Subsampled Rényi Differential Privacy and Analytical Moments Accountant

Yu-Xiang Wang*
UC Santa Barbara
Santa Barbara, CA
yuxiangw@cs.ucsb.edu

Borja Balle
Amazon AI
Cambridge, UK
pigem@amazon.co.uk

Shiva Kasiviswanathan
Amazon AI
Sunnyvale, CA
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) applies 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.

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* standard of privacy definition in the academic world with also wide adoption in the 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 & 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 & Rothblum, 2016; Bun & 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:

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

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 that we uniformly randomly drawn from

*The research is partially completed while Yu-Xiang was a scientist in Amazon AI, Palo Alto.

$\{1, ..., n\}$, 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 & Roth, 2013; Balle & 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 could 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 randomized algorithm first takes a subsample of the dataset generated 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 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 the 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 data set 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 a tight 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 point α^* satisfying $\gamma\alpha^* e^{\epsilon_{\mathcal{M}}(\alpha^*)} \approx 1$. 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

[1] There are different subsampling methods, such as Poisson subsampling, sampling without replacement, sampling with replacement, etc.

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

— a *ternary* version of the Pearson-Vajda divergence (formally defined in Appendix B). 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 3.3). 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. The remarks following the statement of the theorem in Section 3.1 discuss some of the asymptotic implications of this bound, as well as its meaning in several special cases. These provide intuitive explanations justifying the tightness of the bound. In practice, however, asymptotic bounds are of limited interest: concrete bounds with explicit, tight 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"*.

2 Background and Related Work

In this section, we review some background about differential privacy, some related privacy notions, and the technique of moments accountant.

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 datapoints 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 & Rothblum, 2016; Abadi et al., 2016; Balle & Wang, 2018).

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

A common technique for achieving differential privacy while working with a real-valued function $f : \mathcal{X}^n \to \mathbb{R}$ is via 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 & Roth, 2013, Theorem 3.22).

Stochastic Gradient Descent and Subsampling Lemma. A popular way of designing differentially private machine learning models 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). 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 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:

$$K_{\mathcal{M}}(X, X', \lambda) := \log \mathbb{E}_{\theta\sim\mathcal{M}(X)}\left[e^{\lambda \log \frac{\mathcal{M}(X)(\theta)}{\mathcal{M}(X')(\theta)}}\right] = \log \mathbb{E}_{\theta\sim\mathcal{M}(X)}\left[\left(\frac{\mathcal{M}(X)(\theta)}{\mathcal{M}(X')(\theta)}\right)^\lambda\right]. \tag{2}$$

After a change of measure, this is equivalent to:

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

Two random variables have identical CGFs then they are identically distributed (almost everywhere). In other words, this function characterizes the entire distribution of the privacy loss random variable.

[4] The restriction to a scalar-valued function is intended to simplify this presentation, but is not essential.
[5] Balle & Wang (2018) show that a more complicated relation between ϵ and δ yields an if and only if statement.
[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.

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 $\mathcal{M}(X')$.

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 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 Section III of Mironov (2017). Here, we highlight two key properties that are relevant for this paper.

Lemma 5 (Adaptive Composition of RDP, Proposition 1 of (Mironov, 2017)). *If $\mathcal{M}_1$ that takes dataset as input obeys (α, ϵ_1)-RDP, and $\mathcal{M}_2$ that takes the dataset and the output of $\mathcal{M}_1$ as input 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.

Note that $\epsilon_\mathcal{M}(\alpha)$ is equivalent to a data-independent upper bound of the CGF (as defined in (2)),

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

up to a scaling transformation (with $\alpha = \lambda + 1$) as noted by the following remark.

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

The idea of moments accountant (Abadi et al., 2016) is to essentially keep track of the evaluations of CGF at a list of fixed locations through Lemma 5 and then Lemma 6 allows one to find the smallest ϵ given a desired δ or vice versa using:

$$\delta \Rightarrow \epsilon : \qquad \epsilon(\delta) = \min_\lambda \frac{\log(1/\delta) + K_\mathcal{M}(\lambda)}{\lambda}, \tag{3}$$

$$\epsilon \Rightarrow \delta : \qquad \delta(\epsilon) = \min_\lambda e^{K_\mathcal{M}(\lambda) - \lambda\epsilon}. \tag{4}$$

Using the convexity of CGF $K_\mathcal{M}(\lambda)$ and monotonicity of $K_\mathcal{M}(\lambda)/\lambda$ in λ (Van Erven & Harremos, 2014, Corollary 2, Theorem 3), we observe that the optimization problem in (4) is log-convex and the optimization problem (3) is unimodal/quasi-convex. Therefore, the optimization problem in (3) (similarly, in (4)) 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 (3) (similarly, (4)). 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 Appendix G).

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 H.

