As mentioned earlier, previously such a bound (even asymptotically) was only known for the special case of the subsampled Gaussian mechanism (Abadi et al., 2016).

Pure DP. In particular, Theorem 9 also covers pure-DP mechanisms (such as Laplace and randomized response mechanisms) with a bounded $\epsilon(\infty)$. In this case, we can upper bound everything within the logarithm of Theorem 9 with a binomial expansion:

$$1+\sum_{j=1}^{\alpha}\gamma^j\binom{\alpha}{j}e^{j\epsilon(\alpha)}(e^{\epsilon(\infty)}-1)^j = \left(1+\gamma e^{\epsilon(\alpha)}(e^{\epsilon(\infty)}-1)\right)^{\alpha},$$

which results in a bound of the form

$$\epsilon'(\alpha) \leq \frac{\alpha}{\alpha-1}\log\left(1+\gamma e^{\epsilon(\alpha)}(e^{\epsilon(\infty)}-1)\right).$$

As $\alpha \to \infty$ the expression converges to $\log\left(1+\gamma e^{\epsilon(\infty)}(e^{\epsilon(\infty)}-1)\right)$ which gives quantitatively the same result as the privacy amplification result in Lemma 3 for the pure $(\epsilon, 0)-$DP, modulo an extra $e^{\epsilon(\infty)}$ factor whose effect becomes negligible when $\epsilon(\infty)$ is not too big.

Bound under Additional Assumptions. The bound in Theorem 9 could be strengthened under additional assumptions on the RDP guarantee. We defer a detailed discussion on this topic to Section 3.2 (see Theorem 19), but note that a consequence of this is that one can replace $e^{(j-1)\epsilon(j)}\min\{2,(e^{\epsilon(\infty)}-1)^j\}$ in the above bound with an exact evaluation given by the forward finite difference operator of some appropriately defined functional. Also we note that these additional assumptions hold for the Gaussian mechanism.

In particular, applying subsampled Gaussian mechanism for functions with sensitivity 1 (i.e., $\epsilon(\alpha) = \alpha/(2\sigma^2)$) the dominant part of the upper bound on $\epsilon'(\alpha)$ arises from the term $\min\{4(e^{\epsilon(2)}-1), e^{\epsilon(2)}\min\{2,(e^{\epsilon(\infty)}-1)^2\}\}$. Firstly, since the Gaussian mechanism does not have a bounded $\epsilon(\infty)$ term, this term can be simplified as $\min\{4(e^{\epsilon(2)}-1), 2e^{\epsilon(2)}\}$. Let

us consider the regimes: (a) σ^2 "large", (b) σ^2 "small".[8] When σ^2 is large, $4(e^{\epsilon(2)}-1) = 4(e^{1/\sigma^2}-1) \le 8/\sigma^2$ becomes the tight term in $\min\{4(e^{\epsilon(2)}-1), 2e^{\epsilon(2)}\}$. In this case, for small α and γ, the overall $\epsilon'(\alpha)$ bound simplifies to $O(\gamma^2\alpha/\sigma^2)$ (matching the asymptotic bound given in Section 3.4). When σ^2 is small, then the $2e^{\epsilon(2)} = 2e^{1/\sigma^2}$ becomes the tight term in $\min\{4(e^{\epsilon(2)}-1), 2e^{\epsilon(2)}\}$. This (small σ^2) is a regime that the results of Abadi et al. (2016) do not cover.

Integer to Real-valued α. The above calculations rely on a binomial expansion and thus only work for integer α's. To apply it to any real-valued α, we can use the relation between RDP and CGF mentioned in Remark 7, and the fact that CGF is a convex function (see Lemma 31 in Appendix E). The convexity of $K_{\mathcal{M}}(\cdot)$ implies that a piecewise linear interpolation yields a valid upper bound for all $\alpha \in (1, \infty)$.

Corollary 10 . *Let $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denotes the floor and ceiling operators. Then, $K_{\mathcal{M}}(\lambda) \le (1-\lambda+\lfloor\lambda\rfloor)K_{\mathcal{M}}(\lfloor\lambda\rfloor) + (\lambda-\lfloor\lambda\rfloor)K_{\mathcal{M}}(\lceil\lambda\rceil)$.*

Proof. The result is a simple corollary of the convexity of the CGF. Specifically, take $\lambda_1 = \lfloor\lambda\rfloor$, $\lambda_2 = \lceil\lambda\rceil$ and $v := \lambda - \lfloor\lambda\rfloor$. Note that $\lambda = (1-v)\lfloor\lambda\rfloor + v\lceil\lambda\rceil$. The result follows from the definition of convexity. □

The bound on $K_{\mathcal{M}}(\lambda)$ can be translated into a RDP parameter bound as noted in Remark 7.

3.1. **Proof of Theorem 9.** Let us start with the overall proof strategy. The proof is roughly split into three parts. In the first part, we define a new family of privacy definitions called *ternary-$|\chi|^\alpha$-differential privacy* (based on ternary version of Pearson-Vajda divergence) and show that it handles subsampling naturally (Proposition 14). In the second part, we bound the Rényi DP using the ternary-$|\chi|^\alpha$-differential privacy and apply the subsampling lemma from the first part. In the third part, we propose a number of ways of converting the expression stated as ternary-$|\chi|^\alpha$-differential privacy back to that of RDP (Lemmas 15, 16, 17). Each of these conversion strategies yield different coefficients in the sum inside the logarithm defining $\alpha'(\epsilon)$; our bound accounts for all these strategies at once by taking the minimum of these coefficients.

Pearson-Vajda Divergence and the Moments of Linearized Privacy Random Variable. We need to define a few quantities to establish our results. The Pearson-Vajda Divergence (or $|\chi|^\alpha$-divergence) of order α between two distributions (random variables) p and q is defined as follows (Vajda, 1973):

$$D_{|\chi|^\alpha}(p\|q) := \mathbb{E}_q\left[\left|\frac{p}{q} - 1\right|^\alpha\right]. \tag{3.1}$$

Here, $\mathbb{E}_q[\cdot]$ denotes the expectation over the distribution q. This is closely related to the moment of the privacy random variable in that $(p/q-1)$ is the linearized version of $\log(p/q)$. More interestingly, the αth moment of the privacy random variable is the αth derivative of the MGF evaluated at 0:

$$\mathbb{E}_q[\log(p/q)^\alpha] = \frac{\partial^\alpha}{\partial t^\alpha}[e^{K_{\mathcal{M}}(t)}](0),$$

[8] Formally, one could define σ^2 as large when $4(e^{1/\sigma^2}-1) \le 2e^{1/\sigma^2}$ and small otherwise.

while at least for the even order, the $|\chi|^\alpha$-divergence is the αth order *forward finite difference* of the MGF evaluated at 0:

$$\mathbb{E}_q[(p/q-1)^\alpha] = \Delta^{(\alpha)}[e^{K_{\mathcal{M}}(\cdot)}](0). \tag{3.2}$$

In the above expression, the αth order *forward difference operator* $\Delta^{(\alpha)}$ is defined recursively with

$$\Delta^{(\alpha)} := \underbrace{\Delta \circ ... \circ \Delta}_{\alpha\text{-times}}, \tag{3.3}$$

where Δ denotes the first order forward difference operator such that $\Delta[f](x) = f(x+1) - f(x)$ for any function $f : \mathbb{R} \to \mathbb{R}$. See Appendix B for more information on $\Delta^{(\alpha)}$ and its connection to binomial numbers.

Part 1: Ternary-$|\chi|^\alpha$-divergence and Natural Subsampling. Ternary-$|\chi|^\alpha$-divergence is a novel quantity that measures the discrepancy of three distributions instead of two. Let p, q, r be three probability distributions[9], we define

$$D_{|\chi|^\alpha}(p, q\|r) := \mathbb{E}_r\left[\left|\frac{p-q}{r}\right|^\alpha\right].$$

Here, $\mathbb{E}_r[\cdot]$ denotes the expectation over the distribution r. Using this ternary-$|\chi|^\alpha$-divergence notion, we define ζ-ternary-$|\chi|^\alpha$-differential privacy as follows. Analogously with RDP where we considered ϵ as a function of α, we consider ζ as a function of α.

Definition 11 Ternary-$|\chi|^\alpha$-differential privacy**.** *We say that a randomized mechanism $\mathcal{M}$ is ζ-ternary-$|\chi|^\alpha$-DP if for all $\alpha \geq 1$:*

$$\sup_{X,X',X'' \text{ mutually adjacent}} \Big(D_{|\chi|^\alpha}(\mathcal{M}(X), \mathcal{M}(X')\|\mathcal{M}(X''))\Big)^{1/\alpha} \leq \zeta(\alpha).$$

Here, the *mutually adjacent* condition means $d(X, X'), d(X', X''), d(X, X'') \leq 1$, and $\zeta(\alpha)$ is a function from $\mathbb{R}^+$ to $\mathbb{R}^+$. Note that the above definition is a general case of the following binary-$|\chi|^\alpha$-differential privacy definition that works with the standard Person-Vajda $|\chi|^\alpha$-divergences (as defined in (3.1)).

Definition 12 Binary-$|\chi|^\alpha$-differential privacy**.** *We say that a randomized mechanism $\mathcal{M}$ is ξ-binary-$|\chi|^\alpha$-DP if for all $\alpha \geq 1$:*

$$\sup_{X,X':d(X,X')\leq 1} \Big(D_{|\chi|^\alpha}(\mathcal{M}(X)\|\mathcal{M}(X'))\Big)^{1/\alpha} \leq \xi(\alpha).$$

Again, $\xi(\alpha)$ is a function from $\mathbb{R}^+$ to $\mathbb{R}^+$.

As we described earlier, this notion of privacy shares many features of RDP and could have independent interest. It subsumes $(\epsilon, 0)$-DP (for $\alpha \to \infty$) and implies an entire family of $(\epsilon(\delta), \delta)$-DP through Markov's inequality. We provide additional details on this point in Appendix D.

[9]We think of p, q, r as the distributions $\mathcal{M} \circ \mathsf{subsample}(X), \mathcal{M} \circ \mathsf{subsample}(X'), \mathcal{M} \circ \mathsf{subsample}(X'')$, respectively, for mutually adjacent datasets X, X', X''.

For our ternary-$|\chi|^\alpha$-differential privacy, what makes it stand out relative to Rényi DP is how it allows privacy amplification to occur in an extremely clean fashion, as the following proposition states. The proof of the proposition involves conditioning on subsampling events, constructing dummy random variables to match up each of these events, and the use of Jensen's inequality to convert the hard to work with ternary-$|\chi|^\alpha$-DP over three mixture distributions to that of three simple distributions that come from mutually adjacent datasets. The following simple lemma will be helpful.

Lemma 13 . *Bivariate function $f(x,y) = x^j/y^{j-1}$ is jointly convex on $\mathbb{R}_+^2$ for $j > 1$.*

Proof. Note that the function is continuously differentiable on $\mathbb{R}_+^2$. The two eigenvalues of the Hessian matrix

$$0 \quad \text{and} \quad (j^2-j)\frac{x^j}{y^{j+1}}\left(1+\frac{y^2}{x^2}\right)$$

and both are nonnegative in the first quadrant. □

Proposition 14 Subsampling Lemma for Ternary-$|\chi|^\alpha$-DP**.** *Let a mechanism $\mathcal{M}$ obey ζ-ternary-$|\chi|^\alpha$-DP, then the algorithm $\mathcal{M} \circ \mathsf{subsample}$ obeys $\gamma\zeta$-ternary-$|\chi|^\alpha$-DP.*

Proof. If three datasets X, X', X'' of size n are mutually adjacent, they must differ on the same data point (w.l.o.g., let it be the nth), and the remaining $n-1$ data points are the same. Let p, q, r denote the distributions $\mathcal{M}\circ\mathsf{subsample}(X), \mathcal{M}\circ\mathsf{subsample}(X'), \mathcal{M}\circ\mathsf{subsample}(X'')$, respectively.

Let E be the event such that the subsample includes the nth item and E^c be complement event. We have

$$p = \gamma p(\cdot|E) + (1-\gamma)p(\cdot|E^c)$$
$$q = \gamma q(\cdot|E) + (1-\gamma)q(\cdot|E^c).$$

and by construction, $p(\cdot|E^c) = q(\cdot|E^c)$.

Substituting the observation into the ternary-$|\chi|^j$-divergence, we get γ^j to show up:

$$\begin{aligned} D_{|\chi|^j}(p,q\|r) &= \mathbb{E}_r\left[\left(\frac{|p-q|}{r}\right)^j\right] = \gamma^j\mathbb{E}_r\left[\left(\frac{|p(\cdot|E)-q(\cdot|E)|}{r}\right)^j\right] \\ &= \gamma^j D_{|\chi|^j}(p(\cdot|E), q(\cdot|E)\|r). \end{aligned} \tag{3.4}$$

Note that $p(\cdot|E), q(\cdot|E)$ and r are mixture distributions with combinatorially (exponential in n) many mixing components.

Let J be a random subset of size γn chosen by the $\mathsf{subsample}$ operator. In addition, we define an auxiliary dummy variable $i \sim \mathrm{Unif}(1, ..., \gamma n)$. Let i be independent of everything else, so it is clear that $r(\theta|J) = r(\theta|J,i)$. In other words,

$$r(\theta) = \mathbb{E}_{J,i}[r(\theta|J,i)] = \frac{1}{\gamma n\binom{n}{\gamma n}}\sum_{J\subset[n], i\in[\gamma n]} r(\theta|J).$$

Now, define functions g and g' on index set J, i such that:

$$g(J,i) = \begin{cases} p(\theta|J) & \text{if } n \in J \\ p(\theta|J \cup \{n\} \backslash J[i]) & \text{otherwise,} \end{cases} \quad g'(J,i) = \begin{cases} q(\theta|J) & \text{if } n \in J \\ q(\theta|J \cup \{n\} \backslash J[i]) & \text{otherwise.} \end{cases}$$

Check that $p(\theta|E) = \mathbb{E}_{J,i} g(J,i)$ and $q(\theta|E) = \mathbb{E}_{J,i} g'(J,i)$.

The above definitions and the introduction of the dummy random variable i may seem mysterious, but they are critical to the proof. Let us explain the rationale behind them. The dummy random variable i and the way we define $g(J,i)$ and $g'(J,i)$ help to create a coupling among the mixture components in the three mixture distributions $p(\theta|E)$, $q(\theta|E)$ and $q(\theta)$. Note that distributions $p(\theta|E), q(\theta|E)$ have a different number of mixture components compared to $q(\theta)$. In fact, $q(\theta)$ has $\binom{n}{\gamma n}$ components while $p(\theta|E)$ and $q(\theta|E)$ only have $\binom{n-1}{\gamma n-1}$ components due to the conditioning on the event E that fixes the differing (say the nth) datapoint in the sampled set.

The dummy random variable i allows us to define a new sigma-field to redundantly represent both subsampling over $[n-1]$ and $[n]$ under the same uniform probability measure over the random subset $J \subset [n]$ of size γn while establishing a one-to-one mapping between pairs of events such that the corresponding index of the subsample differs by only one datapoint.

This trick allows us to write:

$$\begin{aligned} \mathbb{E}_q \left(\frac{|p(\theta|E) - q(\theta|E)|}{q(\theta)} \right)^j &= \int \frac{[p(\theta|E) - q(\theta|E)]^j}{q(\theta)^{j-1}} d\theta \\ &\underset{\text{Jensen}}{\uparrow} \leq \int \mathbb{E}_{J,i} \left[\frac{|g(J,i) - g'(J,i)|^j}{q(\theta|J)^{j-1}} \right] d\theta \\ &\underset{\text{Fubini}}{\uparrow} = \mathbb{E}_{J,i} \mathbb{E}_q \left[\left(\frac{|g(J,i) - g'(J,i)|}{q(\theta|J)} \right)^j \,\middle|\, J, i \right] \leq \zeta(j)^j. \end{aligned} \tag{3.5}$$

The second but last line uses Jensen's inequality and Lemma 13, which proves the joint convexity of the function $x^j/y(j-1)$ on $\mathbb{R}^2_+$. In the last line, we exchange the order of the integral, from which we get the expression for the ternary-DP directly. By definition of ternary-DP (Definition 11), we get the $\zeta(j)^j$ bound.

