Poisson Subsampled Renyi Differential Privacy

Yuqing Zhu [1] **Yu-Xiang Wang** [1]

Abstract

We consider the problem of "privacy-amplification by subsampling" under the Renyi Differential Privacy (RDP) framework (Mironov, 2017). This is the main workhorse underlying the moments accountant approach for differentially private deep learning (Abadi et al., 2016). Complementing a recent result on this problem that deals with "Sampling without Replacement" (Wang et al., 2019), we address the "Poisson subsampling" scheme which selects each data point independently with probability γ. The seemingly minor change allows us to more precisely characterize the RDP of $\mathcal{M} \circ \mathsf{PoissonSample}$. In particular, we prove an exact analytical formula for the case when $\mathcal{M}$ is the Gaussian mechanism or the Laplace mechanism. For general $\mathcal{M}$, we prove an upper bound that is optimal up to an additive constant of $\log(3)/(\alpha-1)$ and a multiplicative factor of $1+O(\gamma)$. Our result is the first of its kind that makes the moments accountant technique (Abadi et al., 2016) efficient and generally applicable for all Poisson-subsampled mechanisms. An open source implementation is available at `https://github.com/yuxiangw/autodp`.

1. Introduction

"Privacy-amplification by Subsampling" and the Renyi Differential Privacy are the two fundamental techniques that have been driving many exciting recent advances in differentially private learning (Abadi et al., 2016; Park et al., 2016; Papernot et al., 2018; McMahan et al., 2018).

One prominent use case of both techniques is the NoisySGD algorithm (Song et al., 2013; Bassily et al., 2014; Wang et al., 2015; Foulds et al., 2016; Abadi et al., 2016) for differentially private deep learning. NoisySGD iteratively updates the model parameters as follows:

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

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, $\mathcal{I} \subset [n]$ is a randomly subsampled index set and $Z_t \sim \mathcal{N}(0, \sigma^2 I)$. When $\nabla f_i(\theta_t)$ is bounded (or clipped) in ℓ_2-norm, the Gaussian noise-adding procedure is known to ensure (ϵ, δ)-DP for this iteration. ϵ, δ are nonnegative numbers that quantifies the privacy loss incurred from running the algorithm (the smaller the better). But this is clearly not good enough as it takes many iterations to learn the model, and the privacy guarantee deteriorates as the algorithm continues. This is where the "privacy-amplification" and RDP become useful.

The principle of "privacy-amplification by subsampling" works seamlessly with NoisySGD as it allows us to exploit the randomness in choosing the minibatch $\mathcal{I}$ for the interest of a stronger privacy guarantee. Roughly speaking, if the minibatch $\mathcal{I}$ is obtained by selecting each data point with probability γ, then we can "amplify" the privacy guarantee to a stronger $(O(\gamma\epsilon), \gamma\delta)$-DP.

The RDP framework provides a complementary set of benefits that reduce the overall privacy loss over the multiple iterations we run NoisySGD. Notice that the vanilla "strong-composition" is stated for any (ϵ, δ)-DP algorithm. By using the moments accountant techniques (Abadi et al., 2016) that keep track of the RDP of a specific algorithm — subsampled-Gaussian mechanism, one can hope to more efficiently use the privacy budget than what an optimal algorithm would be able to using only (ϵ, δ)-DP (Kairouz et al., 2015).

In general, however, calculating the RDP for the procedure that first subsamples the data set then apply a randomized mechanism $\mathcal{M}$ is highly non-trivial. An exact analytical formula is not known even for the widely-used subsampled-Gaussian mechanism. Existing asymptotic bounds are typically off by a constant, and only apply to a restricted subset of the parameter regimes. To get the most mileage out of the moments accountant, practitioners often resort to numerical integration which calculates and keep track of a

[1]UC Santa Barbara, Department of Computer Science. Correspondence to: Yuqing Zhu <yuqingzhu@ucsb.edu>, Yu-Xiang Wang <yuxiangw@cs.ucsb.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

fixed list of RDP values (Abadi et al., 2016; Park et al., 2016).

Wang et al. (2019) took a first stab at this problem and provided a general "RDP-amplification" bound that applies to any $\mathcal{M}$. Their result, however, is still a constant factor away from being optimal. A more subtle difference is that Wang et al. (2019) considered "Subsampling without Replacements" — finding a random subset of size m at random — rather than the "Poisson subsampling" that was used by Abadi et al. (2016), which includes each data points independently at random with probability γ. The difference is substantial enough that it introduces several new technical hurdles.