Other Related Work. A closely related notion to RDP is that of *zero-concentrated differential privacy* (zCDP) introduced in (Bun & Steinke, 2016) (see also (Dwork & 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 $\lambda\xi + \lambda(\lambda+1)\rho \geq K_{\mathcal{M}}(\lambda)$.*

In general, the RDP view of privacy is broader than the CDP view as it captures finer information. 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. While this independent work attempts to solve a problem closely related to ours, they are not directly comparable in that they deal with the amplification properties of tCDP while we deal with that of Rényi DP (and therefore CDP without truncation). A simple consequence of this difference is that the popular subsampled Gaussian mechanism explained above, that is covered by our analysis, is not directly covered by the amplification properties of tCDP.

3 Our Results

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. Based on these bounds, in Section 3.3, we discuss an idea for implementing a data structure that can efficiently track privacy parameters under composition.

3.1 "Privacy Amplification" for RDP

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 subsample procedure. Missing details from this Section are collected in Appendix B.

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: subsample without replacement m datapoints of the dataset (sampling parameter $\gamma = m/n$), and (2) apply $\mathcal{M}$: a randomized algorithm taking the subsampled dataset as the input. For all integers $\alpha \geq 2$, if $\mathcal{M}$ obeys $(\alpha, \epsilon(\alpha))$-RDP, then this new randomized algorithm $\mathcal{M} \circ \mathsf{subsample}$ obeys $(\alpha, \epsilon'(\alpha))$-RDP where,*

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

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 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 Appendix G. The phase transition behavior of this bound, noted in the introduction, is probably most easily observed through Figure 1 (Section 4), where

we empirically illustrates the behavior of this bound for the commonly used subsampled mechanisms. Now before discussing the proof idea, we mention 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 I). 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 which becomes negligible as $\epsilon(\infty)$ gets smaller.

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 Appendix B.5 (see Theorem 27), 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, with 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". When σ^2 is large, $4(e^{\epsilon(2)}-1)=4(e^{1/\sigma^2}-1)\leq 8/\sigma^2$ 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 Appendix C). 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, we can use the relation between RDF and CGF mentioned in Remark 7, and the fact that CGF is a convex function (see Lemma 36 in Appendix H). 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, $K_{\mathcal{M}}(\lambda)\leq(1-\lambda+\lfloor\lambda\rfloor)K_{\mathcal{M}}(\lfloor\lambda\rfloor)+(\lambda-\lfloor\lambda\rfloor)K_{\mathcal{M}}(\lceil\lambda\rceil)$.*

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

Proof Idea The proof of this theorem is roughly split into three parts (see Appendix B.1). 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 16, Appendix B.1). 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 17, 18, 19, Appendix B.1). 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.

3.2 A lower bound of 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.

Proposition 11. *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 $\Pr[\mathsf{subsample}(X) = [x, x, ..., x, x'] = \gamma$. 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 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 $x/(1+x) \leq \log(1+x) \leq x$ 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 4.

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, 2018). The only difference from the standard RDP is that now ϵ is parameterized by a pair of fixed adjacent datasets. This point is in illustrated in Appendix C, where we discuss an asymptotic approximation of the Rényi divergence for the subsampled Gaussian mechanism.

3.3 Analytical Moments Accountant

Our theoretical results above allow us to build an analytical moments accountant for composing differentially private mechanisms. This is a data structure that 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 (3) (or (4)).

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 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 with $O(1)$ 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 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 takes only $O(\log(\alpha))$ evaluations of $\epsilon_{\mathcal{M}}(\cdot)$. More details about our analytical moments accountant and substantiations to the above claims are provided in Appendix G.

4 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) :