Combining (3.4) with (3.5) gives the claimed result because the definitions of g and g' ensure that each inner expectation is a ternary Pearson-Vajda divergence of the original mechanism on a triple of mutually adjacent datasets. □

Part 2: Bounding RDP with Ternary-$|\chi|^\alpha$-DP. We will now show that (a transformation of) the quantity of interest — RDP of the subsampled mechanism — can be expressed as a linear combination of a sequence of binary-$|\chi|^\alpha$-DP parameters $\xi(\alpha)$ for integer $\alpha = 2, 3, ...$ through Newton's series expansion of the moment generating function:

$$\mathbb{E}_q \left[\left(\frac{p}{q} \right)^\alpha \right] = 1 + \binom{\alpha}{1} \mathbb{E}_q \left[\frac{p}{q} - 1 \right] + \sum_{j=2}^{\alpha} \binom{\alpha}{j} \mathbb{E}_q \left[\left(\frac{p}{q} - 1 \right)^j \right]. \tag{3.6}$$

Observe that $\mathbb{E}_q \left[\frac{p}{q} - 1 \right] = 0$, so it suffices to bound $\mathbb{E}_q \left[\left(\frac{p}{q} - 1 \right)^j \right]$ for $j \geq 2$.

Note that $\frac{p}{q} - 1$ is a special case of $(p-q)/r$ with $q = r$, therefore,

$$\max_{p,q} \mathbb{E}_q\left[\left(\frac{p-q}{q}\right)^j\right] \le \max_{p,q,r} \mathbb{E}_r\left[\left(\frac{p-q}{r}\right)^j\right] \le \max_{p,q,r} D_{|\chi|^j}(p,q\|r).$$

The same holds if we write $\mathcal{M}' = \mathcal{M} \circ \mathsf{subsample}$ and restrict the maximum on the left to $p = \mathcal{M}'(X)$ and $q = \mathcal{M}'(X')$ with X, X' adjacent, and the maximum on the right to $p = \mathcal{M}'(X)$, $q = \mathcal{M}'(X')$ and $r = \mathcal{M}'(X'')$ with mutually adjacent X, X' and X''. For the subsampled mechanism, the right-hand side of the above equation can be bounded by Proposition 14. Putting these together, we can bound (3.6) as

$$\mathbb{E}_q\left[\left(\frac{p}{q}\right)^\alpha\right] \le 1 + \sum_{j=2}^{\alpha} \binom{\alpha}{j} \gamma^j \zeta(j)^j,$$

where mechanism $\mathcal{M}$ satisfies ζ-ternary-$|\chi|^\alpha$-DP and p, q denote the distributions $\mathcal{M} \circ \mathsf{subsample}(X), \mathcal{M} \circ \mathsf{subsample}(X')$, respectively, for adjacent datasets X, X'. Using this result along with the definition of Rényi differential privacy (from Definition 4) implies the RDP parameter following bound,

$$\epsilon_{\mathcal{M}\circ\mathsf{subsample}}(\alpha) \le \frac{1}{\alpha - 1} \log\left(1 + \sum_{j=2}^{\alpha} \binom{\alpha}{j} \gamma^j \zeta(j)^j\right). \tag{3.7}$$

Part 3: Bounding Ternary-$|\chi|^\alpha$-DP using RDP. Considering (3.7), it remains to bound

$$\zeta(j)^j := \sup_{p,q,r} \mathbb{E}_r\left[\frac{|p-q|^j}{r^j}\right]$$

using RDP. We provide several ways of doing so and plugging them into (3.7) shows how the various terms in the bound of Theorem 9 arise.

(a) The $4(e^{\epsilon(2)} - 1)$ Term. To begin with, we show that the binary-$|\chi|^\alpha$-DP and ternary-$|\chi|^\alpha$-DP are equivalent up to a constant of 4.

Lemma 15 . *If a randomized mechanism $\mathcal{M}$ is ξ-binary-$|\chi|^\alpha$-DP, then it is ζ-ternary-$|\chi|^\alpha$-DP for some ζ satisfying $\xi(\alpha)^\alpha \le \zeta(\alpha)^\alpha \le 4\xi(\alpha)^\alpha$.*

Proof. The first inequality follows trivially by definition. We now prove the second. Let p, q, r be three probability distributions over the same support. Consider the following four sets:

$$E_1 = \{\theta \,|\, p(\theta) \ge q(\theta), q(\theta) \ge r(\theta)\}$$
$$E_2 = \{\theta \,|\, p(\theta) \ge q(\theta), q(\theta) < r(\theta)\},$$
$$E_3 = \{\theta \,|\, p(\theta) < q(\theta), p(\theta) \ge r(\theta)\},$$
$$E_4 = \{\theta \,|\, p(\theta) < q(\theta), p(\theta) \ge r(\theta)\}.$$

Let $p = p(\theta), q = q(\theta), r = r(\theta)$. For $\theta \in E_1$, $|p-q|^j/r^{j-1} = (p-q)^j/r^{j-1} \le (p-r)^j/r^{j-1}$. For $\theta \in E_2$, $|p-q|^j/r^{j-1} \le (p-q)^j/q^{j-1}$. Similarly, for $\theta \in E_3$ and $\theta \in E_4$, $|p-q|^j/r^{j-1}$

is bounded by $(q-r)^j/r^{j-1}$ and $(q-p)^j/p^{j-1}$ respectively. It then follows that,

$$\begin{aligned}
&\mathbb{E}_{\theta\sim r}[|p-q|^j/r^j] \\
=&\mathbb{E}_{\theta\sim r}[|p-q|^j/r^j\mathbf{1}_{\theta\in E_1}] + \mathbb{E}_{\theta\sim r}[|p-q|^j/r^j\mathbf{1}_{\theta\in E_2}] + \mathbb{E}_{\theta\sim r}[|p-q|^j/r^j\mathbf{1}_{\theta\in E_3}] + \mathbb{E}_{\theta\sim r}[|p-q|^j/r^j\mathbf{1}_{\theta\in E_4}] \\
\leq&\mathbb{E}_{\theta\sim r}[|p-r|^j/r^j\mathbf{1}_{\theta\in E_1}] + \mathbb{E}_{\theta\sim q}[|p-q|^j/q^j\mathbf{1}_{\theta\in E_2}] + \mathbb{E}_{\theta\sim r}[|q-r|^j/r^j\mathbf{1}_{\theta\in E_3}] + \mathbb{E}_{\theta\sim p}[|q-p|^j/p^j\mathbf{1}_{\theta\in E_4}] \\
\leq&D_{|\chi|^j}(p\|r) + D_{|\chi|^j}(p\|q) + D_{|\chi|^j}(q\|r) + D_{|\chi|^j}(q\|p) \leq 4\xi(j).
\end{aligned}$$

□

In Lemma 15, for the special case of $\alpha = 2$, we have

$$\mathbb{E}_q[|p/q-1|^2] = \mathbb{E}_q[(p/q)^2] - 2\mathbb{E}_q[p/q] + 1 = e^{\epsilon(2)} - 1.$$

Using the bound from Lemma 15 relating the binary and ternary-$|\chi|^\alpha$-DP, gives that $\zeta(2) \leq 4(e^{\epsilon(2)}-1)$.

(b) The $e^{(j-1)\epsilon(j)}\min\{2, (e^{\epsilon(\infty)}-1)^j\}$ Term. Now, we provide a bound for $j \geq 2$. We start with the following simple lemma.

Lemma 16 . *Let X, Y be nonnegative random variables, for any $j \geq 1$,*

$$\mathbb{E}[|X-Y|^j] \leq \mathbb{E}[X^j] + \mathbb{E}[Y^j].$$

Proof. Using that the $X, Y \geq 0$

$$\begin{aligned}
\mathbb{E}[|X-Y|^j] &= \mathbb{E}[(X-Y)^j\mathbf{1}(X \geq Y)] + \mathbb{E}[(X-Y)^j\mathbf{1}(X < Y)] \\
&\leq \mathbb{E}[X^j \cdot \mathbf{1}(X \geq Y)] + \mathbb{E}\left[Y^j \cdot \mathbf{1}(X < Y)\right] \leq \mathbb{E}[X^j] + \mathbb{E}[Y^j]
\end{aligned}$$

□

This "triangular inequality"-like result exploits the nonnegativity of X, Y and captures the intrinsic cancellations of the 2^j terms of a Binomial expansion. If we do not have non-negativity, the standard expansion will provide a 2^j factor rather than 2 in the $\min\{2, (e^{\epsilon(\infty)}-1)^j\}$ term (see e.g., Proposition 3.2 of Bobkov et al. (2019)).
Next we show an alternative bound that is tighter in cases when X and Y are related to each other with a multiplicative bound. Note that this bound is only going to be useful when $\mathcal{M}$ has a bounded $\epsilon(\infty)$; i.e. when $\mathcal{M}$ satisfies $(\epsilon, 0)$-DP guarantee.

Lemma 17 . *Let X, Y be nonnegative random variables and with probability 1, $e^{-\varepsilon}Y \leq X \leq e^{\varepsilon}Y$. Then for any $j \geq 1$,*

$$\mathbb{E}[|X-Y|^j] \leq \mathbb{E}[Y^j](e^{\varepsilon}-1)^j.$$

Proof. The multiplicative bound implies that: $-Y(1-e^{-\varepsilon}) \leq X - Y \leq Y(e^{\varepsilon}-1)$, which gives that with probability 1

$$|X-Y| \leq \max\{e^{\varepsilon}-1, 1-e^{-\varepsilon}\}Y = (e^{\varepsilon}-1)Y,$$

and the claimed result follows. □

Take $X = p/r$ and $Y = q/r$. Applying Lemma 16 gives $\zeta(j) \le 2e^{(j-1)\epsilon(j)}$. Using Lemma 17 instead with $\varepsilon = \epsilon(\infty)$ provided by the mechanism $\mathcal{M}$, we have $\zeta(j) \le e^{(j-1)\epsilon(j)}(e^{\epsilon(\infty)}-1)^j$. Using these bounds together, we get the overall bound

$$\zeta(j) \le e^{(j-1)\epsilon(j)} \min\{2, (e^{\epsilon(\infty)}-1)^j\}.$$

Note that at $j = 2$, $e^{(j-1)\epsilon(j)} \min\{2, (e^{\epsilon(\infty)}-1)^j\}$ simplifies to $e^{\epsilon(2)} \min\{2, (e^{\epsilon(\infty)}-1)^2\}$.

Plugging in the above derivations in (3.7) yields the bound stated in Theorem 9.

3.2. **Improving the Bound in Theorem 9.** We note that we can improve the bound in Theorem 9 under some additional assumptions on the RDP guarantee. We formalize this idea in this section. We use $d(X, X') \le 1$ to represent neighboring datasets. We start with some additional conditions on the mechanism $\mathcal{M}$ as defined below.

Definition 18 Tightness and Self-consistency**.** *We say a mechanism $\mathcal{M}$ and its corresponding RDP privacy guarantee $\epsilon_{\mathcal{M}}(\cdot)$ are* tight *if $\max_{X,X':d(X,X')\le 1} D_\ell(\mathcal{M}(X)\|\mathcal{M}(X')) = \epsilon_{\mathcal{M}}(\ell)$ for every $\ell = 1, 2, 3, ...$ We say that a tight pair $(\mathcal{M}, \epsilon_{\mathcal{M}}(\cdot))$ is* self-consistent *with respect to $|\chi|^\alpha$-divergence, if*

$$\Big(\cap_{\ell=1,2,...,\alpha} \operatorname*{argmax}_{X,X':d(X,X')\le 1} D_\ell(\mathcal{M}(X)\|\mathcal{M}(X'))\Big) \cap \operatorname*{argmax}_{X,X':d(X,X')\le 1} D_{|\chi|^\alpha}(\mathcal{M}(X)\|\mathcal{M}(X')) \neq \emptyset.$$

The tightness condition requires that the RDP function $\epsilon_{\mathcal{M}}(\cdot)$ be attainable by two distributions induced by a pair of adjacent datasets and the self-consistency condition requires that *the same* pair of distributions attains the maximal $|\chi|^\alpha$-divergence for a given range of parameters. Self-consistency is a non-trivial condition in general but it is true in most popular cases such as the Gaussian mechanism, Laplace mechanism, etc., where we know the Rényi divergence analytically and the difference of two datasets are characterized by one numerical number, e.g., sensitivity. (See Appendix C for a discussion.)

Define,

$$B(\epsilon, l) := \Delta^{(l)}\left[e^{(\cdot-1)\epsilon(\cdot)}\right](0) = \sum_{i=0}^{l}(-1)^i\binom{l}{i}e^{(i-1)\epsilon(i)},$$

as the lth order forward finite difference (see (3.3)) of the functional $e^{(\cdot-1)\epsilon(\cdot)}$ evaluated at 0.

Theorem 19 Tighter RDP Parameter Bounds**.** *Given a dataset of n points drawn from a domain $\mathcal{X}$ and a (randomized) mechanism $\mathcal{M}$ that takes an input from $\mathcal{X}^m$ for $m \le n$, let $\mathcal{M}\circ$subsample be the randomized mechanism defined in Theorem 9. If $\mathcal{M}$ obeys $(\alpha, \epsilon(\alpha))$-RDP and additionally the RDP guarantee is tight and $(\alpha+1)$-self-consistent as per Definition 18, then for all integer $\alpha \ge 2$, the mechanism $\mathcal{M} \circ$ subsample obeys $(\alpha, \epsilon'(\alpha))$-RDP with*

$$\epsilon'(\alpha) \le \frac{1}{\alpha-1}\log\Bigg(1 + \gamma^2\binom{\alpha}{2}\min\Big\{4(e^{\epsilon(2)}-1), e^{\epsilon(2)}\min\{2,(e^{\epsilon(\infty)}-1)^2\}\Big\} + 4\sum_{j=3}^{\alpha}\gamma^j\binom{\alpha}{j}\sqrt{B(\epsilon, 2\lfloor j/2\rfloor))\cdot B(\epsilon, 2\lceil j/2\rceil)}\Bigg).$$

Proof. The proof is identical to that of Theorem 9. The part where it differs is in Part 3, i.e., bounding $\zeta(j)^j$ using RDP. As a result of the assumptions in Definition 18, we know that there exist a pair of adjacent datasets, which give rise to a pair of distribution p and q, that simultaneously achieves the upper bound in the definition of both $\xi(j)$ and $\epsilon(j)$ divergences for all j of interest. For even j, the χ^j-divergence can be written in an analytical form as a Rényi divergence (Nielsen and Nock, 2014) using a binomial expansion. Using Lemma 15 along with this expansion, gives rise to the $4\Delta^{(j)}[e^{(\cdot-1)\epsilon(\cdot)}](0) = 4B(\epsilon, j)$ bound for even j. For odd j, we reduce it to the even j case through the Cauchy-Schwartz inequality

$$\mathbb{E}_q[|p/q-1|^j] = \mathbb{E}_q[|p/q-1|^{(j-1)/2}|p/q-1|^{(j+1)/2}] \leq \sqrt{\mathbb{E}_q[(p/q-1)^{j-1}]\mathbb{E}_q[(p/q-1)^{j+1}]},$$

where each of the terms in the square root can now be bounded by the binomial expansion. Putting these together, one notices that one can replace $e^{(j-1)\epsilon(j)}\min\{2, (e^{\epsilon(\infty)}-1)^j\}$ with a more exact evaluation given by $4\sqrt{B(\epsilon, 2\lfloor j/2\rfloor))\cdot B(\epsilon, 2\lceil j/2\rceil)}$ in the bound of Theorem 9. We use this bound only for $j \geq 3$ because for $j = 2$, as discussed in the paragraph after Lemma 15, we have an easy alternative way of bounding $\zeta(2)$ that does not require these additional assumptions. □

When applicable, Theorem 19 could be substantially stronger than Theorem 9 [10]. The only difference is in the trailing terms for $j \geq 3$. Specifically, the $\gamma^j\binom{\alpha}{j}e^{(j-1)\epsilon(j)}\max\{2, (e^{\epsilon(\infty)}-1)^j\}$ term in Theorem 9 only goes to zero as j gets larger when γ or $\epsilon(\infty)$ are very small. On the other hand, Theorem 19, allows the bound to benefit from the cancellations of positive and negative terms when $B(\epsilon, l)$ is evaluated, rather than naively upper bounding the negative terms by 0, hence allowing a more precise calculation of the Ternary Pearson-Vajda divergences in Equation 3.7.