In this paper, we provide the first general result of "privacy-amplification" of RDP via Poisson subsampling. Our main contributions are the following.

1. First, we prove a nearly optimal upper bound on the RDP of $\mathcal{M} \circ \mathsf{PoissonSample}$ as a function of the sampling probability γ, RDP order α, and the RDP of $\mathcal{M}$ up to α. The bound matches a lower bound up to an additive factor of $\log(3)/(\alpha-1)$, where α is the order of RDP. When α is small relative to $1/\gamma$ with γ being the sampling probability, our upper bound is optimal up to a multiplicative factor of $1 + O(\gamma\alpha e^{\epsilon(\alpha)})$. The result tightens and generalizes Lemma 3 of (Abadi et al., 2016), which addresses only the case when $\mathcal{M}$ is Gaussian mechanism and applies only to the cases when γ is very small.

2. Second, we identify a novel condition on the odd order Pearson-Vajda χ^α-Divergences under which we can exactly attain the lower bound. We show that Gaussian mechanism and Laplace mechanism fall under this category, but there exists $\mathcal{M}$ that samples from an exponential family distribution where the condition is false and the lower bound is not attainable. Practically, our analytical characterization simplifies the moments accountant approach for differentially private deep learning by avoiding numerical integration and pre-specifying a list of moments. On the theory front, our result corroborates the observation of Wang et al. (2019) that the Pearson-Vajda Divergences are natural quantities for understanding the subsampling in differential privacy.

3. Lastly, knowing that exactly evaluating the analytical subsampled RDP bound of αth order takes α calls of the RDP subroutine $\epsilon_{\mathcal{M}}(\cdot)$, we propose an efficiently τ-term approximation scheme that uses only τ call of $\epsilon_{\mathcal{M}}(\cdot)$. We conduct numerical experiments to compare our general bounds, tight bound, and τ-term approximations for a variety of problem setup and showcasing the use of these bounds in moments accountant-based strong composition.

2. Background and Problem Setup

In this section, we provide some background on differential privacy, privacy-amplification by subsampling, RDP and the moments accountant technique so as to formally set up the problem. We will also introduce symbols and notations as we proceed.

Differential Privacy. Let $\mathcal{X}$ be the space of all data sets. One representation of such a data set is to take $\mathcal{X} = \{0,1\}^N$ where N is the size of the population and each $X \in \mathcal{X}$ is an indicator vector that describes each individual's participation in the data set. We say $X, X' \in \mathcal{X}$ are neighbors if X' can be constructed by adding or removing one individual from X, or equivalently, $\|X - X'\|_1 = 1$.

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** *A randomized algorithm $\mathcal{M} : \mathcal{X} \to \Theta$ is (ϵ, δ)-DP (differentially private) if for every pair of neighboring datasets $X, X' \in \mathcal{X}$, 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 places an information-theoretic limit on an adversary's ability to infer whether the input dataset is X or X', and as a result, guarantees a degree of *plausible deniability* to any individual in the population. ϵ, δ are privacy loss parameters that quantify the strength of privacy protection. In practice, we consider the privacy guarantee marginally meaningful if $\epsilon \approx 1$ and $\delta = o(1/n)$[1], where n denotes the size of data set and $o(\cdot)$ is the standard little-o notation. When $\delta = 0$, we say that $\mathcal{M}$ obeys ϵ-(pure) DP.

One important property of DP relevant to this paper is that it composes gracefully over multiple access. Roughly speaking, if we run k sequentially chosen (ϵ, δ)-DP algorithm on a dataset, the overall *composed* privacy loss is $(\tilde{O}(\sqrt{k}\epsilon), k\delta + \delta')$-DP where the $\tilde{O}$ notation hides logarithmic terms in k, $1/\delta$ and $1/\delta'$. Part of the reason for writing this paper is to enable sharper algorithm-dependent composition for a popular class of algorithms that subsamples the data first. Before we get there, let us describe the RDP framework and the moments accountant that the make these algorithm-dependent composition possible.

Renyi Differential Privacy and Moments Accountant. Renyi differential privacy (RDP) is a refinement of DP that uses Renyi-divergence as a distance metric in the place of the sup-divergence.

Definition 2 (Rényi Differential Privacy (Mironov, 2017))**.**

[1] It is traditionally required that δ to be cryptographically small, e.g., $o(\text{poly}(1/n))$, but in practice, with a big data set, $\delta = 1/n^2$ is typically considered acceptable.