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

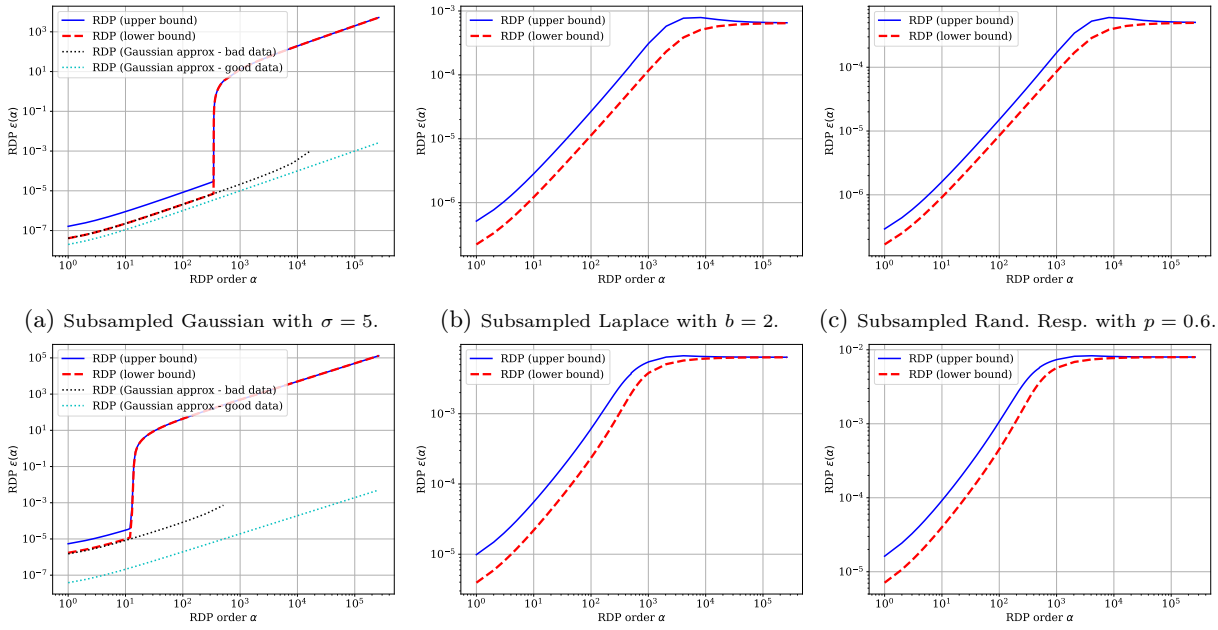
Here σ^2 represents the variance of the Gaussian perturbation, $2b^2$ the variance of the Laplace perturbation, and p the probability of replying truthfully in randomized response. We considered two groups of parameters σ, b, p for the three base mechanisms $\mathcal{M}$.

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.

In Figure 1, 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 1a and 1d, 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 $\epsilon(\infty)$.



(a) Subsampled Gaussian with $\sigma = 5$. (b) Subsampled Laplace with $b = 2$. (c) Subsampled Rand. Resp. with $p = 0.6$.

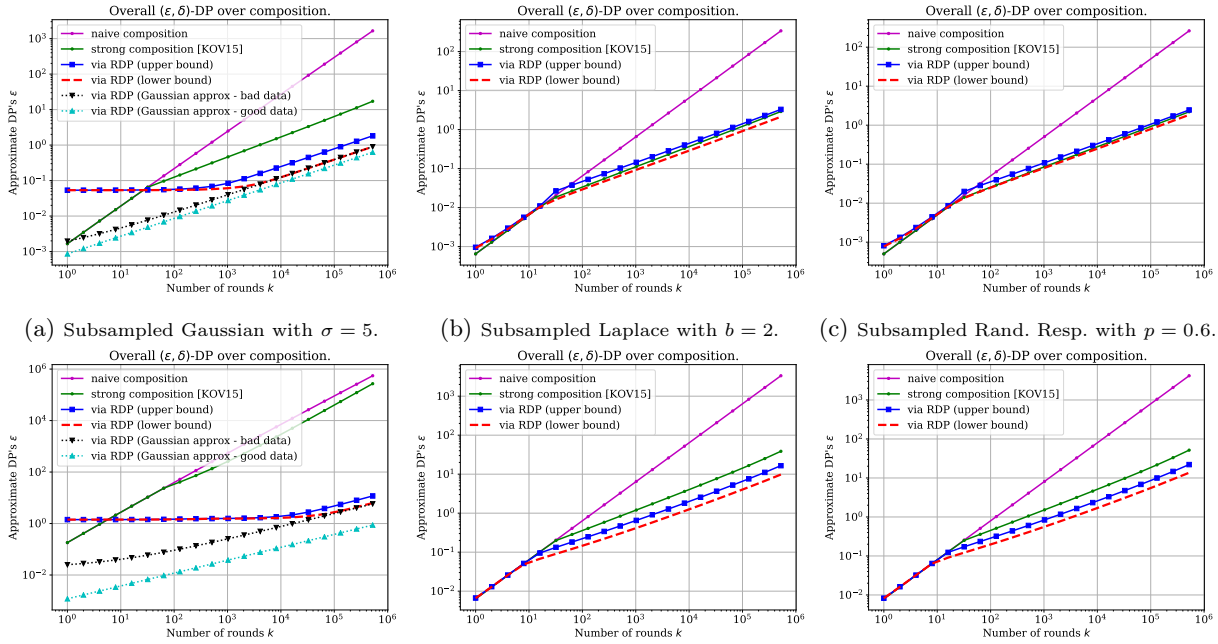
(d) Subsampled Gaussian with $\sigma = 0.5$. (e) Subsampled Laplace with $b = 0.5$. (f) Subsampled Rand. Resp. with $p = 0.9$

Figure 1: 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 11 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 Appendix C.

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 Appendix C. The asymptotic Gaussian approximation with the "bad" data in Example 28 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 2, we plot the overall (ϵ, δ)-DP for $\delta = 1e{-}8$ 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 ccountant. 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. (2015) obeys that $\delta < 1e-8$ 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 (3).