3.3. **A Lower Bound on the RDP for Subsampled Mechanisms.** We now discuss whether our bound in Theorem 9 can be improved. First, we provide a short answer: it cannot be improved in general. This is demonstrated in the following proposition.

Proposition 20 . *Let $\mathcal{M}$ be a randomized algorithm that takes a dataset in $\mathcal{X}^{\gamma n}$ as an input. If $\mathcal{M}$ obeys $(\alpha, \epsilon(\alpha))$-RDP for a function $\epsilon : \mathbb{R}_+ \to \mathbb{R}_+$ and that there exists $x, x' \in \mathcal{X}$ such that $\epsilon(\alpha) = D_\alpha\big(\mathcal{M}([x, x, ..., x, x'])\|\mathcal{M}([x, x, ..., x, x])\big)$ for all integer $\alpha \geq 1$ (e.g., this condition is true for all output perturbation mechanisms for counting queries), then the RDP function ϵ' for $\mathcal{M} \circ \mathsf{subsample}$ obeys the following lower bound for all integers $\alpha \geq 1$:*

$$\epsilon'(\alpha) \geq \frac{\alpha}{\alpha-1}\log(1-\gamma) + \frac{1}{\alpha-1}\log\left(1 + \alpha\frac{\gamma}{1-\gamma} + \sum_{j=2}^{\alpha}\binom{\alpha}{j}\left(\frac{\gamma}{1-\gamma}\right)^j e^{(j-1)\epsilon(j)}\right).$$

Proof. Consider two datasets $X, X' \in \mathcal{X}^n$ where X' contains n data points that are identically x and X is different from X' only in its last data point. By construction, $\mathsf{subsample}(X') \equiv [x, x, ..., x]$, $\Pr[\mathsf{subsample}(X) = [x, x, ..., x]] = 1-\gamma$ and $\Pr[\mathsf{subsample}(X) = [x, x, ..., x, x'] =$

[10] It may not be strictly stronger due to the factor of 4 that comes from the conversion from ternary to binary versions of Pearson-Vajda divergences.

γ. In other words, $\mathcal{M} \circ \mathsf{subsample}(X') = \mathcal{M}([x, x, ..., x]) := p$ and $\mathcal{M} \circ \mathsf{subsample}(X) = (1-\gamma)p + \gamma\mathcal{M}([x, x, ..., x, x']) := (1-\gamma)p + \gamma q$. It follows that

$$\begin{aligned}\mathbb{E}_q\left[\left(\frac{(1-\gamma)q+\gamma p}{q}\right)^\alpha\right] =& \mathbb{E}_q\left[\left(1-\gamma+\gamma\frac{p}{q}\right)^\alpha\right] = (1-\gamma)^\alpha \mathbb{E}_q\left[\left(1+\frac{\gamma}{1-\gamma}\frac{p}{q}\right)^\alpha\right] \\ =& (1-\gamma)^\alpha\left(1+\alpha\frac{\gamma}{1-\gamma}+\sum_{j=2}^{\alpha}\binom{\alpha}{j}\left(\frac{\gamma}{1-\gamma}\right)^j \mathbb{E}_q\left[\left(\frac{p}{q}\right)^j\right]\right).\end{aligned}$$

When we take x, x' to be the one in the assumption that attains the RDP $\epsilon(\cdot)$ upper bound, then by RDP definition (see Definition 4) we can replace $\mathbb{E}_q\left[(p/q)^j\right]$ in the above bound with $e^{(j-1)\epsilon(j)}$ as claimed. □

Let us compare the above lower bound to our upper bound in Theorem 9 in two regimes. When $\alpha\gamma e^{\epsilon(\alpha)} \ll 1$, such that $\alpha^2\gamma^2 e^{\epsilon(2)} < 1$ is the dominating factor in the summation, we can use the bounds $x/(1+x) \le \log(1+x) \le x$ to get that both the upper and lower bound are $\Theta(\alpha\gamma^2 e^{\epsilon(2)})$. In other words, they match up to a constant multiplicative factor. For other parameter configurations, note that $\gamma/(1-\gamma) > \gamma$, our bound in Theorem 9 (with the $2e^{(j-1)\epsilon(j)}$) is tight up to an additive factor $\frac{\alpha}{\alpha-1}\log((1-\gamma)^{-1}) + \frac{\log(2)}{\alpha-1}$ which goes to 0 as $\gamma \to 0$ and $\alpha \to \infty$. We provide explicit comparisons of the upper and lower bounds in the numerical experiments presented in Section 5.

The longer answer to this question of optimality is more intricate. The RDP bound can be substantially improved when we consider more fine-grained per-instance RDP in the same flavor as the per-instance (ϵ, δ)-DP (Wang, 2019). The only difference from the standard RDP is that now ϵ is parameterized by a pair of fixed adjacent datasets. This point is illustrated below, where we discuss an asymptotic approximation of the Rényi divergence for the subsampled Gaussian mechanism.

3.4. **Asymptotic Approximation of Rényi Divergence for Subsampled Gaussian Mechanism.** We now focus on an asymptotic upper bound on the Rényi divergence for the subsampled Gaussian mechanism. The results from this section are also used in our numerical experiments detailed in Section 5.

Let $\mathcal{X}$ denote the input domain. Let $f : \mathcal{X} \to \Theta$ be some statistical query. We consider a subsampled Gaussian mechanism which releases the answers to f by adding Gaussian noise to the mean of a subsampled dataset. In this case, the output θ of the subsampled Gaussian mechanism is a sample from $\mathcal{N}(\mu_J, \sigma^2/|J|^2)$ where μ_J is short for $\mu(X_J) := \frac{1}{|J|}\sum_{i\in J} f(x_i)$ and J is a random subset of size γn. The distribution of J induces a discrete prior distribution of μ_J. Under certain regularity conditions on $f(x_i)$ for $i = 1, ..., n$, as $n \to \infty$ and $|J| = \gamma n$ with $0 < \gamma < 1$, then the sampling without replacement version of the central limit theorem (CLT) (see, e.g., Madow, 1948) implies that $\sqrt{|J|}(\mu(X_J) - \frac{1}{n}\sum_{i=1}^n f(x_i))$ converges in distribution to $\mathcal{N}(0, \frac{1-\gamma}{n}\sum_{i=1}^n (f(x_i) - \mu(X))^2)$. It follows from the independence of J and the added noise that we can approximate the complex mixture distribution of the output θ

with its asymptotic distribution[11] of

$$\mathcal{N}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i), \frac{1-\gamma}{n|J|}\sum_{i=1}^{n}(f(x_i)-\mu(X))^2 + \frac{\sigma^2}{|J|^2}\right).$$

This allows us to use the analytical formula of the Rényi divergence between two Gaussians (see Appendix F) as an asymptotic approximation of the Rényi divergence between the more complex mixture distributions.

A disclaimer is that this is a truly asymptotic approximation and would only be true when $|J|, n \to \infty$ and the (sequences of) "population" dataset $\{f(x_i)|i \in [n]\}$ satisfies the regularity conditions of the sample-without-replacement CLT as n gets large. Moreover, convergence in distribution in general does not imply convergence in higher moments, thus the approximation of the Rényi divergence using that of the limiting distribution may only be a *good* approximation for small Rényi divergence orders. This approximation is nevertheless interesting as it allows us to understand the dependence of different parameters in the bound with a simple data-dependent formula.

One important observation is that the part of the variance due to the dataset can be either bigger or smaller than that of the added noise, and this could imply a vastly different Rényi divergence. We give examples here of two contrasting situations. Let $\alpha^\star$ be defined through $\gamma\alpha^* e^{\epsilon_{\mathcal{M}}(\alpha^*)} = 1$ where $\mathcal{M}$ is the Gaussian mechanism. Without loss of generality, we assume that $f(x_i) \le 1/2$, which implies that the global sensitivity of μ is $1/|J|$.

Example 21 Gaussian approximation - a "bad" data case**.** *Let $f(x_1) = f(x_2) = ... = f(x_{n-1}) = f(x_n) = -1/2$ for the elements in X', and for X the only difference (from X') is that in X we have $f(x_n) = 1/2$. Then the two asymptotic distributions are $p = \mathcal{N}(-\frac{1}{2}+\frac{1}{n}, \frac{(1-\gamma)(n-1)}{n^2|J|} + \frac{\sigma^2}{|J|^2})$ and $q = \mathcal{N}(-\frac{1}{2}, \frac{\sigma^2}{|J|^2})$, and the corresponding Rényi divergence equals*

$$D_\alpha(p\|q) = \begin{cases} +\infty & \text{if } \alpha \ge \alpha^* := \frac{\sigma^2}{\gamma(1-\gamma)}\frac{n}{n-1} + 1, \\ \frac{\alpha\gamma^2}{2\sigma^2}\left(\frac{\alpha^*-1}{\alpha^*-\alpha}\right) + \frac{1}{2}\log\left(\frac{\alpha^*-1}{\alpha^*}\right) + \frac{1}{2(\alpha-1)}\log(\frac{\alpha^*-1}{\alpha^*-\alpha}) & \text{otherwise.} \end{cases}$$

Example 22 Gaussian approximation - a "good" data case**.** *Let n be an odd number, and let X' be such that $f(x_i) = 1/2$ for $i \le \lfloor n/2 \rfloor$ and $f(x_i) = -1/2$ otherwise, and for X the only difference (from X') is that in X we have $f(x_n) = 1/2$. The two asymptotic distributions are $p = \mathcal{N}(\frac{1}{2n}, \frac{\sigma^2}{|J|^2} + \frac{1-\gamma}{4|J|} - \frac{1-\gamma}{4n^2|J|})$ and $q = \mathcal{N}(-\frac{1}{2n}, \frac{\sigma^2}{|J|^2} + \frac{1-\gamma}{4|J|} - \frac{1-\gamma}{4n^2|J|})$, and the corresponding Rényi divergence equals*

$$D_\alpha(p\|q) = \frac{\alpha\gamma^2}{2\sigma^2 + \gamma(1-\gamma)(n-n^{-1})/2}.$$

The first example (a "bad" data case) is closely related to our construction in the proof of Proposition 20. For $\alpha \ll \sigma^2/\gamma$, the example shows that the Rényi divergence of order

[11] Technically speaking, we can only talk about the limiting distribution of a de-meaned and appropriately scaled version of θ, e.g., $\sqrt{|J|}(\theta - \frac{1}{n}\sum_{i=1}^{n} f(x_i))$. We use the stated distribution (that is not well-defined at the limit) for an informal discussion that avoids specifying the scaling factor.

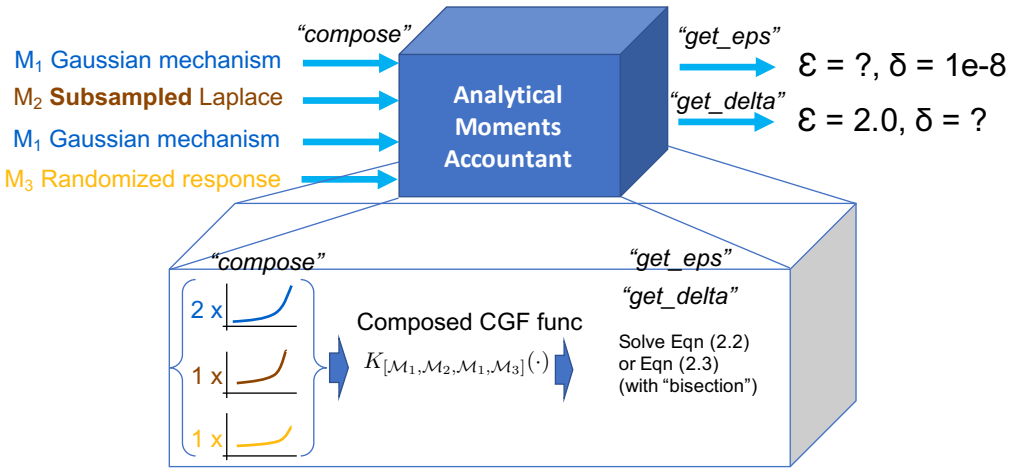


FIGURE 1. Illustration of the user-interface and inner workings of the analytical moments accountant.

α between the outputs on the two datasets has an $O(\alpha\gamma^2/\sigma^2)$ rate, matching our upper bound from Theorem 9 (see the "Bound under Additional Assumptions" remark following Theorem 9) in the small α, large σ regime. The second example corresponds to a "good" data case where the dataset has a variety of different data points, and as we can see from the expression of the asymptotic distribution, the variance that comes from subsampling the dataset dominates the noise from Gaussian mechanism and the per-instance RDP loss for this particular pair of X and X' can be γn times smaller than the bad case.

4. Analytical Moments Accountant

Our theoretical results above allow us to build an analytical moments accountant data structure for composing differentially private mechanisms. The data structure tracks the CGF function $K_{\mathcal{M}}(\cdot)$ of a (potentially adaptive) sequence of mechanisms $\mathcal{M}$ in symbolic form (or as an evaluation oracle). It supports subsampling before applying $\mathcal{M}$ and the $K_{\mathcal{M}}(\cdot)$ will be adjusted accordingly using the RDP amplification bound in Theorem 9. The data structure allows data analysts to query the smallest ϵ from a given δ (or vice versa) for (ϵ, δ)-DP using (2.2) (or (2.3)). An illustration of the user-interface of an analytical moments accountant is given in Figure 1. The more detailed design of its inner workings are given in pseudocodes in Algorithms 1, 4, 5. In particular, the "analytical moments accountant" is designed to compose a sequence of "mechanisms" — described in Algorithm 2 symbolically in terms of their CGF functions. The privacy amplification by subsampling is handled as a transformation from one "mechanism" to another, as described in Algorithms 2 and 3 (also, see the illustration in Figure 2).

A working prototype of the analytical moments accountant is released as an open-source code package "autodp" in `https://github.com/yuxiangw/autodp`.

Practically, our analytical moments accountant is better than the moment accountants proposed by Abadi et al. (2016) in several noteworthy ways: (1) our approach allows one to keep track the CGF's of all $\lambda \geq 1$ in symbolic form without paying infinite memory, whereas moments account (Abadi et al., 2016) requires a predefined list of λ's and pays a memory

Algorithm 1 Analytical Moments Accountant (AMA) class

Attributes:

- "Set of mechanisms" MechSet: a set of mechanisms that the AMA has tracked so far
- "Coefficient vector" coeffs: a dictionary ("Mechanism" $\mathcal{M}$, "Coefficient" $c \geq 0$) of coefficients counting the number of times each mechanism in MechSet has been composed.

Private Methods:

- "CGF" cgf: A function that takes parameter $\lambda > 0$, outputs the CGF of the composed mechanisms: $\sum_{\mathcal{M} \in \text{MechSet}} \text{coeffs}[\mathcal{M}] K_{\mathcal{M}}(\lambda)$.

Public Methods:

- "Compose mechanism" compose: Takes a mechanism $\mathcal{M}$ and a coefficient c as inputs and updates the AMA attributes.
- "Query ϵ" get_eps: Takes $\delta > 0$ as an input, calculates the smallest ϵ using (2.2).
- "Query δ" get_delta: Takes $\epsilon > 0$ as an input, calculates the smallest δ using (2.3).