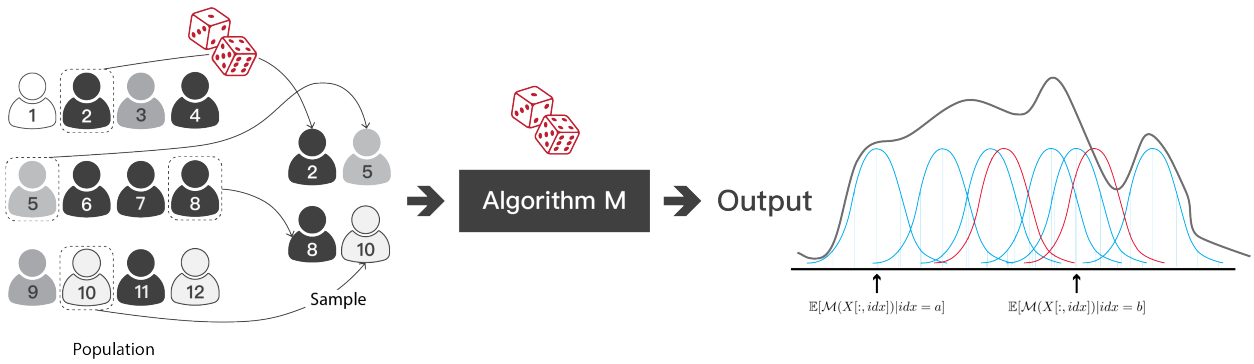


Figure 1. Illustration of the subsampled-mechanism and the key underlying idea that enables "privacy-amplification". The diagram on the left illustrate the two parts of randomization. Part (1): PoissonSample: Each person toss a random coin to select whether they are included in the data set; Part (2): The subsampled data set is analyzed by a randomized algorithm $\mathcal{M}$. The figure on the right illustrates the fact that the distribution of output is a mixture distribution indexed by the different potential subset selected by the subsampling, and that when we change the original data set by adding or removing one person, only a small fraction of the mixture components that *happen to be affected* by that change will be different, hus opening up the possibility of "privacy amplifying".

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

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

In this paper, we do not treat each α in isolation but instead take a functional view of RDP where we use $\epsilon_{\mathcal{M}}(\alpha)$ to denote that randomized algorithm $\mathcal{M}$ obeys $(\alpha, \epsilon_{\mathcal{M}}(\alpha))$-RDP. The function $\epsilon_{\mathcal{M}}(\cdot)$ can be viewed as a more elaborate description of the privacy loss incurred by running $\mathcal{M}$. It subsumes pure-DP as an RDP algorithm is $\epsilon(+\infty)$-DP.

The moments accountant technique (Abadi et al., 2016) can be thought of as a data structure that keeps track of the RDP (function) for the sequence of data accesses. Composition is trivial in RDP as

$$\epsilon_{\mathcal{M}_1\times\mathcal{M}_2}(\cdot) = [\epsilon_{\mathcal{M}_1} + \epsilon_{\mathcal{M}_2}](\cdot).$$

At any given time, let the composition of all algorithms being $\mathcal{M}$, the moments accountant can be used to produce an (ϵ, δ)-DP certificate using

$$\delta \Rightarrow \epsilon: \qquad \epsilon(\delta) = \min_{\alpha>1} \frac{\log(1/\delta)}{\alpha-1} + \epsilon_{\mathcal{M}}(\alpha-1), \tag{2}$$

$$\epsilon \Rightarrow \delta: \qquad \delta(\epsilon) = \min_{\alpha>1} e^{(\alpha-1)(\epsilon_{\mathcal{M}}(\alpha-1)-\epsilon)}. \tag{3}$$

This approach is simpler and often produces more favorable composed privacy parameters than the advanced composition approach for (ϵ, δ)-DP. As the moments accountant gain popularity, many classes of randomized algorithms with exact analytical RDP are becoming available, e.g., the exponential family mechanisms (Geumlek et al., 2017).

As a side note, the initial moments accountant (Abadi et al., 2016) keeps track of a vector of log-moment (equivalent to RDP up to a rescaling) associated with a pre-defined list of order αs. Wang et al. (2019) observes that these optimization problems are unimodal and proposes an *analytical moments accountant* that solves (2) and (3) using bisections can be solved using bisection with a doubling trick. This avoids the need to pre-define the list of moments to track. Wang et al. (2019) also observes that $(\alpha-1)\epsilon(\alpha)$ is a convex function in α and any such discretization scheme (e.g., all integer α) can be extended into a continuous function in α by simply doing linear interpolation.