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



(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 2: 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 $\delta = 1e-8$ (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., 2015) 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 11. 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 Appendix C.

analysis of the Gaussian mechanism (Balle & 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 & 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 2a) and five orders of magnitude smaller ϵ in the low privacy regime (see Figure 2d).

The results for composing subsampled Laplace mechanisms and subsampled randomized response mechanisms are shown in Figures 2b, 2c, 2e, and 2f. 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.

5 Conclusion

In this paper, we have 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 data set 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 efficiently conversion of RDP to (ϵ, δ)-DP for any δ (or ϵ) of choice. These results substantially expands 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 the new approach with subsampling-based sublinear algorithms for exploratory data analysis.

Acknowledgment

The authors 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, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS-16)*, (pp. 308–318). ACM.

Apple, D. (2017). Learning with privacy at scale. *Apple Machine Learning Journal*.

Balle, B., Barthe, G., & Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *NIPS*.

Balle, B., & Wang, Y.-X. (2018). Improving gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. *International Conference in Machine Learning (ICML)*.

Bassily, R., Smith, A., & Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS-14)*, (pp. 464–473). IEEE.

Beimel, A., Nissim, K., & Stemmer, U. (2013). Characterizing the sample complexity of private learners. In *Conference on Innovations in Theoretical Computer Science (ITCS-13)*, (pp. 97–110). ACM.

Bernard, T. S., Hsu, T., Perlroth, N., & Lieber, R. (2017). Equifax says cyberattack may have affected 143 million in the us. *The New York Times, Sept*, *7*.

Bobkov, S., Chistyakov, G., & Götze, F. (2016). R\'enyi divergence and the central limit theorem. *arXiv preprint arXiv:1608.01805*.

Bun, M., Dwork, C., Rothblum, G. N., & Steinke, T. (2018). Composable and versatile privacy via truncated cdp. In *to appear in STOC-18*.

Bun, M., Nissim, K., Stemmer, U., & Vadhan, S. (2015). Differentially private release and learning of threshold functions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, (pp. 634–649). IEEE.

Bun, M., & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, (pp. 635–658). Springer.

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The Guardian*, *17*.

Dajani, A., Lauger, A., Singer, P., Kifer, D., Reiter, J., Machanavajjhala, A., Garfinkel, S., Dahl, S., Graham, M., Karwa, V., Kim, H., Leclerc, P., Schmutte, I., Sexton, W., Vilhuber, L., & Abowd, J. (2017). The modernization of statistical disclosure limitation at the u.s. census bureau. *Census Scientific Advisory Commitee Meetings*.
URL https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., & Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *International Conference on the Theory and Applications of Cryptographic Techniques*, (pp. 486–503). Springer.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, (pp. 265–284). Springer.

Dwork, C., & Roth, A. (2013). The algorithmic foundations of differential privacy. *Theoretical Computer Science*, *9*(3-4), 211–407.

Dwork, C., & Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

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

Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, (pp. 1054–1067). ACM.

European Parliament, & Council of the European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*.

Foulds, J., Geumlek, J., Welling, M., & Chaudhuri, K. (2016). On the theory and practice of privacy-preserving bayesian data analysis. In *Conference on Uncertainty in Artificial Intelligence (UAI-16)*, (pp. 192–201). AUAI Press.

Geumlek, J., Song, S., & Chaudhuri, K. (2017). Renyi differential privacy mechanisms for posterior sampling. In *Advances in Neural Information Processing Systems*, (pp. 5295–5304).

Gil, M., Alajaji, F., & Linder, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, *249*, 124–131.

Kairouz, P., Oh, S., & Viswanath, P. (2015). The composition theorem for differential privacy. In *International Conference on Machine Learning (ICML-15)*.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., & Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing*, *40*(3), 793–826.

Lukacs, E. (1970). *Characteristic functions*. Griffin.

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

Murtagh, J., & Vadhan, S. (2016). The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, (pp. 157–175). Springer.

Nielsen, F., & Nock, R. (2014). On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, *21*(1), 10–13.

Song, S., Chaudhuri, K., & Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. In *Conference on Signal and Information Processing*.

Sweeney, L. (2015). Only you, your doctor, and many others may know. *Technology Science*, *2015092903*.

Uber Security (2017). Uber releases open source project for differential privacy. https://medium.com/uber-security-privacy/differential-privacy-open-source-7892c82c42b6.

Ullman, J. (2017). Cs7880: Rigorous approaches to data privacy, spring 2017. http://www.ccs.neu.edu/home/jullman/PrivacyS17/HW1sol.pdf.

Vajda, I. (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, T., & Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, *60*(7), 3797–3820.