Algorithm 2 Mechanism class

Attributes:

- "Mechanism Name" name: a string describing its name, e.g., "Gaussian mechanism"
- "Parameters" params: a dictionary of the native parameters of this mechanism, e.g., σ for the Gaussian mechanism, b for the Laplace mechanism and so on.
- "Tightness and Self-Consistency Check" flag: a boolean indicating whether the mechanism satisfies Definition 18.

Public Methods:

- "CGF function" cgf: A function that takes parameter $\lambda > 0$, outputs the CGF $K_{\mathcal{M}}(\lambda)$.

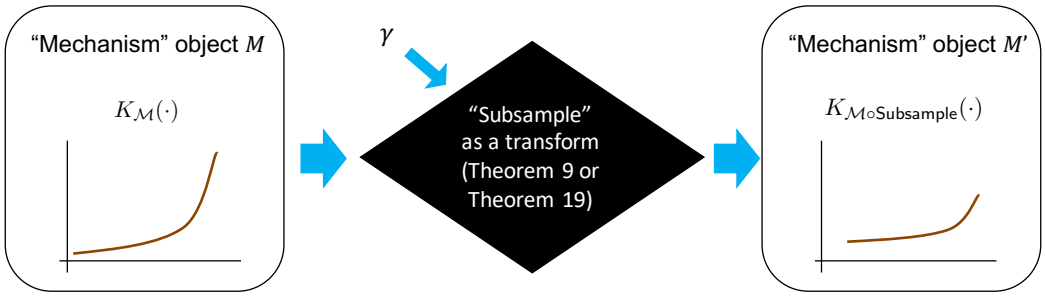


FIGURE 2. Illustration of the "subsample" operation as a transformation of one "Mechanism" object into another.

proportional to the size of the list; (2) our approach completely avoids numerical integration used by moments account; and finally (3) our approach supports subsampling for generic RDP mechanisms while the moments accountant was built for supporting only Gaussian mechanisms. All of this translates into an efficient and accurate way for tracking ϵ's and δ's when composing differentially private mechanisms.

We design the data structure to be numerically stable, and efficient in both space and time. In particular, it tracks CGFs and takes $O(1)$ time to compose a new mechanism and uses space only linear in the number of *unique* mechanisms applied (rather than the number of total mechanisms applied). Using the convexity of CGFs and the monotonicity of RDP, we

Algorithm 3 "subsample" function

Input: Mechanism object $\mathcal{M}$, sampling ratio γ.
Output: Mechanism object $\mathcal{M}'$ representing $\mathcal{M} \circ \textbf{Subsample}_\gamma$.

Declare Mechanism object $\mathcal{M}'$
$\mathcal{M}'$.name $\leftarrow \mathcal{M}$.name + "_Subsample_" + γ.
$\mathcal{M}'$.params $\leftarrow \mathcal{M}$.params + {"gamma" : γ }
if $\mathcal{M}$.flag **then**
  Set $\mathcal{M}'$.cgf according to Theorem 19 and Remark 7.
else
  Set $\mathcal{M}'$.cgf according to Theorem 9 and Remark 7.
end if
$\mathcal{M}'$.flag $\leftarrow$ False
Return $\mathcal{M}'$.

Algorithm 4 "compose" method in Analytical Moments Accountant

Input: Mechanism object $\mathcal{M}$, coefficient $c \geq 0$.

if $\mathcal{M}$ is in this.MechSet **then**
  this.coeffs[$\mathcal{M}$] $\leftarrow$ this.coeffs[$\mathcal{M}$] + c.
else
  Add $\mathcal{M}$ to this.MechSet
  this.coeffs[$\mathcal{M}$] $\leftarrow c$.
end if
Update this.cgf method.

Algorithm 5 "get_eps" method in Analytical Moments Accountant

Input: "delta" $\delta \geq 0$ (represented numerically as $\log(1/\delta)$).
Output: "epsilon" ϵ parameter.

Use doubling trick on λ to find the a window that encloses the optimal solution.
Do bisection on λ to find the argmin of (2.2).

are able to provide $\delta \Rightarrow \epsilon$ conversion to (ϵ, δ)-DP to within accuracy τ in oracle complexity $O(\log(\lambda^*/\tau))$, where λ^* is the optimal value for λ. Similarly, for $\epsilon \Rightarrow \delta$ queries.

Note that for subsampled mechanisms the direct evaluation $\epsilon_{\mathcal{M}\circ\mathsf{subsample}}(\alpha)$ of the upper bounds in Theorem 9 is already polynomial in α. To make the data structure truly scalable, we devise a number of ways to approximate the bounds that then takes only $O(\log(\alpha))$ evaluations of $\epsilon_{\mathcal{M}}(\cdot)$. In the remainder of the section, we will describe all the practically relevant details in the design of this data structure.

Design of Analytical Moments Accountant Data Structure. Let $\mathcal{M}_1, \mathcal{M}_2, .., \mathcal{M}_k$ be a sequence of (possibly adaptively chosen) randomized mechanisms that one applies to the dataset and let $K_{\mathcal{M}_1}, ..., K_{\mathcal{M}_k}$ be the corresponding CGFs. The analytical moments accountant maintains $K = K_{\mathcal{M}_1} + ... + K_{\mathcal{M}_k}$ in symbolic forms and it can evaluate $K(\lambda)$ at any $\lambda > 0$. The two main usages of the analytical moments accountant are for keeping track of: (a) RDP parameter $\epsilon(\alpha)$ for all α, and (b) $(\epsilon(\delta), \delta)$-DP for all $0 \leq \delta < 1$, for

a heterogeneous sequence of adaptively chosen randomized mechanisms. The conversion to RDP is straightforward using the one-to-one relationship between CGF and RDP (see Remark 7) with the exception of RDP at $\alpha = 1$ (Kullback-Leibler privacy) and $\alpha = +\infty$ (pure DP), which are tracked separately. The conversion to (ϵ, δ)-DP is obtained by solving the univariate optimization problems described in (2.2) and (2.3).

Space and Time Complexity for Tracking Mechanisms and for (ϵ, δ)-DP Query. We start by analyzing the space and time complexity for basic operations of this data structure.

Proposition 23 . *The analytical moments accountant takes $O(1)$ time to compose a new mechanism. At any point in time after the analytical moments accountant has been declared and is in operation, let L denote the total number of unique mechanisms that it has seen so far. Then the analytical moments accountant takes $O(L)$ space. Evaluating the CGF (at a given λ) takes $O(L)$ time. Evaluating a (ϵ, δ)-DP query to accuracy within τ (in terms of absolute difference in the argument $|\lambda - \lambda^*|$) takes $O(\log(\lambda^*/\tau))$ CGF evaluation calls, where λ^* is the corresponding minimizer in (2.2) or (2.3).*

Proof. We keep track of a dictionary of functions where the (key,value)-pair is effectively $(\mathcal{M}, (K_{\mathcal{M}}, c_{\mathcal{M}}))$ where $K_{\mathcal{M}}$ is a function that returns the CGF given any positive input, and $c_{\mathcal{M}}$ is the coefficient denoting how many times $\mathcal{M}$ appeared. This naturally allows $O(1)$ time to add a new mechanism and $O(L)$ space.

Since CGFs composes by simply adding up the functions, the overall CGF is $\sum_{i=1}^{L} c_{\mathcal{M}_i} K_{\mathcal{M}_i}$. Evaluating this function takes L CGF queries. We think of the problems of solving for ϵ given δ and solving for δ given ϵ as *zero*th order optimization problem using these queries. These problems are efficiently solvable due to the geometric properties of CGFs that we mention in Section 2 and Appendix E.

When solving for ϵ given δ (from (2.2)), starting with a small positive value $\lambda_{\max}$, we keep doubling the candidate $\lambda_{\max}$ and calculating $\frac{1/\delta + K_{\mathcal{M}}(\lambda_{\max})}{(\lambda_{\max})} - \frac{1/\delta + K_{\mathcal{M}}(\lambda_{\max}-1)}{(\lambda_{\max}-1)}$ until we find that it is positive. This procedure is guaranteed to detect a bounded interval containing λ^* in $O(\log \lambda^*)$ time thanks to the monotonicity of RDP. Then we use the bisection method to find the optimal λ^*, using the unimodal property of the objective function. Note that $\lambda_{\max} \leq 2\lambda^*$. This ensures that the total oracle evaluation complexity to find a τ-optimal solution (i.e., to within accuracy τ) of λ^* is $O(\log(\lambda^*/\tau))$. We can solve for δ given ϵ using the same bisection algorithm with the same time complexity, by using the fact that (2.3) is a log-convex problem. □

The results are compared to a naïve implementation of the standard moments accountant that keeps track of an array of size $\lambda_{\max}$ and handles $\delta \Rightarrow \epsilon$ queries without regarding the geometry of CGFs. The latter will take $O(\lambda_{\max})$ time and space for tracking a new mechanism, and $O(\lambda_{\max})$ time to find a 1-optimal solution. In addition, it does not allow a dynamic choice of $\lambda_{\max}$. The analytical moments accountant described here, despite its simplicity, is an exponential improvement over the naïve version, besides being more flexible and adaptive.

There are still several potential problems. First, the input may be an upper bound to a CGF which is not be an actual CGF function of any random variable, therefore breaking

the computational properties. Secondly, when we need to handle subsampled mechanisms, even just evaluating the RDP bound in Theorem 9 once at α will cost $O(\alpha^2)$ (therefore $O(\lambda^2)$). Lastly, the quantities in the bound of Theorem 9 could be exponentially large and operating on them naïvely might cause floating point overflows or underflows. We address these problems below.

"Projecting" a CGF Upper Bound into a Feasible Set. Note that an upper bound of the CGF does not necessarily have the standard properties associated with CGF that we note in Appendix E, however, we can "project" it to another valid upper bound using the proposition below so that it satisfies the properties from Appendix E.

Proposition 24 . *Let $\bar{K}_{\mathcal{M}}$ be an upper bound of $K_{\mathcal{M}}$. There exists a functional F such that $F[\bar{K}_{\mathcal{M}}] \le K_{\mathcal{M}}$ and $F[\bar{K}_{\mathcal{M}}]$ obeys that $F[\bar{K}_{\mathcal{M}}]$ is convex, monotonically increasing, evaluates to 0 at 0, and $\frac{1}{\lambda}F[\bar{K}_{\mathcal{M}}](\lambda)$ is monotonically increasing on $\lambda \ge 0$.*

Proof. We prove this by constructing such an F explicitly. First define $g := \text{convexhull}(\bar{K}_{\mathcal{M}})$. By definition, g is the pointwise largest convex function that satisfies the given upper bound. Secondly, we find the largest β such that $\beta\lambda \le g(\lambda)$, $\forall\lambda$. Let the smallest λ such that $g(\lambda) = \beta\lambda$ be $\tilde{\lambda}$. Then, we define

$$F[\bar{K}_{\mathcal{M}}](\lambda) = \begin{cases} 0 & \text{when } \lambda \le 0, \\ \beta\lambda & \text{when } 0 < \lambda \le \tilde{\lambda}, \\ g(\lambda) & \text{when } \lambda > \tilde{\lambda}. \end{cases}$$

Clearly, this is the largest function that satisfies the shape constraints, and therefore must be an upper bound of the actual true CGF of interest. □

This ensures that if we replace $K_{\mathcal{M}}$ with $F[\bar{K}_{\mathcal{M}}]$ for any upper bound $\bar{K}_{\mathcal{M}}$, the computational properties of (2.2) and (2.3) remain unchanged.

At the moment, these projection steps are not implemented in autodp. It is therefore the user's responsibility to ensure that the base CGF of a mechanism $\mathcal{M}$ is implemented correctly and satisfy the convexity and monotonicity.

Approximate Computation of Theorem 9. The evaluation of the RDP parameter for a subsampled mechanism according to our bounds in Theorem 9 could still depend polynomially in α. We resolve this by only calculating this bound exactly up to some reasonable α_{thresh}[12], and then for $\alpha > \alpha_{\text{thresh}}$ we could use an optimization based-upper bound that we briefly describe below.

Noting that the expression in (3.7) (that is used in Theorem 9) can be written as a log-sum-exp or softmax function of $\alpha + 1$ items, where the jth item corresponds to:

$$\log \binom{\alpha}{j} + j \log \gamma + j \log \zeta(j).$$

Here, $\zeta(j)$ is the smallest of the upper bounds that we have of the ternary $|\mathcal{X}|^j$-DP of order j using RDP.

[12]The choice of α_{thresh} could change depending on the desired application of these bounds.

For any vector x of length $\alpha + 1$ we can use the following approximation:

$$\max(x) \leq \text{softmax}(x) \leq \max(x) + \log(\alpha).$$

When $\exp(x - \max(x))$ is dominated by a geometric series (which it often is for most mechanism $\mathcal{M}$ of interest), then we can further improve $\log(\alpha)$ by something independent to α.

The $\max(x)$ can be solved efficiently in $O(\log(\alpha))$ time as the function can have at most two local minima. This observation follows from the fact that $\log \zeta(j)$ (or any reasonable upper bound of it) is monotonically increasing, $j \log \gamma$ is monotonically decreasing, and that $\log \binom{\alpha}{j}$ is unimodal. Furthermore, we can use the Stirling approximation for $\log \binom{\alpha}{j}$ when α is large.

Numerical Stability in Computing the bound in Theorem 9. Since log-sum-exp is involved, we use the standard numerically stable implementation of the log-sum-exp function via: $\log(\sum_i \exp(x_i)) = \max_j x_j + \log(\sum_i \exp(x_i - \max_j(x_j)))$.

We also run into new challenges. For instance, the $\sum_{\ell=0}^{j} \binom{j}{\ell}(-1)^{j-\ell} e^{(\ell-1)\epsilon(\ell)}$ term involves taking structured differences of very large numbers that ends up being very small. We find that the alternative higher order finite difference operator representation $\Delta^{(j)}[e^{(\cdot-1)\epsilon(\cdot)}](0)$ and a polar representation of real numbers with a sign and log absolute value allows us to avoid floating point number overflow. However, the latter approach still suffers from the problem of error propagation and does not accurately compute the expression for large j.

To the best of our knowledge, the numerical considerations and implementation details of the moments accountant have not been fully investigated before, and accurately computing the closed form expression of χ^j-divergences using Rényi Divergences for large j remains an open problem of independent interest.