Privacy amplification by subsampling. As we discussed in the introduction, "privacy amplification by subsampling" is the other workhorse (besides RDP / moments accountant) that drove much of the recent advances in differentially private deep learning. We would like to add that, it was also used as a key technical hammer for analyzing DP algorithms for empirical risk minimization (Bassily et al., 2014) and Bayesian learning (Wang et al., 2015), as well as for studying learning-theoretic questions with differential privacy constraints (Kasiviswanathan et al., 2011; Beimel et al., 2013; Bun et al., 2015; Wang et al., 2016).

We now furnish a bit more details on this central property and highlight some subtleties in the types. The privacy amplification lemma was derived in (Kasiviswanathan et al., 2011; Beimel et al., 2013; Li et al., 2012), where all three authors adopted what Balle et al. (2018) calls Poisson subsampling:

Definition 3 (PoissonSample)**.** *Given a dataset X, the procedure* PoissonSample *outputs a subset of the data $\{x_i|\sigma_i = 1, i \in [n]\}$ by sampling $\sigma_i \sim \mathrm{Ber}(\gamma)$ independently for $i = 1, ..., n$.*

The procedure is equivalent to the "sampling without replacement" scheme with $m \sim \text{Binomial}(\gamma, n)$. At the limit of $n \to \infty$, $\gamma \to 0$ while $\gamma n \to \lambda$, the Binomial distribution converges to a Poisson distribution with parameter λ. This is probably the reason why it is called Poisson sampling to begin with[2].

Here we cite the tight privacy amplification bound for PoissonSample as it first appears.

Lemma 4 ((Li et al., 2012, Theorem 1))**.** *If $\mathcal{M}$ is (ϵ, δ)-DP, then $\mathcal{M}'$ that applies $\mathcal{M} \circ \mathsf{PoissonSample}$ obeys (ϵ', δ')-DP with $\epsilon' = \log\left(1 + \gamma(e^{\epsilon} - 1)\right)$ and $\delta' = \gamma\delta$.*

The lemma implies that if the base privacy loss $\epsilon \leq 1$, then the amplified privacy loss obeys that $\epsilon' \leq 2\gamma\epsilon$.

Poisson subsampling is different from the "sampling without replacement" scheme that outputs a subset with size γn uniformly at random. Interestingly, it was shown that the latter also enjoys the same bound with respect to the "replace-one" version of the DP definition. In general, we find that the "add/remove" version of the DP definition works more naturally with Poisson sampling, while the "replace-one" version works well with "sampling without replacement". We defer a more comprehensive account of the subsampling lemma for (ϵ, δ)-DP to (Balle & Wang, 2018) and the references therein.

Subsampled RDP and friends. A small body of recent work focuses on deriving algorithm-specific subsampling Lemma so that this classical wisdom can be combined with more modern techniques such as RDP and Concentrated DIfferential privacy (CDP) (Bun & Steinke, 2016) (also (Dwork & Rothblum, 2016)). Abadi et al. (2016) obtains the first such results for subsampled-Gaussian mechanism under Poisson subsampling. Wang et al. (2019) provides a general subsampled RDP bound that supports any $\mathcal{M}$ but under the "sampling without replacement" scheme. The objective of this paper is to come up with results of a similar flavor for the Poisson sampling scheme. The main differences in our setting include:

(a) Poisson sampling goes naturally with add/remove version of the DP definition, which is independent to the size of the data.

(b) The size of the random subset m itself is a Binomial random variable.

(c) It is asymmetric, the Renyi divergence of P against Q is different from the Renyi divergence of Q against P.

As we will see in the our results, the third difference brings about some major technical challenges.

Finally, Bun et al. (2018) studies subsampling in CDP with a conclusion that subsampling does not amplify the CDP parameters in general. A truncated version of CDP was then proposed, called tCDP, which does get amplified up to a threshold. CDP and tCDP are closely related to RDP in that they are linear upper bounds of $\epsilon(\alpha)$ on $(1, \infty]$ and on $(1, \tau]$ for some threshold τ respectively. RDP captures finer information about the underlying mechanism. The experimental results in (Wang et al., 2019) suggest that unlike the case for the Gaussian mechanism (in which case CDP is tight), there isn't a good linear approximation of $\epsilon(\alpha)$ for the subsampled-Gaussian mechanism due to the phase transition. Our results on the Poisson-sampling model echoes the same phenomenon.