Wang, Y.-X. (2018). Per-instance differential privacy. *Journal of Confidentiality and Privacy, to appear*.

Wang, Y.-X., Fienberg, S., & Smola, A. (2015). Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning (ICML-15)*, (pp. 2493–2502).

Wang, Y.-X., Lei, J., & Fienberg, S. E. (2016). Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *Journal of Machine Learning Research*, *17*(183), 1–40.

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 12 (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 13 (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. (2015) 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 & Vadhan, 2016).

B Proofs and Missing Details from Section 3.1

In this section, we fill in the missing details and proofs from Section 3.1. We first define a few quantities needed to establish our results.

Pearson-Vajda Divergence and the Moments of Linearized Privacy Random Variable. The Pearson-Vajda Divergence (or $|\chi|^\alpha$-divergence) of order α is defined as follows (Vajda, 1973):

$$D_{|\chi|^\alpha}(p\|q) := \mathbb{E}_q\left[\left|\frac{p}{q} - 1\right|^\alpha\right]. \tag{5}$$

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 derivate of the

MGF evaluated at 0:

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

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

$$\mathbb{E}[(p/q-1)^\alpha] = \Delta^{(\alpha)}[e^{K_{\mathcal{M}}(\cdot)}](0). \tag{6}$$

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

$$\Delta^{(\alpha)} := \underbrace{\Delta \circ ... \circ \Delta}_{\alpha\text{-times}}, \tag{7}$$

where Δ denote 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 D for more information on $\Delta^{(\alpha)}$ and its connection to binomial numbers.

B.1 A Sketch of the Proof of Theorem 9

In this section, we present a sketch of the proof of our main theorem. The arguments are divided into three parts. In the first part, we define a new family of privacy definitions called *ternary-$|\chi|^\alpha$-differential privacy* and show that it handles subsampling naturally. In the second part, we bound the Rényi DP using the ternary-$|\chi|^\alpha$-differential privacy and apply their subsampling lemma. In the third part, we propose several different ways of converting the expression stated as ternary-$|\chi|^\alpha$-differential privacy back to that of RDP, hence giving rise to the stated results in the remarks following Theorem 9.

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[8], we define

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

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

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

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 (5)).

Definition 15 (Binary-$|\chi|^\alpha$-differential privacy)**.** *We say that a randomized mechanism $\mathcal{M}$ is ξ-binary-$|\chi|^\alpha$-DP if for all $\alpha \geq 1$:*

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

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 F.

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:

[8] 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''.

Proposition 16 (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.*

The entire proof is presented in Appendix B.2. The key idea involves using 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 intractable ternary-$|\chi|^\alpha$-DP of a mixture distribution to that of three simple distributions that come from 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:

$$\mathbb{E}_q\left[\left(\frac{p}{q}\right)^\alpha\right] = 1 + \binom{\alpha}{1}\mathbb{E}_q\left[\frac{p}{q} - 1\right] + \sum_{j=2}^{\alpha}\binom{\alpha}{j}\mathbb{E}_q\left[\left(\frac{p}{q} - 1\right)^j\right]. \tag{8}$$

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,

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

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 16. Putting these together, we can bound (8) as

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

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,

$$\epsilon_{\mathcal{M}\circ\mathsf{subsample}}(\alpha) \leq \frac{1}{\alpha - 1}\log\left(1 + \sum_{j=2}^{\alpha}\binom{\alpha}{j}\gamma^j\zeta(j)^j\right), \tag{9}$$

Part 3: Bounding Ternary-$|\chi|^\alpha$-DP using RDP. It remains to bound $\zeta(j)^j := \sup_{p,q,r}\mathbb{E}_r[\frac{|p-q|^j}{r^j}]$ using RDP. We provide several ways of doing so and plugging them into (9) show how the various terms in the bound of Theorem 9 arise. Missing proofs are presented in Appendix B.3.

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

For the special case of $j = 2$, 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 17 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 18. *Let X, Y be nonnegative random variables, for any $j \geq 1$*

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

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 have a 2^j factor rather than 2 (see e.g., Proposition 3.2 of Bobkov et al. (2016)).
An alternative bound that is tighter in cases when X and Y is related to each other with a multiplicative bound. Note that this bound is only going to be useful when $\mathcal{M}$ has a bounded $\epsilon(\infty)$, such as when $\mathcal{M}$ satisfies $(\epsilon, 0)$-DP guarantee.

Lemma 19. *Let X, Y be nonnegative random variables and with probability 1, $e^{-\varepsilon}Y \le X \le e^{\varepsilon}Y$. Then for any $j \ge 1$*

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

Take $X = p/r$ and $Y = q/r$. Applying Lemma 18 gives $\zeta(j) \le 2e^{(j-1)\epsilon(j)}$. Using Lemma 19 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 of,

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

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

B.2 Proof of the Subsampling Lemma for Ternary-$|\chi|^\alpha$-DP

In this section, we prove Proposition 16. The proof uses the following simple lemma.