5. Experiments and Discussion

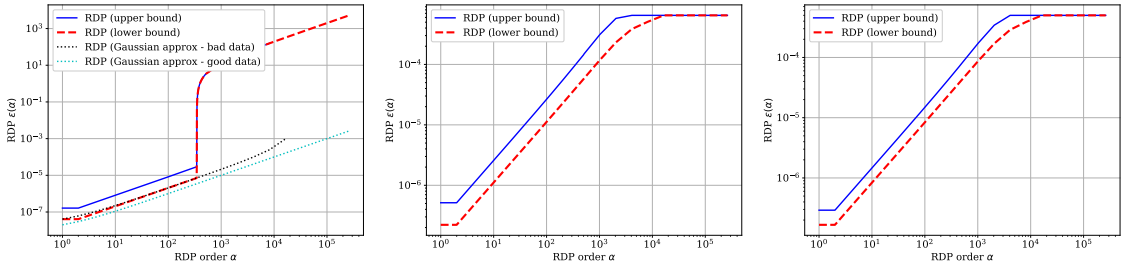
In this section, we present numerical experiments to demonstrate our upper and lower bounds of RDP for subsampled mechanisms and the usage of analytical moments accountant. In particular, we consider three popular randomized privacy mechanisms: (1) Gaussian mechanism, (2) Laplace mechanism, and (3) randomized response mechanism, and investigate the amplification effect of subsampling with these mechanisms on RDP. The RDP of these three mechanisms are known in analytical forms (See Mironov, 2017, Table II) :

$$\epsilon_{\text{Gaussian}(\alpha)} = \frac{\alpha}{2\sigma^2},$$

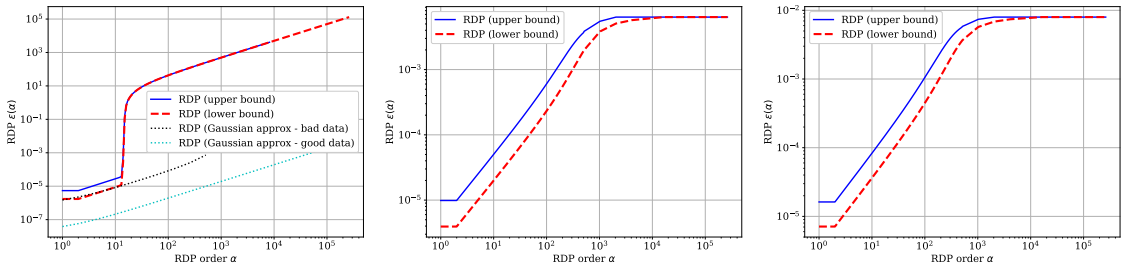
$$\epsilon_{\text{Laplace}(\alpha)} = \frac{1}{\alpha - 1} \log \left(\left(\frac{\alpha}{2\alpha - 1} \right) e^{(\alpha-1)/\lambda} + \left(\frac{\alpha - 1}{2\alpha - 1} \right) e^{-\alpha/\lambda} \right) \text{ for } \alpha > 1,$$

$$\epsilon_{\text{RandResp}(\alpha)} = \frac{1}{\alpha - 1} \log \left(p^\alpha (1-p)^{1-\alpha} + (1-p)^\alpha p^{1-\alpha} \right) \text{ for } \alpha > 1.$$

Here σ^2 represents the variance of the Gaussian perturbation, $2b^2$ the variance of the Laplace perturbation, and p the probability of replying truthfully in randomized response. We considered two groups of parameters σ, b, p for the three base mechanisms $\mathcal{M}$.



(A) Subsampled Gaussian mechanism with $\sigma = 5, \gamma = 0.001$ (B) Subsampled Laplace mechanism with $b = 2, \gamma = 0.001$ (C) Subsampled Rand. Resp. mechanism with $p = 0.6, \gamma = 0.001$



(D) Subsampled Gaussian mechanism with $\sigma = 0.5, \gamma = 0.001$ (E) Subsampled Laplace mechanism with $b = 0.5, \gamma = 0.001$ (F) Subsampled Rand. Resp. mechanism with $p = 0.9, \gamma = 0.001$

FIGURE 3. The RDP parameter ($\epsilon(\alpha)$) of the three subsampled mechanisms as a function of order α, with subsampling rate $\gamma = 0.001$ in all the experiments. The top row illustrates the case where the base mechanism $\mathcal{M}$ (before amplification using subsampling) is in a relatively high privacy regime (with $\epsilon \approx 0.5$) and the bottom row shows the low privacy regime with $\epsilon \approx 2$. RDP upper bound obtained through Theorem 9 is represented as the blue curve, and the corresponding lower bound obtained through Proposition 20 is represented as the red dashed curve. For the Gaussian case, we also present the RDP bound obtained through the asymptotic Gaussian approximation idea explained in Section 3.4.

High Privacy Regime: We set $\sigma = 5, b = 2$ and $p = 0.6$. These correspond to $(0.2\sqrt{2\log(1.25/\delta)}, \delta)$-DP, $(0.5, 0)$-DP, and approximately $(0.41, 0)$-DP for the Gaussian, Laplace, and Randomized response mechanisms, respectively, using the standard differential privacy calibration.

Low Privacy Regime: We set $\sigma = 1$, $b = 0.5$ and $p = 0.9$. These correspond to $(\sqrt{2\log(1.25/\delta)}, \delta)$-DP, $(2, 0)$-DP, and approximately $(2.2, 0)$-DP for the Gaussian, Laplace, and Randomized response mechanisms, respectively, using the standard differential privacy calibration.

The subsampling ratio γ is taken to be 0.001 for both regimes. More extensive experiments can be found in Appendix G, where we have also considered other subsampling ratios $\gamma \in \{0.01, 0.1, 0.5\}$ in both the low-noise and high-noise regimes.

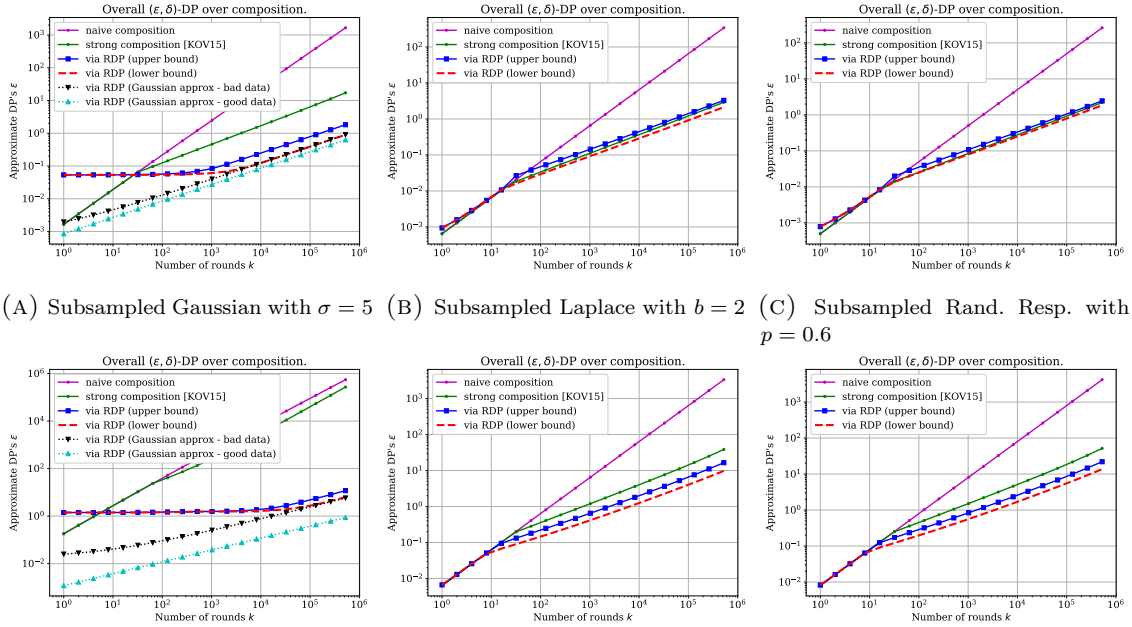
In Figure 3, we plot the upper and lower bounds (as well as asymptotic approximations whenever applicable) of RDP parameter $\epsilon'(\alpha)$ for the subsampled mechanism $\mathcal{M} \circ \mathsf{subsample}$

as a function of α. As we can see, the upper and lower bounds match up to a multiplicative constant for all the three mechanisms. There is a phase transition in the subsampled Gaussian case as we expect in both the upper and lower bound, which occurs at about $\gamma\alpha e^{\epsilon(\alpha)} < 1$. Note that our upper bound (the blue curve) matches the lower bound up to a multiplicative constant throughout in all regimes. For subsampled Gaussian mechanism in Plots 3a and 3d, the RDP parameter matches up to an (not visible in log scale) additive factor for large α. The RDP parameter for subsampled Laplace and subsampled randomized response (in the second and third column) are both linear in α at the beginning, then they flatten as $\epsilon(\alpha)$ approaches $\epsilon(\infty)$.

For the Gaussian mechanism we also plot an asymptotic approximation obtained under the assumption that the size of the input dataset grows $n \to \infty$ while the subsampling ratio $\gamma = m/n$ is kept constant. In fact, we derive two asymptotic approximations: one in the case of "good" data and one for "bad" data. The approximations and the definitions of "good" and "bad" data can be found in Section 3.4. The asymptotic Gaussian approximation with the "bad" data in Example 21 matches almost exactly with lower bound up to the phase transition point both in the high- and low-privacy regimes. The Gaussian approximation for the "good" data (with $n = 100/\gamma$) is smaller than the lower bound, especially in the low-privacy regime, highlighting that we could potentially gain a lot by performing a dataset-dependent analysis.

In Figure 4, we plot the overall (ϵ, δ)-DP for $\delta = 10^{-8}$ as we compose each of the three subsampled mechanisms for $600,000$ times. The ϵ is obtained as a function of δ for each k separately by calling the $\delta \Rightarrow \epsilon$ query in our analytical moments accountant. Our results are compared to the algorithm-independent techniques for differential privacy including naïve composition and strong composition. The strong composition baseline is carefully calibrated for each k by choosing an appropriate pair of $(\tilde{\epsilon}, \tilde{\delta})$ for $\mathcal{M}$ such that the overall (ϵ, δ)-DP guarantee that comes from composing k rounds of $\mathcal{M} \circ \mathsf{subsample}$ using Kairouz et al. (2017) obeys $\delta < 10^{-8}$ and ϵ is minimized. Each round is described by the $\big(\log(1+\gamma(e^{\tilde{\epsilon}}-1)), \gamma\tilde{\delta}\big)$-DP guarantee using the standard subsampling lemma (Lemma 3) and $\tilde{\epsilon}$ is obtained as a function of $\tilde{\delta}$ via (2.2).

Not surprisingly, both our approach and strong composition give an $\sqrt{k}$ scaling while the naïve composition has an $O(k)$ scaling throughout. An interesting observation for the subsampled Gaussian mechanism is that the RDP approach initially performs worse than the naïve composition and strong composition with the standard subsampling lemma. Our RDP lower bound certifies that this is not due to an artifact of our analysis but rather a fundamental limitation of the approach that uses RDP to obtain (ϵ, δ)-DP guarantees. We believe this is a manifestation of the same phenomenon that leads to the sub-optimality of the classical analysis of the Gaussian mechanism (Balle and Wang, 2018), which also relies on the conversion of a bound on the CGF of the privacy loss into an (ϵ, δ)-DP guarantee, and might be addressed using the necessary and sufficient condition for (ϵ, δ)-DP in terms of tail probabilities of the privacy loss random variable given in Balle and Wang (2018, Theorem 5). Luckily, such an artifact does not affect the typical usage of RDP: as the number of rounds of composition continues to grow, we end up having about an order of magnitude smaller ϵ than the baseline approaches in the high privacy regime (see Figure 4a) and five orders of magnitude smaller ϵ in the low privacy regime (see Figure 4d).



(A) Subsampled Gaussian with $\sigma = 5$ (B) Subsampled Laplace with $b = 2$ (C) Subsampled Rand. Resp. with $p = 0.6$

(D) Subsampled Gaussian with $\sigma = 0.5$ (E) Subsampled Laplace with $b = 0.5$ (F) Subsampled Rand. Resp. with $p = 0.9$

FIGURE 4. Comparison of techniques for strong composition of (ϵ, δ)-DP over $600,000$ data accesses with three different subsampled mechanisms. We plot ϵ as a function of the number of rounds of composition k with $\delta = 10^{-8}$ (note that smaller ϵ is better). The top row illustrates the case where the base mechanism $\mathcal{M}$ (before amplification using subsampling) is in a relatively high privacy regime (with $\epsilon \approx 0.5$) and the bottom row shows the low privacy regime with $\epsilon \approx 2$. We consider two baselines: the naïve composition that simply adds up (ϵ, δ) and the strong composition is through the result of Kairouz et al. (2017) with an optimal choice of per-round δ parameter computed for every k. The blue curve is based on the composition applied to the RDP upper bound obtained through Theorem 9, and the red dashed curve is based on the composition applied to the lower bound on RDP obtained through Proposition 20. For the Gaussian case, we also present the curves based on applying the composition on the RDP bound obtained through the Gaussian approximation idea explained in Section 3.4.

The results for composing subsampled Laplace mechanisms and subsampled randomized response mechanisms are shown in Figures 4b, 4c, 4e, and 4f. Unlike the subsampled Gaussian case, the RDP-based approach achieves about the same or better ϵ bound for all k when compared to what can be obtained using a subsampling lemma and strong composition.

6. Conclusion

In this paper, we studied the effect of subsampling (without replacement) in amplifying Rényi differential privacy (RDP). Specifically, we established a tight upper and lower bound for the RDP parameter for the randomized algorithm $\mathcal{M} \circ \mathsf{subsample}$ that first subsamples

the dataset then applies $\mathcal{M}$ to the subsample, in terms of the RDP parameter of $\mathcal{M}$. Our analysis also reveals interesting theoretical insight into the connection of subsampling to a linearized privacy random variable, higher order discrete differences of moment generating functions, as well as a ternary version of Pearson-Vajda divergence that appears fundamental in understanding and analyzing the effect of subsampling. In addition, we designed a data structure called *analytical moments accountant* which composes RDP for randomized algorithm (including subsampled ones) in symbolic forms and allows efficient conversion of RDP to (ϵ, δ)-DP for any δ (or ϵ) of choice. These results substantially expand the scope of the mechanisms with RDP guarantees to cover subsampled versions of Gaussian mechanism, Laplace mechanism, Randomized Responses, posterior sampling and so on, which facilitates flexible differentially private algorithm design. We compared our approach to the standard approaches that use subsampling lemma on (ϵ, δ)-DP directly and then applies strong composition, and in our experiments we notice an order of magnitude improvement in the privacy parameters with our bounds when we compose the subsampled Gaussian mechanism over multiple rounds.

Future work includes applying this technique to more advanced mechanisms for differentially private training of neural networks, addressing the data-dependent per-instance RDP for subsampled mechanisms, connecting the problem more tightly with statistical procedures that uses subsampling/resampling as key components such as *bootstrap* and *jackknife*, as well as combining this approach with subsampling-based sublinear algorithms for exploratory data analysis.

Acknowledgment

The authors thank the anonymous reviewers and the journal's editor for the helpful comments that led to significant improvements of the paper's presentation. We also thank Ilya Mironov and Kunal Talwar for helpful discussions and the clarification of their proof of Lemma 3 in (Abadi et al., 2016).

References

Abadi, Martín, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang (2016) "Deep learning with differential privacy," in *ACM Conference on Computer and Communications Security (CCS-16)*, DOI: 10.1145/2976749.2978318.

Apple, Differential Privacy Team (2017) "Learning with privacy at scale," *Apple Machine Learning Journal*, URL: https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.

Balle, Borja, Gilles Barthe, and Marco Gaboardi (2018) "Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences," in *Advances in Neural Information Processing Systems (NeurIPS)*, URL: http://papers.nips.cc/paper/7865-privacy-amplification-by-subsampling-tight-analyses-via-couplings-and-divergences.

Balle, Borja and Yu-Xiang Wang (2018) "Improving Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising," *International Conference in Machine Learning (ICML)*, URL: http://proceedings.mlr.press/v80/balle18a.html.

Bassily, Raef, Adam Smith, and Abhradeep Thakurta (2014) "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Foundations of Computer Science (FOCS-14)*, DOI: 10.1109/FOCS.2014.56.

Beimel, Amos, Kobbi Nissim, and Uri Stemmer (2013) "Characterizing the sample complexity of private learners," in *Conference on Innovations in Theoretical Computer Science (ITCS)*, DOI: 10.1145/2422436.2422450.

Bobkov, Sergey G, GP Chistyakov, Friedrich Götze et al. (2019) "Rényi divergence and the central limit theorem," *The Annals of Probability*, Vol. 47, No. 1, pp. 270–323, DOI: 10.1214/18-AOP1261.

Bun, Mark, Cynthia Dwork, Guy N. Rothblum, and Thomas Steinke (2018) "Composable and versatile privacy via truncated CDP," in *ACM Symposium on Theory of Computing, (STOC-18)*, URL: https://doi.org/10.1145/3188745.3188946, DOI: 10.1145/3188745.3188946.

Bun, Mark, Kobbi Nissim, Uri Stemmer, and Salil Vadhan (2015) "Differentially private release and learning of threshold functions," in *Foundations of Computer Science (FOCS-15)*, URL: https://doi.org/10.1109/FOCS.2015.45, DOI: 10.1109/FOCS.2015.45.

Bun, Mark and Thomas Steinke (2016) "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference (TCC)*, DOI: 10.1007/978-3-662-53641-4_24.

Dwork, Cynthia, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor (2006a) "Our data, ourselves: Privacy via distributed noise generation," in *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT-16)*, Springer, URL: https://doi.org/10.1007/11761679_29.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006b) "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography (TCC)*, DOI: 10.1007/11681878_14.

Dwork, Cynthia and Aaron Roth (2013) "The algorithmic foundations of differential privacy," *Theoretical Computer Science*, Vol. 9, No. 3-4, DOI: 10.1561/0400000042.

Dwork, Cynthia and Guy N Rothblum (2016) "Concentrated differential privacy," *arXiv preprint arXiv:1603.01887*, URL: http://arxiv.org/abs/1603.01887.

Dwork, Cynthia, Guy N Rothblum, and Salil Vadhan (2010) "Boosting and differential privacy," in *Foundations of Computer Science (FOCS-10)*, IEEE, DOI: 10.1109/FOCS.2010.12.

Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova (2014) "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *ACM conference on computer and communications security (CCS-14)*, DOI: 10.1145/2660267.2660348.

Foulds, James, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri (2016) "On the Theory and practice of privacy-preserving Bayesian data analysis," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, AUAI Press, URL: http://auai.org/uai2016/proceedings/papers/45.pdf.

Geumlek, Joseph, Shuang Song, and Kamalika Chaudhuri (2017) "Renyi Differential Privacy Mechanisms for Posterior Sampling," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 5289–5298, URL: http://papers.nips.cc/paper/7113-renyi-differential-privacy-mechanisms-for-posterior-sampling.

Gil, Manuel, Fady Alajaji, and Tamas Linder (2013) "Rényi divergence measures for commonly used univariate continuous distributions," *Information Sciences*, Vol. 249, DOI: 10.1016/j.ins.2013.06.018.

Kairouz, Peter, Sewoong Oh, and Pramod Viswanath (2017) "The Composition Theorem for Differential Privacy," *IEEE Transactions on Information Theory*, Vol. 63, No. 6, pp. 4037–4049, DOI: 10.1109/TIT.2017.2685505.

Kasiviswanathan, Shiva Prasad, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith (2011) "What can we learn privately?" *SIAM Journal on Computing*, Vol. 40, No. 3, DOI: 10.1137/090756090.

Lukacs, Eugene (1970) *Characteristic functions*: Griffin.

Madow, William G (1948) "On the limiting distributions of estimates based on samples from finite universes," *The Annals of Mathematical Statistics*, pp. 535–545, DOI: 10.1214/aoms/1177730149.

Mironov, Ilya (2017) "Rényi differential privacy," in *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, DOI: 10.1109/CSF.2017.11.

Murtagh, Jack and Salil P. Vadhan (2018) "The Complexity of Computing the Optimal Composition of Differential Privacy," *Theory of Computing*, Vol. 14, No. 1, pp. 1–35, DOI: 10.4086/toc.2018.v014a008.

Nielsen, Frank and Richard Nock (2014) "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, Vol. 21, No. 1, DOI: 10.1109/LSP.2013.2288355.

Song, Shuang, Kamalika Chaudhuri, and Anand D Sarwate (2013) "Stochastic gradient descent with differentially private updates," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248, IEEE, DOI: 10.1109/GlobalSIP.2013.6736861.

Uber Security (2017) "Uber Releases Open Source Project for Differential Privacy.," https://medium.com/uber-security-privacy/differential-privacy-open-source-7892c82c42b6.

Ullman, Jonathan (2017) "CS7880: Rigorous Approaches to Data Privacy, Spring 2017.," http://www.ccs.neu.edu/home/jullman/PrivacyS17/HW1sol.pdf.

Vajda, Igor (1973) "χ^α-divergence and generalized Fisher information," in *Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, p. 223, Academia.

Van Erven, Tim and Peter Harremos (2014) "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, Vol. 60, No. 7, DOI: 10.1109/TIT.2014.2320500.

Wang, Yu-Xiang (2019) "Per-instance Differential Privacy," *Journal of Privacy and Confidentiality*, Vol. 9, No. 1, DOI: 10.29012/jpc.662.

Wang, Yu-Xiang, Stephen Fienberg, and Alex Smola (2015) "Privacy for free: Posterior sampling and stochastic gradient monte carlo," in *International Conference on Machine Learning (ICML-15)*, pp. 2493–2502, URL: http://proceedings.mlr.press/v37/wangg15.html.

Wang, Yu-Xiang, Jing Lei, and Stephen E. Fienberg (2016) "Learning with Differential Privacy: Stability, Learnability and the Sufficiency and Necessity of ERM Principle," *Journal of Machine Learning Research*, Vol. 17, No. 183, URL: http://jmlr.org/papers/v17/15-313.html.

Appendix A. Composition of Differentially Private Mechanisms

Composition theorems for differential privacy allow a modular design of privacy preserving mechanisms based on mechanisms for simpler sub tasks:

Theorem 25 Naïve composition, Dwork et al. (2006a)**.** *A mechanism that permits k adaptive interactions with mechanisms that preserves (ϵ, δ)-differential privacy (and does not access the database otherwise) ensures $(k\epsilon, k\delta)$-differential privacy.*

A stronger composition is also possible as shown by Dwork et al. (2010).

Theorem 26 Strong composition, Dwork et al. (2010)**.** *Let $\epsilon, \delta, \delta^* > 0$ and $\epsilon \leq 1$. A mechanism that permits k adaptive interactions with mechanisms that preserves (ϵ, δ)-differential privacy ensures $(\epsilon\sqrt{2k\ln(1/\delta^*)} + 2k\epsilon^2, k\delta + \delta^*)$-differential privacy.*

Kairouz et al. (2017) recently gave an optimal composition theorem for differential privacy, which provides an exact characterization of the best privacy parameters that can be guaranteed when composing a number of (ϵ, δ)-differentially private mechanisms. Unfortunately, the resulting optimal composition bound is quite complex to state exactly, and indeed is even #P-complete to compute exactly when composing mechanisms with different (ϵ_i, δ_i) parameters (Murtagh and Vadhan, 2018).

Appendix B. Discrete Difference Operators and Newton's Series Expansion

In this section, we provide more details of the discrete calculus objects that we used in the proof, and also illustrate how the interesting identity (3.2) comes about.

Discrete Difference Operators. Discrete difference operators are linear operators that transform a function into its discrete derivatives. Let f be a function $\mathbb{R} \to \mathbb{R}$, the first order forward difference operator of f is a function such that

$$\Delta[f](x) = f(x+1) - f(x).$$

The αth order forward difference operator $\Delta^{(\alpha)}$ can be constructed recursively by

$$\Delta^{(\alpha)} = \Delta \circ \Delta^{(\alpha-1)}$$

for all $\alpha = 1, 2, 3, ...$ with $\Delta^{(1)} := \mathrm{Id}$.

The forward difference operators are linear transformation of functions that can be thought of as a convolution (denoted by $\star$) with a linear combination of Dirac-delta functions (δ_{dirac}), which we call filters.

$$\Delta[f] = f \star (\delta_{\mathrm{dirac}}(x-1) - \delta_{\mathrm{dirac}}(x)).$$

From the linear combination point of view, the first order forward difference operator is the linear combination of the (infinite) basis functions of Dirac-delta functions supported on all integers with coefficient sequence $[..., 0, -1, 1, 0, ...]$. This sequence of coefficients uniquely defines the difference operators. For example, when $\alpha = 2$, the coefficients that construct operator $\Delta^{(\alpha)}$ are

$$\ldots, 0, 0, 1, -2, 1, 0, 0 \ldots$$

and when $\alpha = 3$ and $\alpha = 4$, we get

$$\ldots, 0, 0, -1, 3, -3, 1, 0, 0 \ldots$$

and

$$\ldots, 0, 0, 1, -4, 6, -4, 1, 0, 0 \ldots$$

respectively. In general, these convolution operators can be constructed by Pascal's triangle of the αth order, or simply the binomial coefficients with alternating signs.

When computing the bound in Theorem 9 we need to calculate $\Delta^{(\ell)}[f](0)$ for all integer $\ell \leq \alpha$. The recursive definition of the bound above allows us to compute all finite differences up to order α by $O(\alpha^2)$ evaluation of f rather than the naïve direct calculation of $O(\alpha^3)$. In Section 4, we describe some further speed-ups with approximate evaluation.

Newton Series Expansion. Newton series expansion is the discrete analogue of the continuous Taylor series expansion, with all derivatives replaced with discrete difference operators and all monomials replaced with falling factorials.

Consider infinitely differentiable function $f : \mathbb{R} \to \mathbb{R}$. The Taylor series expansion of f at 0 and the Newton series expansion of f at 0 are respectively:

$$f(x) = f(0) + \frac{\partial}{\partial x}[f](0)x + \frac{\partial^2}{\partial x^2}[f](0)\frac{x^2}{2!} + ... + \frac{\partial^k}{\partial x^k}[f](0)\frac{x^k}{k!} + ...$$

$$f(x) = f(0) + \Delta^{(1)}[f](0)x + \Delta^{(2)}[f](0)\frac{x(x-1)}{2!} + ... + \Delta^{(k)}[f](0)\frac{(x)_k}{k!} + ...$$

where $(x)_k$ denotes the falling factorials $x(x-1)(x-2)...(x-k+1)$. For integer x, it is clear that the Newton's series expansion has a finite number of terms.

Appendix C. On Tightness and Self-consistency Guarantees

When specifying a sequence of RDP guarantees for $\mathcal{M}$ in terms of

$$\sup_{X,X':d(X,X')\leq 1} D_\alpha(\mathcal{M}(X)\|\mathcal{M}(X')) \leq \epsilon(\alpha)$$

it really matters whether $\epsilon(\alpha)$ is the exact analytical form of some underlying pairs of distributions induced by a pair of adjacent datasets X, X' or just a sequence of conservative estimates. If it is the latter, then it is unclear at which α the slacks are bigger and at which α the slacks are smaller. And the sequence of $\epsilon(\cdot)$ might not be realizable by any pairs distributions. For example, if we use a polynomial upper bound of $\epsilon(\cdot)$, we know from the theory of CGF that no distribution has a CGF of polynomial order higher than 2 and the only distribution that has polynomial order exactly two is the Gaussian distribution (Lukacs, 1970).

In this section, we provide an example proof that the analytical Rényi DP bound of the Gaussian mechanisms (defined in Section 2) are self-consistent. Again for simplicity, for the Gaussian mechanism, we assume that the sensitivity of function f is 1.

Lemma 27 . *For the Gaussian mechanism, $\epsilon(\alpha) = \alpha/(2\sigma^2)$ is tight and self-consistent.*

Proof. The Gaussian mechanism with variance σ^2 has a tight RDP parameter bound $\epsilon(\alpha) = \frac{\alpha}{2\sigma^2}$ (Gil et al., 2013). This is achieved by the distributions $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(1, \sigma^2)$.

For self-consistency, it suffices to show that the $|\chi|^\alpha$-divergence's maximum for every even α are also achieved by the same pair of distributions. Consider $q = \mathcal{N}(0, \sigma^2)$ and $p = \mathcal{N}(\mu, \sigma^2)$ for $0 \le \mu \le 1$

$$D_{|\chi|^\alpha}(p\|q) = \mathbb{E}_q[(p/q-1)^\alpha] = \mathbb{E}_q[(e^{-\frac{-2x\mu+\mu^2}{2\sigma^2}}-1)^\alpha] = \Delta^{(\alpha)}[e^{(\ell^2-\ell)\mu^2}](0)$$

Differentiating w.r.t. μ, we get

$$2\mu(\ell^2-\ell)\Delta^{(\alpha)}[\mathbb{E}_q[e^{(\ell^2-\ell)\mu^2}]](0) \ge 0$$

for $\mu > 0$. In other words, the divergence is monotonically increasing in μ. □

In general, verifying the self-consistency is not straightforward, but since $|\chi|^\alpha$-divergence is a proper f-divergence, it is jointly convex in its arguments. When the set of distributions is a convex polytope, it suffices to check for this condition at all the vertices of the polytope.

Appendix D. Other Properties of Ternary-$|\chi|^\alpha$-DP

When $\alpha = 1$, both the binary- and ternary-$|\chi|^\alpha$-divergence become the total variation distance. When $\alpha = 2$ the binary-$|\chi|^\alpha$-divergence becomes the χ^2-distance.

The following lemma shows that we can convert binary-$|\chi|^\alpha$-DP (and therefore, ternary-$|\chi|^\alpha$-DP) to the more standard (ϵ, δ)-DP using the tail bound of a privacy random variable.

Lemma 28 $|\chi|^\alpha$-differential privacy $\Rightarrow$ (ϵ, δ)-DP**.** *If an algorithm is ξ-binary-$|\chi|^\alpha$-DP, then it is also $\left(\epsilon, \left(\frac{\xi(\alpha)}{e^\epsilon - 1}\right)^\alpha\right)$-DP for all $\epsilon > 0$ and equivalently, $(\log \xi(\alpha) - 1 + \frac{\log(1/\delta)}{\alpha}, \delta)$ for all $\delta > 0$.*

Proof. By Markov's inequality,

$$\Pr[|p/q - 1| > t] \le \mathbb{E}[|p/q-1|^\alpha]/t^\alpha = \left(\frac{\xi(\alpha)}{t}\right)^\alpha.$$

The results follows from changing the variable from p/q to $e^{\log(p/q)}$. □

The following lemma shows that we can bound the above by a quantity that depends on the Rényi divergence and the Pearson-Vajda divergence. It also generalizes Lemma 17 that we used in the proof of Theorem 9.

Lemma 29 . *Let p, q, r are three distributions. For all conjugate pair $u, v \ge 1$ such that $1/u + 1/v = 1$, and all integer $j \ge 2$ we have that*

$$\mathbb{E}_r\left[\left(\frac{|p-q|}{r}\right)^j\right] \le e^{(j-1)D_{(j-1)v+1}(q\|r)} D_{|\chi|^{ju}}(p\|q)^{1/u}.$$

Proof. The proof is a straightforward application of Hölder's inequality.

$$\begin{aligned}\mathbb{E}_r\left[\left(\frac{|p-q|}{r}\right)^j\right] &= \int r\left(\frac{q}{r}\right)^j \left|\frac{p}{q}-1\right|^j d\theta \underset{\text{Change of measure}}{=} \int q\left(\frac{q}{r}\right)^{j-1}\left|\frac{p}{q}-1\right|^j d\theta \\ &\underset{\text{Hölder}}{\leq} \left(\mathbb{E}_q\left[\left(\frac{q}{r}\right)^{(j-1)v}\right]\right)^{1/v}\left(\mathbb{E}_q\left[\left(\frac{p}{q}-1\right)^{ju}\right]\right)^{1/u} \\ &= e^{(j-1)D_{(j-1)v+1}(q\|r)} D_{|\chi|^{ju}}(p\|q)^{1/u}.\end{aligned}$$

□

Remark 30 . *When we take $v=\infty$ and $u=1$, we recover the result from Lemma 17. When we take $u=v=2$, this guarantees that ju is an even number and the above results becomes*

$$\mathbb{E}_r\left[\left(\frac{|p-q|}{r}\right)^j\right] \leq e^{(j-1)D_{2j-1}(q\|r)}\sqrt{\Delta^{(2j)}[e^{(\cdot-1)D_{(\cdot)}(p\|q)}](0)},$$

where $\Delta^{(2j)}$ is the finite difference operator of order $2j$. Note that $e^{(\cdot-1)D_{(\cdot)}(q\|r)}$ can be viewed as the moment generating function of the random variable $\log(p(\theta)/q(\theta))$ induced by $\theta\sim q$. The $2j$th order discrete derivative of the MGF at 0 is $\mathbb{E}_q[(\frac{p}{q}-1)^{2j}]$, which very nicely mirrors the corresponding $2j$th order continuous derivative of the MGF evaluated at 0, which by the property of an MGF is $\mathbb{E}_q[\log(p/q)^{2j}]$.