Lemma 20. *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 21 (Proposition 16 Restated)**.** *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.

$$\begin{aligned} D_{|\chi|^j}(p, q\|r) &= \mathbb{E}_r\left[\left(\frac{|p - q|}{r}\right)^j\right] = \gamma^j \mathbb{E}_r\left[\left(\frac{|p(\cdot|E) - q(\cdot|E)|}{r}\right)^j\right] \\ &= \gamma^j D_{|\chi|^j}(p(\cdot|E), q(\cdot|E)\|r). \end{aligned} \tag{10}$$

Note that $p(\cdot|E), q(\cdot|E)$ and r are mixture distributions with combinatorially 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 to everything else, so it is clear that $r(\theta|J) = r(\theta|J, i)$. In other words,

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

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

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

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. Let us explain the rationale behind them. Note that mixture distributions $p(\theta|E), q(\theta|E)$ have a different number of mixture components comparing to $q(\theta)$. $q(\theta)$ 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 σ-field to redundantly represent both subsampling over $[n-1]$ and $[n]$ under the same uniform probability measure 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:

$$\begin{aligned} \mathbb{E}_q \left(\frac{|p(\theta|E) - q(\theta|E)|}{q(\theta)} \right)^j &= \int \frac{[p(\theta|E) - q(\theta|E)]^j}{q(\theta)^{j-1}} d\theta \\ &\underset{\text{Jensen}}{\uparrow} \leq \int \mathbb{E}_{J,i} \left[\frac{|g(J,i) - g'(J,i)|^j}{q(\theta|J)^{j-1}} \right] d\theta \\ &\underset{\text{Fubini}}{\uparrow} = \mathbb{E}_{J,i} \mathbb{E}_q \left[\left(\frac{|g(J,i) - g'(J,i)|}{q(\theta|J)} \right)^j \middle| J, i \right] \leq \zeta(j)^j. \end{aligned} \tag{11}$$

The second but last line uses Jensen's inequality and Lemma 20, which proves the joint convexity of 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. Combining (10) with (11) gives the claimed result because the definitions of g and g' ensure that each inner expectation is a ternary Liese–Vajda divergence of the original mechanism on a triple of mutually adjacent datasets. □

B.3 Missing Proofs on Bounding Ternary-$|\chi|^\alpha$-DP using RDP

Lemma 22 (Lemma 17 Restated)**.** *If a randomized mechanism $\mathcal{M}$ is ξ-binary-$|\chi|^\alpha$-DP, then it is ζ-ternary-$|\chi|^\alpha$-DP for some ζ satisfying $\xi(\alpha)^\alpha \leq \zeta(\alpha)^\alpha \leq 4\xi(\alpha)^\alpha$.*

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

$$\{x|p \geq q, q \geq r\}, \{x|p \geq q, q < r\}, \{x|p < q, p \geq r\}, \{x|p < q, p \geq r\}$$

Under the first event $|p-q|^j/r^{j-1} = (p-q)^j/r^{j-1} \leq (p-r)^j/r^{j-1}$. Under the second event $|p-q|^j/r^{j-1} \leq (p-q)^j/q^j$.Similarly, under the third and fourth event, $|p-q|^j/r^{j-1}$ is bounded by $(q-r)^j/r^{j-1}$ and $(q-p)^j/p^{j-1}$ respectively. It then follows that:

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

□

Lemma 23 (Lemma 18 Restated). *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 $X, Y \geq 0$

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

□

Lemma 24 (Lemma 19 Restated). *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. □

B.4 Proof of Corollary 10

Corollary 25 (Corollary 10 Restated). *Let $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denotes the floor and ceiling operators*

$$K_{\mathcal{M}}(\lambda) \leq (1 - \lambda + \lfloor \lambda \rfloor) K_{\mathcal{M}}(\lfloor \lambda \rfloor) + (\lambda - \lfloor \lambda \rfloor) K_{\mathcal{M}}(\lceil \lambda \rceil).$$

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. □

B.5 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 $d(X, X') \leq 1$ to represent neighboring datasets. We start with some additional conditions on the mechanism $\mathcal{M}$ as defined below.

Definition 26 (Tightness and Self-consistency). *We say a mechanism $\mathcal{M}$ and its corresponding RDP privacy guarantee $\epsilon_{\mathcal{M}}(\cdot)$ are* tight *if $\max_{X, X' : d(X, X') \leq 1} D_\ell(\mathcal{M}(X) \| \mathcal{M}(X')) = \epsilon_{\mathcal{M}}(\ell)$ for every $\ell = 1, 2, 3, \ldots$ We say that a tight pair $(\mathcal{M}, \epsilon_{\mathcal{M}}(\cdot))$ is* self-consistent *with respect to $|\chi|^\alpha$-divergence, if*

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

The tightness condition requires that the RDP function $\epsilon_{\mathcal{M}}(\cdot)$ to 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 E 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 (7)) of the functional $e^{(\cdot - 1)\epsilon(\cdot)}$ evaluated at 0.