Appendix E. Properties of CGFs and Rényi Divergence

In this section, we highlight some interesting properties of CGF, which in part enables our analytical moments accountant data structure described in Section 4.

Lemma 31 . *The CGF of a random variable (if finite for $\lambda\in\mathbb{R}$), obeys that:*

(a) *It is infinitely differentiable.*

(b) *$\frac{\partial}{\partial\lambda}K_{\mathcal{M}}(\lambda)$ monotonically increases from the infimum to the supremum of the support of the random variable.*

(c) *It is convex (and strictly convex for all distributions that are not a single point mass).*

(d) *$K_{\mathcal{M}}(0)=0$, i.e., it passes through the origin.*

(e) *The CGF of a privacy loss random variable further obeys that $K_{\mathcal{M}}(-1)=0$.*

These properties are used in establishing the computational properties of the analytical moments accountant as we have seen before.

We provide a first-principle proof of convexity (c), which is elementary and does not use a variational characterization of the Rényi divergence as in the Corollary 2 of Van Erven and Harremos (2014).

Proof. We use the definition of convex functions. By definition, for all $\lambda \geq 0$, we have

$$K_{\mathcal{M}}(\lambda) = \log \mathbb{E}_p[e^{\lambda \log \frac{p(\theta)}{q(\theta)}}] = \log \mathbb{E}_p \left[\left(\frac{p(\theta)}{q(\theta)} \right)^{\lambda} \right].$$

Let $\lambda_1, \lambda_2 \geq 0$ and $v \in [0,1]$. Take $\lambda = (1-v)\lambda_1 + v\lambda_2$ and apply Hölder's inequality with the exponents being the conjugate pairs $1/(1-v)$ and $1/v$:

$$\begin{aligned}
\mathbb{E}_p \left[\left(\frac{p(\theta)}{q(\theta)} \right)^{\lambda} \right] &= \mathbb{E}_p \left[\left(\frac{p(\theta)}{q(\theta)} \right)^{(1-v)\lambda_1 + v\lambda_2} \right] = \mathbb{E}_p \left[\left(\frac{p(\theta)}{q(\theta)} \right)^{(1-v)\lambda_1} \left(\frac{p(\theta)}{q(\theta)} \right)^{v\lambda_2} \right] \\
&\leq \mathbb{E}_p \left[\left(\frac{p(\theta)}{q(\theta)} \right)^{\lambda_1} \right]^{1-v} \mathbb{E}_p \left[\left(\frac{p(\theta)}{q(\theta)} \right)^{\lambda_2} \right]^{v} \\
&= \exp[K_{\mathcal{M}}(\lambda_1)]^{1-v} \exp[K_{\mathcal{M}}(\lambda_2)]^{v}.
\end{aligned}$$

Take logarithm on both sides, we get

$$K_{\mathcal{M}}((1-v)\lambda_1 + v\lambda_2) \leq (1-v)K_{\mathcal{M}}(\lambda_1) + vK_{\mathcal{M}}(\lambda_2)$$

and the proof is complete. □

Corollary 32 . *Optimization problem* (2.3) *is log-convex. Optimization problem* (2.2) *is unimodal / quasi-convex.*

Proof. To see the first claim, check that the logarithm of (2.3) is the sum of a convex function and an affine function, which is convex. To see the second claim, first observe $1/\lambda$ is monotonically decreasing in $\mathbb{R}_+$. It suffices to show that $\frac{K_{\mathcal{M}}(\lambda)}{\lambda}$ is monotonically increasing. Let $\partial K_{\mathcal{M}}(\lambda)$ be a subgradient of $K_{\mathcal{M}}(\lambda)$, we can take the "derivative" of the function

$$\lim_{\delta \to 0} \frac{1}{\delta} \left(\frac{K_{\mathcal{M}}(\lambda+\delta)}{\lambda+\delta} - \frac{K_{\mathcal{M}}(\lambda)}{\lambda} \right) \geq \frac{\partial K_{\mathcal{M}}(\lambda)}{\lambda} - \frac{K_{\mathcal{M}}(\lambda)}{\lambda^2} \geq 0$$

The last inequality follows from the first order condition of a convex function

$$K_{\mathcal{M}}(0) \geq K_{\mathcal{M}}(\lambda) + (0-\lambda) \cdot \partial K_{\mathcal{M}}(\lambda)$$

and that $K_{\mathcal{M}}(0) = 0$. □

The corollary implies that optimization problems defined in (2.2) and (2.3) have unique minimizers and they can be solved efficiently using bisection or convex optimization to arbitrary precision even if all we have is (possibly noisy) blackbox access to $K_{\mathcal{M}}(\cdot)$ or its derivative.

Appendix F. Rényi Divergence of Exponential Family Distributions and RDP

Exponential Family Distributions. Let θ be a random variable whose distribution is by ϕ. It is an exponential family distribution if the probability density function can be written as

$$p(\theta; \phi) = h(\theta) \exp(\eta(\phi)^\top T(\theta) - F(\phi)).$$

If we re-parameterize, we can rewrite the exponential family distribution as a *natural* exponential family

$$p(\theta;\eta) = h(\theta)\exp(\eta^\top T(\theta) - A(\eta))$$

where the normalization constant A is called the log-partition function.

Rényi Divergence of Two Natural Exponential Family Distributions. Let $\mathcal{S}$ be the natural parameter space, i.e., every $\eta \in \mathcal{S}$ defines a valid distribution. Then for $\eta_1, \eta_2 \in \mathcal{S}$, the Rényi divergence between the two exponential family distribution $p_{\eta_1} := p(\theta;\eta_1)$ and $p_{\eta_2} := p(\theta;\eta_2)$ is:

(1) If $\alpha \notin \{0,1\}$ and $\alpha\eta_1 + (1-\alpha)\eta_2 \in \mathcal{S}$,

$$D_\alpha(p_{\eta_1}\|p_{\eta_2}) = \frac{1}{\alpha-1}\log\left(\frac{A(\alpha\eta_1+(1-\alpha)\eta_2)}{A(\eta_1)^\alpha A(\eta_2)^{1-\alpha}}\right).$$

(2) If $\alpha \notin \{0,1\}$ and $\alpha\eta_1 + (1-\alpha)\eta_1 \notin \mathcal{S}$,

$$D_\alpha(p_{\eta_1}\|p_{\eta_2}) = +\infty$$

(3) If $\alpha = 1$,

$$D_\alpha(p_{\eta_1}\|p_{\eta_2}) = D_{KL}(p_{\eta_1}\|p_{\eta_2}) = (\eta_1-\eta_2)^\top \nabla_\eta A(\eta_1) + A(\eta_2) - A(\eta_1),$$

namely, the Kullback-Liebler divergence of the two distributions and also the Bregman divergence with respect to convex function A.

(4) If $\alpha = 0$,

$$D_\alpha(p_{\eta_1}\|p_{\eta_2}) = -\log(\Pr_{\eta_2}[p_{\eta_1} > 0]).$$

For example, the Rényi divergence between multivariate normal distributions $\mathcal{N}(\mu_1,\Sigma_1), \mathcal{N}(\mu_2,\Sigma_2)$ equals (Gil et al., 2013)

$$\begin{aligned} &D_\alpha(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)) \\ =&\begin{cases} +\infty, & \text{if } \Sigma_\alpha := \alpha\Sigma_2 + (1-\alpha)\Sigma_1 \text{ is not positive definite.} \\ \frac{\alpha}{2}(\mu_1-\mu_2)^\top\Sigma_\alpha^{-1}(\mu_1-\mu_2) - \frac{1}{2(\alpha-1)}\log\left(\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}\right), & \text{otherwise.} \end{cases} \end{aligned}$$

Exponential Family Mechanisms and its Rényi-DP. Let the differentially private mechanism to release θ be sampling from an exponential family. Let

$$p(\theta) = h(\theta)\exp(\eta(X)^\top T(\theta) - A(\eta(X)))$$

denote the distribution induced by this differentially private mechanism on dataset X, and similarly let

$$q(\theta) = h(\theta)\exp(\eta(X')^\top T(\theta) - A(\eta(X'))).$$

be the corresponding distribution when the dataset is X'.

In this case, the privacy random variable $\log(p/q)$ has a specific form

$$\varphi(\theta) = [\eta(X) - \eta(X')]^\top T(\theta) - [A(\eta(X)) - A(\eta(X'))].$$

Using this, it can be shown that the α-Rényi divergence between p and q is

$$\begin{aligned} D_\alpha(p\|q) &= \log \mathbb{E}_q \left[e^{\alpha\varphi(\theta)}\right]^{\frac{1}{\alpha-1}} \\ &= \frac{1}{\alpha-1}\left[A(\alpha\eta(X)+(1-\alpha)\eta(X')) - \alpha A(\eta(X)) - (1-\alpha)A(\eta(X'))\right]. \end{aligned}$$

A special case of the exponential family mechanisms of particular interest is the posterior sampling mechanisms where $\eta(X)$ has a specific form (Geumlek et al., 2017).

To obtain RDP from the above closed-form Rényi divergence, it remains to maximize over two adjacent datasets X, X'. We make a subset of the following three assumptions.

(A) Bounded parameter difference: $\sup_{X,X':d(X,X')\le 1} \|\eta(X)-\eta(X')\| \le \Delta$ with respect a norm $\|\cdot\|$.

(B) (B,κ)-Local Lipschitz: The log-partition function A is (B,κ)-Local Lipschitz with respect to $\|\cdot\|$ if for all dataset X and all η such that $\|\eta-\eta(X)\| \le \kappa$, we have

$$A(\eta) \le A(\eta(X)) + B\|\eta-\eta(X)\|.$$

(C) (L,κ)-Local smoothness: The log-partition function A is (L,κ)-smooth with respect to $\|\cdot\|$ if for all dataset X and all η such that $\|\eta-\eta(X)\| \le \kappa$, we have

$$A(\eta) \le A(\eta(X)) + \langle \nabla A(\eta(X)), \eta-\eta(X)\rangle + L\|\eta-\eta(X)\|^2.$$

The following proposition refines the results of (Geumlek et al., 2017, Lemma 3).

Proposition 33 RDP of exponential family mechanisms**.** *Let $\mathcal{M}$ be an exponential family mechanism that obeys Assumption (A)(B)(C) with parameter Δ, B, L, κ with a common norm $\|\cdot\|$. If in addition, $\kappa \ge \Delta$, then $\mathcal{M}$ obeys $(\alpha, \epsilon(\alpha))$-RDP for all $\alpha \in (1, \kappa/\Delta+1]$ with*

$$\epsilon(\alpha) \le \min\left\{\frac{\alpha L\Delta^2}{2}, 2B\Delta\right\}.$$

Remark 34 . *We can view B and L as (nondecreasing) functions of κ. For any fixed α of interest, we can optimize over all feasible choice of κ:*

$$\epsilon(\alpha) \le \min_{\kappa:\alpha\Delta\le\kappa} \min\left\{\alpha L(\kappa)\Delta^2, 2B(\kappa)\Delta\right\} = \min\left\{\alpha L(\alpha\Delta)\Delta^2, 2B(\alpha\Delta)\Delta\right\}.$$

In fact, as can be seen clearly from the proof, $2B(\alpha\Delta)\Delta$ can be improved to $[B((\alpha-1)\Delta) + B(\Delta)]\Delta$.

Proof of Proposition 33. Assumption (A) implies that $\|\eta(X)-\eta(X')\| \le \Delta$. Note that for all $\alpha \le \kappa/\Delta+1$, $\|\alpha\eta(X)+(1-\alpha)\eta(X')-\eta(X)\| \le \kappa$. Assumption (B) implies that

$$A(\alpha\eta(X)+(1-\alpha)\eta(X')) \le A(\eta(X)) + (\alpha-1)B\|\eta(X')-\eta(X))\| \le A(\eta(X)+(\alpha-1)B\Delta,$$

and that

$$A(\eta(X')) \le A(\eta(X)) + B\Delta.$$

Substitute these into the definition of $D_\alpha(p\|q)$ we get that

$$D_\alpha(p\|q) \le \frac{1}{\alpha-1}[A(\eta(X)) + (\alpha-1)B\Delta - A(\eta(X)) + (\alpha-1)B\Delta] = 2B\Delta. \tag{F.1}$$

Assumption (C) implies that for all $\alpha \le \kappa/\Delta + 1$

$$\begin{aligned}
&A(\alpha\eta(X) + (1-\alpha)\eta(X')) = A(\eta(X) + (\alpha-1)(\eta(X) - \eta(X'))) \\
\le& A(\eta(X)) + (\alpha-1)\langle \nabla A(\eta(X), \eta(X) - \eta(X'))\rangle + \frac{(\alpha-1)^2 L}{2}\|\eta(X) - \eta(X')\|^2 \\
\le& A(\eta(X)) + (\alpha-1)\langle \nabla A(\eta(X), \eta(X) - \eta(X'))\rangle + \frac{(\alpha-1)^2 L\Delta^2}{2}
\end{aligned}$$

where the last step uses Assumption (A). Assumption (C) also implies that

$$\begin{aligned}
A(\eta(X')) - A(\eta(X) &\le \langle \nabla A(\eta(X), \eta(X') - \eta(X))\rangle + \frac{L\|\eta(X) - \eta(X')\|^2}{2} \\
&\le \langle \nabla A(\eta(X), \eta(X) - \eta(X'))\rangle + \frac{L\Delta^2}{2}.
\end{aligned}$$

Substitute these into the definition of $D_\alpha(p\|q)$ we get that

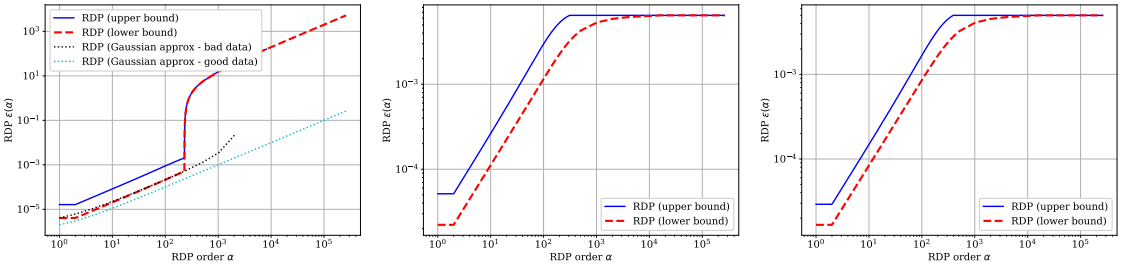
$$\begin{aligned}
D_\alpha(p\|q) \le \frac{1}{\alpha-1}\Big[&A(\eta(X)) + (\alpha-1)\langle \nabla A(\eta(X), \eta(X) - \eta(X'))\rangle + \frac{(\alpha-1)^2 L\Delta^2}{2} \\
&- A(\eta(X)) + (\alpha-1)\langle \nabla A(\eta(X), \eta(X') - \eta(X))\rangle + \frac{(\alpha-1)L\Delta^2}{2}\Big] = \frac{\alpha L\Delta^2}{2},
\end{aligned}$$

which, together with (F.1), produces the bound as claimed. □

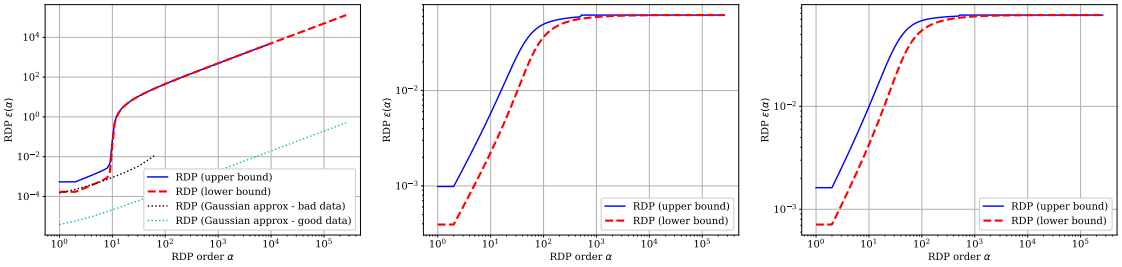
Appendix G. Experiments with varying sampling ratio

In this section, we present additional experiments to demonstrate the behaviors of the analytical moments accountant and our bounds in Theorem 9 and 19 under a variety of sampling ratios γ. As we can see, we observe qualitatively the same behavior as in the case of $\gamma = 0.001$. The gap between the upper and lower bounds of the RDP of the subsampled mechanisms widens by a factor of $(1-\gamma)^{-1}$ as we increase γ, but the experiments have demonstrated that provided that γ is bounded away from 1, the privacy amplification bounds remain meaningful. In particular, in the low-noise regime, it still provides substantial improvements over the standard advanced composition with (ϵ, δ)-DP.