Theorem 27 (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 $m \leq n$, let the randomized algorithm $\mathcal{M} \circ \mathsf{subsample}$ be defined as: (1) $\mathsf{subsample}$: subsample without replacement m datapoints of the dataset (sampling parameter $\gamma = m/n$), and (2) apply $\mathcal{M}$: a randomized algorithm taking the subsampled dataset as the input. If $\mathcal{M}$ obeys $(\alpha, \epsilon(\alpha))$-RDP and additionally the RDP guarantee is tight and $(\alpha+1)$-self-consistent as per Definition 26, then for all integer $\alpha \geq 2$, this new randomized algorithm $\mathcal{M} \circ \mathsf{subsample}$ obeys $(\alpha, \epsilon'(\alpha))$-RDP where,*

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

Proof Idea. The proof is identical to that of Theorem 9 as laid out in Appendix B.1. 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 26, we know that there exist a pair of adjacent data sets, 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 & Nock, 2014) using a binomial expansion. Using Lemma 17 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 term 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 Appendix B.1, we have an alternative way of bounding $\zeta(2)$ that does not require these additional assumptions.

C Asymptotic Approximation of Rényi Divergence for Subsampled Gaussian Mechanism

In this section, we present 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 4.

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. Without loss of generality, we assume that $f(x_i) \leq 1/2$, which implies that the global sensitivity of μ is $1/|J|$. By the sampling without replacement version of the central limit theorem[9], $\sqrt{|J|}(\mu(X_J) - \frac{1}{n}\sum_{i=1}^n f(x_i))$ converges in distribution to $\mathcal{N}(0, \frac{1}{n}\sum_{i=1}^n (f(x_i)-\mu(X))^2)$. In other words, the distribution of θ asymptotically converges to

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

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

[9] Under boundedness of $f(x_i)$, the regularity conditions holds.

We disclaim that this is a truly asymptotic approximation and should only be true when $|J|, n \to \infty$ and $\gamma = |J|/n \to 0$, but it is nevertheless interesting as it allows us to understand the dependence of different parameters in the bound. 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.

Example 28 (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 $p = \mathcal{N}(-\frac{1}{2} + \frac{1}{n}, \frac{n-1}{n^2|J|} + \frac{\sigma^2}{|J|^2})$ and $q = \mathcal{N}(-\frac{1}{2}, \frac{\sigma^2}{|J|^2})$, and the corresponding Rényi divergence equals*

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

Example 29 (Gaussian approximation - a "good" data case)**.** *Let n be an odd number, and let X' be such that $f(x_i) = 1/2$ for $i \leq \lfloor n/2 \rfloor$ 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 $p = \mathcal{N}(\frac{1}{2n}, \frac{\sigma^2}{|J|^2} + \frac{1}{4|J|} - \frac{1}{4n^2|J|})$ and $q = \mathcal{N}(-\frac{1}{2n}, \frac{\sigma^2}{|J|^2} + \frac{1}{4|J|} - \frac{1}{4n^2|J|})$, and the corresponding Rényi divergence equals*

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

The first example (a "bad" data case) is closely related to our construction in the proof of Proposition 11. For $\alpha \ll \sigma^2/\gamma$, the example shows an $O(\alpha\gamma^2/\sigma^2)$ rate, matching our upper bound from Theorem 9 (see Remark "Bound under Additional Assumptions" in Section 3.1) in the small α, large σ regime. The second example corresponds to a "good" data case where the dataset has a variety of different datapoints, and as we can see, the variance of the asymptotic distribution 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.

D 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 (6) 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.

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

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 Appendix G we will describe 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.

E 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 have 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 30. *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 $0 \leq \mu \leq 1$

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

Take derivative w.r.t. μ, we get

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

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.

F Other Properties of Ternary-$|\chi|^\alpha$-DP

When $\alpha = 1$, both the binary- and ternary-$|\chi|^\alpha$-divergence reduces to the total variation distance. When $\alpha = 2$ the binary-$|\chi|^\alpha$-divergence become 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 31 ($|\chi|^\alpha$-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,

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

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 19 that we used in the proof of Theorem 9.

Lemma 32. *Let p, q, r are three distributions. For all conjugate pair $u, v \geq 1$ such that $1/u + 1/v = 1$, and all integer $j \geq 2$ we have that*

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

Proof. The proof is a straightforward application of the 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 33. *When we take $v = \infty$ and $u = 1$, we recover the result from Lemma 19. 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 $\mathbb{E}_q[\log(p/q)^{2j}]$.