We note that small sampling ratio is often desirable in cases such as running stochastic gradient methods to privately train machine learning models, as it allows the algorithm to update the parameters with many more iterations. The artifact of our bounds at $\gamma \to 1$ is basically saying that whenever there is a need to set $\gamma > 0.5$, the learner will be better off just running full-gradient descent (with noise added for privacy).

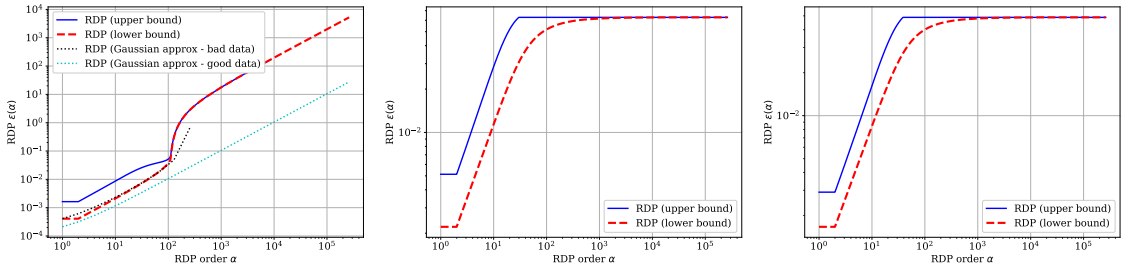


(A) Subsampled Gaussian mechanism with $\sigma = 5, \gamma = 0.01$ (B) Subsampled Laplace mechanism with $b = 2, \gamma = 0.01$ (C) Subsampled Rand. Resp. mechanism with $p = 0.6, \gamma = 0.01$



(D) Subsampled Gaussian mechanism with $\sigma = 0.5, \gamma = 0.01$ (E) Subsampled Laplace mechanism with $b = 0.5, \gamma = 0.01$ (F) Subsampled Rand. Resp. mechanism with $p = 0.9, \gamma = 0.01$

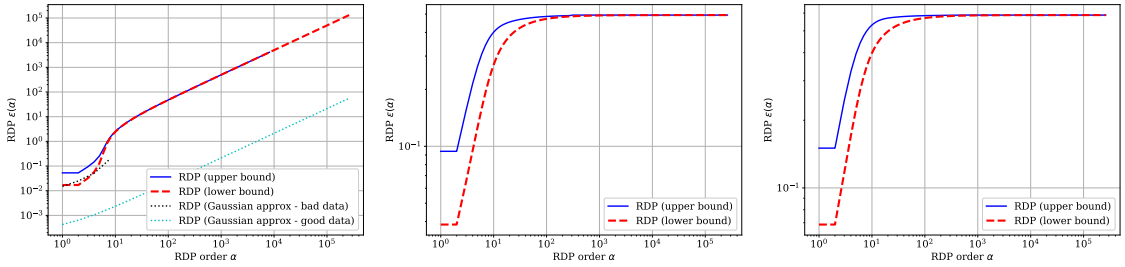
FIGURE 5. The same experiments reported in Figure 3 (the RDP as a function of α) but with $\gamma = 0.01$.



(A) Subsampled Gaussian mechanism with $\sigma = 5, \gamma = 0.1$

(B) Subsampled Laplace mechanism with $b = 2, \gamma = 0.1$

(C) Subsampled Rand. Resp. mechanism with $p = 0.6, \gamma = 0.1$

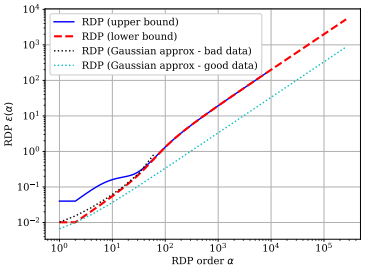


(D) Subsampled Gaussian mechanism with $\sigma = 0.5, \gamma = 0.1$

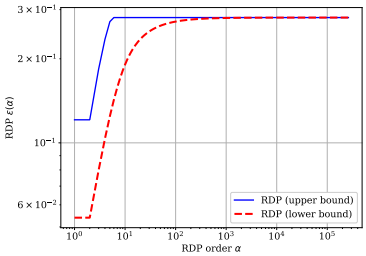
(E) Subsampled Laplace mechanism with $b = 0.5, \gamma = 0.1$

(F) Subsampled Rand. Resp. mechanism with $p = 0.9, \gamma = 0.1$

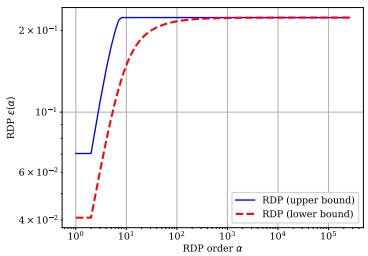
FIGURE 6. The same experiments reported in Figure 3 (the RDP as a function of α) but with $\gamma = 0.1$.



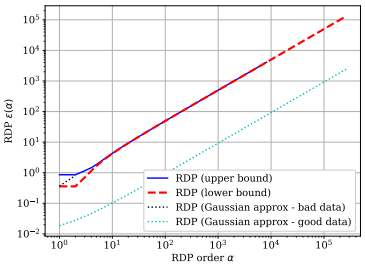
(A) Subsampled Gaussian mechanism with $\sigma = 5, \gamma = 0.5$



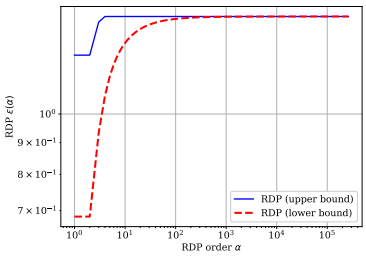
(B) Subsampled Laplace mechanism with $b = 2, \gamma = 0.5$



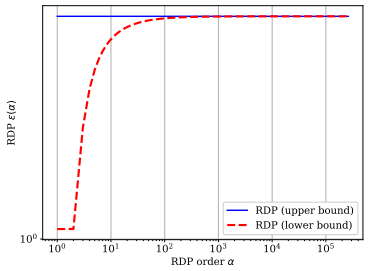
(C) Subsampled Rand. Resp. mechanism with $p = 0.6, \gamma = 0.5$



(D) Subsampled Gaussian mechanism with $\sigma = 0.5, \gamma = 0.5$

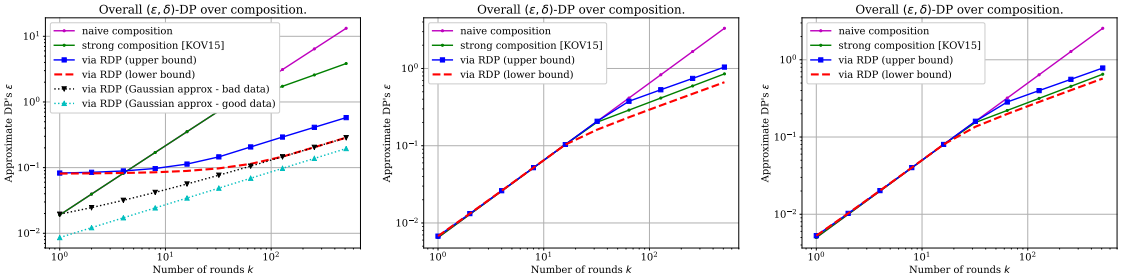


(E) Subsampled Laplace mechanism with $b = 0.5, \gamma = 0.5$



(F) Subsampled Rand. Resp. mechanism with $p = 0.9, \gamma = 0.5$

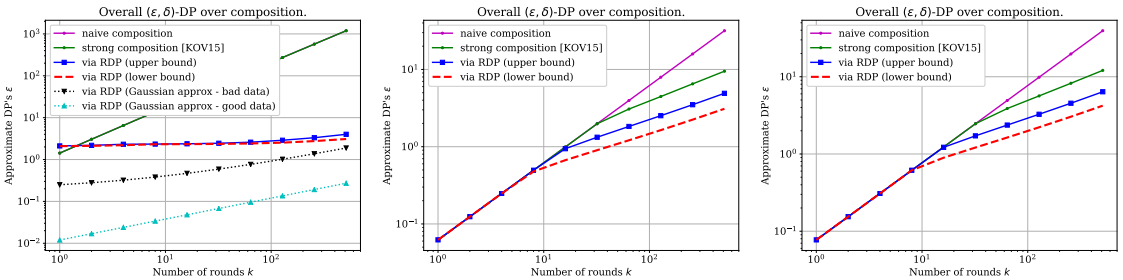
FIGURE 7. The same experiments reported in Figure 3 (the RDP as a function of α) but with $\gamma = 0.5$.



(A) Subsampled Gaussian mechanism with $\sigma = 5, \gamma = 0.01$

(B) Subsampled Laplace mechanism with $b = 2, \gamma = 0.01$

(C) Subsampled Rand. Resp. mechanism with $p = 0.6, \gamma = 0.01$

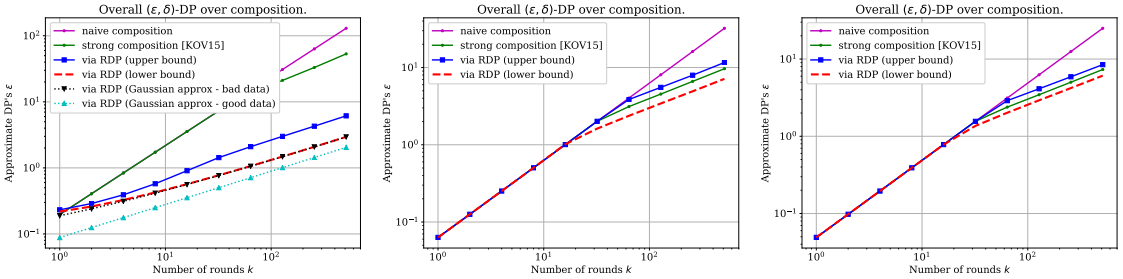


(D) Subsampled Gaussian mechanism with $\sigma = 0.5, \gamma = 0.01$

(E) Subsampled Laplace mechanism with $b = 0.5, \gamma = 0.01$

(F) Subsampled Rand. Resp. mechanism with $p = 0.9, \gamma = 0.01$

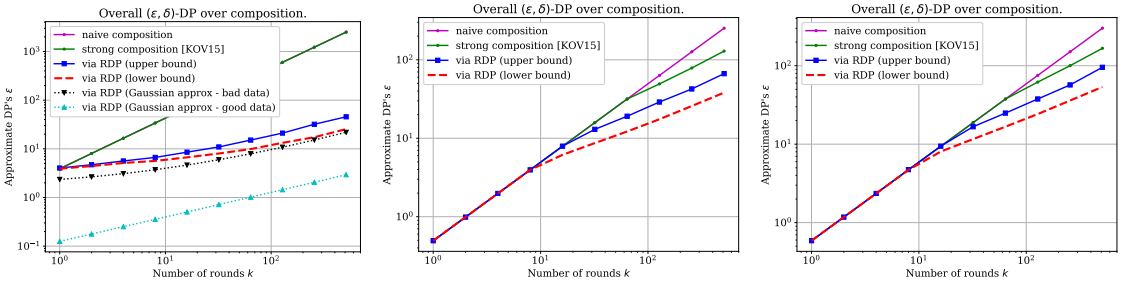
FIGURE 8. The same experiments reported in Figure 4 but with $\gamma = 0.01$.



(A) Subsampled Gaussian mechanism with $\sigma = 5, \gamma = 0.1$

(B) Subsampled Laplace mechanism with $b = 2, \gamma = 0.1$

(C) Subsampled Rand. Resp. mechanism with $p = 0.6, \gamma = 0.1$

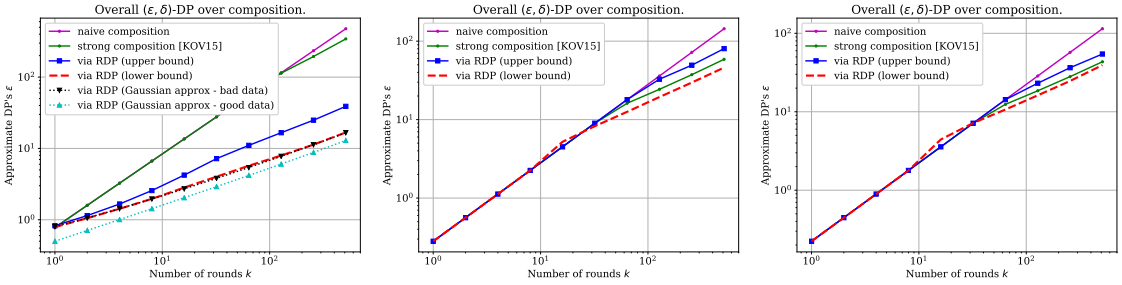


(D) Subsampled Gaussian mechanism with $\sigma = 0.5, \gamma = 0.1$

(E) Subsampled Laplace mechanism with $b = 0.5, \gamma = 0.1$

(F) Subsampled Rand. Resp. mechanism with $p = 0.9, \gamma = 0.1$

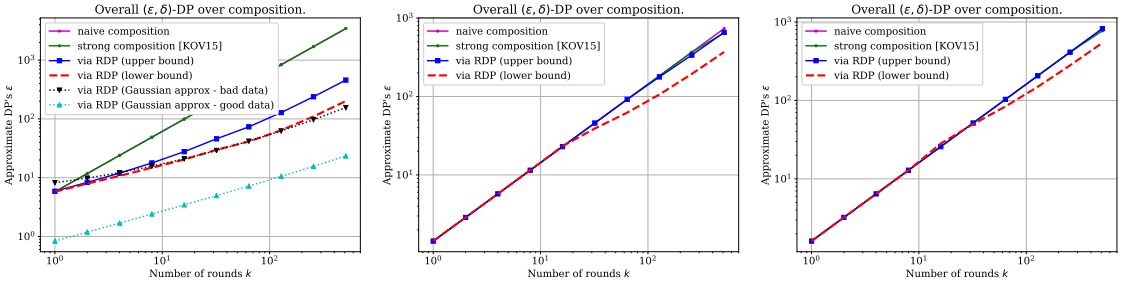
FIGURE 9. The same experiments reported in Figure 4 but with $\gamma = 0.1$.



(A) Subsampled Gaussian mechanism with $\sigma = 5, \gamma = 0.5$

(B) Subsampled Laplace mechanism with $b = 2, \gamma = 0.5$

(C) Subsampled Rand. Resp. mechanism with $p = 0.6, \gamma = 0.5$



(D) Subsampled Gaussian mechanism with $\sigma = 0.5, \gamma = 0.5$

(E) Subsampled Laplace mechanism with $b = 0.5, \gamma = 0.5$

(F) Subsampled Rand. Resp. mechanism with $p = 0.9, \gamma = 0.5$

FIGURE 10. The same experiments reported in Figure 4 but with $\gamma = 0.5$.