G Analytical Moments Accountant and Numerically Stable Computation

In this section, we provide more details on the *analytical moments accountant* that we described briefly in Section 3.3. Recall that the analytical moments accountant is a data structure that one can attach to a dataset to keep track of the privacy loss over a sequence of differentially private data accesses. The data structure caches the CGF of the privacy random variables in symbolic form and permits efficient (ϵ, δ)-DP calculations for any desired δ or ϵ. Here is how it works.

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 the $K_{\mathcal{M}_1}, ..., K_{\mathcal{M}_k}$ be the corresponding CGF. 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 usage 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 we keep track of separately. The conversion to (ϵ, δ)-DP is obtained by solving the univariate optimization problems described in (3) and (4).

We note that our analytical moments accountant is conceptually the same as the moments accountant used by Abadi et al. (2016) and the RDP composition used by Mironov (2017). Both prior work however considered only a predefined discrete list of orders λ (or α's). Our main difference is that, for every mechanism, we keep track of the CGF for all $\lambda \in \mathbb{R}_+$ at the same time.

In the remainder of the section, we will describe specific designs of this data structure and substantiate our claims described earlier in Section 3.3.

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

Proposition 34. *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 in operation, let the total number of unique mechanisms that it has seen so far be L. Then the analytical moments accountant takes $O(L)$ space . The CGF queries (at a given λ) takes time $O(L)$. (ϵ, δ)-DP query to accuracy τ (in terms of absolute difference in the argument $|\lambda - \lambda^*|$) takes time $O(L)$ and $O(L \log(\lambda^*)/\tau)$ CGF evaluation calls respectively, where λ^* is the corresponding minimizer in (3) or (4).*

Proof. We keep track of a dictionary of λ 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 zeroth 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 H.

When solving for ϵ given δ, 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 that guarantees to contain λ^* in $O(\log \lambda^*)$ time thanks to the monotonicity of RDP. Then we do bisection to find the optimal λ^*, using the unimodal property of the objective function. Note that $\lambda_{\max} \leq 2\lambda^*$. This ensures that the 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 (4) 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 an 1-suboptimal 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 could be an upper bound which may 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 for 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 dealing them naïvely will cause floating point numbers overflow or underflow. 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 H, however, we can "project" it to another valid upper bound using the proposition below so that it satisfies the properties from Appendix H.

Proposition 35. *Let $\bar{K}_{\mathcal{M}}$ be an upper bound of $K_{\mathcal{M}}$, there is a functional F such that $F[\bar{K}_{\mathcal{M}}] \leq K_{\mathcal{M}}$ 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 $\lambda \geq 0$.*

Proof. We prove 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 $\beta\lambda \leq g(\lambda),\ \forall\lambda$. Let the smallest λ such that $g(\lambda) = \beta\lambda$ be $\tilde{\lambda}$. Then, we define

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

Clearly, this is the largest function that satisfy 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 (3) and (4) remain unchanged.

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

Noting that the expression 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 $|\mathcal{X}|^j$-privacy of order j using RDP.

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 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.

H Properties of Cumulant Generating Functions 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 Appendix G.

Lemma 36. *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 is not a single point mass).

(d) $K_{\mathcal{M}}(0) = 0$, e.g., 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 & 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 $\lambda = (1-v)\lambda_1 + v\lambda_2)/2$ and apply Hölder's inequality with the exponents being the conjugate pair $1/(1-v)$ 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

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

and the proof is complete. □

Corollary 37. *Optimization problem (4) is log-convex. Optimization problem (3) is unimodal / quasi-convex.*

Proof. To see the first claim, check that the logarithm of (4) 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 $\frac{K_{\mathcal{M}}(\lambda)}{\lambda}$ (this is RDP!) 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

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

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

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

and that $K_{\mathcal{M}}(0) = 0$. □

The corollary implies that optimization problems defined in (3) and (4) 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.

I Rényi Divergence of Exponential Family Distributions and RDP

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

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

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

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

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 $\alpha = 1$,

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

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)

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

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

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

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

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

be the corresponding distribution when the dataset is X'.

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

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

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 data sets 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 data set 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 data set 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 38 (RDP of exponential family mechanisms)**.** *Let $\mathcal{M}$ is 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 39. *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 38. Assumption (A) implies that $\|\eta(X)-\eta(X')\| \le \Delta$. Note that for all $\alpha \le \kappa/\Delta$, $\|\alpha\eta(X)+(1-\alpha)\eta(X')-\eta(X)\| \le \kappa$. 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

$$D_\alpha(p\|q) \le \frac{1}{\alpha-1}[A(\eta(X)) + (\alpha-1)B\Delta - A(\eta(X)) + (\alpha-1)B\Delta] = 2B\Delta. \tag{12}$$

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

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

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

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

which, together with (12), produces the bound as claimed. □