More symbols and notations. We end the section with a quick summary of the notations that we introduced. X, X' denotes two neighboring datasets. $\mathcal{M}$ is a randomized algorithm and $\epsilon_{\mathcal{M}}(\cdot)$ is the RDP function of $\mathcal{M}$ (the subscript may be dropped when it's clear from the context). n, m are reserved for the size of the original and subsampled data. We note that neither is public and m is random. Greek letters $\alpha, \gamma, \epsilon, \delta$ are reserved for the order of RDP, the sampling probability as well as the two privacy loss parameters. $\mathcal{M} \circ \mathsf{PoissonSample}(X)$ is used to mean the composition function $\mathcal{M}(\mathsf{PoissonSample}(X))$.

Let us also define a few shorthands. We will denote p to be the density function of $\mathcal{M} \circ \mathsf{PoissonSample}(X)$, and q to be the density from data set $\mathcal{M} \circ \mathsf{PoissonSample}(X')$. Similarly, we will define μ_0 and μ_1 as two generic density functions of $\mathcal{M}(X)$ and $\mathcal{M}(X')$.

3. Main results

Before we present our main result, we would like to warn the readers that the presented bounds might not be as interpretable. We argue that this is a feature rather than an artifact of our proof because we need the messiness to state the bound exactly. These bounds are meant to be *implemented* to achieve the tightest possible privacy composition numerically in the Moments Accountant, rather than being made easily interpretable. After all, "constant matters in differential privacy!" For the interest of interpretability, we provide figures that demonstrate the behaviors of the bound for prototypical mechanisms in practice.

Theorem 5 (General upper bound)**.** *Let $\mathcal{M}$ be any randomized algorithm that obeys $(\alpha, \epsilon(\alpha))$-RDP. Let γ be the sub-*

[2] We noticed that the original definition of Poisson sampling in the survey sampling theory is slightly more general. It allows a different probability of sampling each person (Särndal et al., 2003). Our results apply trivially to that setting as well with a personalized RDP bound for individual i that depends on γ_i.

sampling probability and then we have for integer $\alpha \geq 2$,

$$\epsilon_{\mathcal{M}\circ\mathsf{PoissonSample}}(\alpha) \leq \frac{1}{\alpha-1}\log\Bigg\{(1-\gamma)^{\alpha-1}(\alpha\gamma-\gamma+1) + \binom{\alpha}{2}\gamma^2(1-\gamma)^{\alpha-2}e^{\epsilon(2)} + 3\sum_{\ell=3}^{\alpha}\binom{\alpha}{\ell}(1-\gamma)^{\alpha-\ell}\gamma^{\ell}e^{(\ell-1)\epsilon(\ell)}\Bigg\}.$$

The proof is revealing but technically involved. One main difference from Wang et al. (2019) is that in Poisson sampling we need to bound both $D_\alpha(p\|q)$ and $D_\alpha(q\|p)$. Existing arguments via the quasi-convexity of Renyi divergence allows us to easily bound $D_\alpha(p\|q)$ tightly using RDP for the case when p has one more data points than q, but $D_\alpha(q\|p)$ turns out to be very tricky. A big part of our novelty in the proof is about analyzing $D_\alpha(q\|p)$. We defer more details of the proof to Appendix A.

Theorem 6 (Lower bound)**.** *$\mathcal{M}$ and pairs of adjacent data sets such that*

$$\epsilon_{\mathcal{M}\circ\mathsf{PoissonSample}}(\alpha) \geq \frac{1}{\alpha-1}\log\Bigg\{(1-\gamma)^{\alpha-1}(\alpha\gamma-\gamma+1) + \sum_{\ell=2}^{\alpha}\binom{\alpha}{\ell}(1-\gamma)^{\alpha-\ell}\gamma^{\ell}e^{(\ell-1)\epsilon(\ell)}\Bigg\}.$$

Proof. The construction effectively follows Proposition 11 of (Wang et al., 2019), while adjusting for the details. Let $\mathcal{M}$ be Laplace noise adding of a counting query $f(X') = \sum_{x\in X'}\mathbf{1}[x>0]$. Let everyone in the data set X' obeys that $x<0$. In the adjacent dataset $X' = X\cup\{x_{n+1}\}$ with $x_{n+1}>0$. Let μ_0 be the Laplace distribution centered at 0, μ_1 be the one that is centered at 1. Then we know that $\mathcal{M}(X')\sim\mu_0 = q$ and $\mathcal{M}(X)$ $(1-\gamma)\mu_0+\gamma\mu_1 = p$. It follows that

$$\begin{aligned}&\mathbb{E}_q[(p/q)^\alpha] = \mathbb{E}_{\mu_0}[((1-\gamma)+\gamma\mu_1/\mu_0)^\alpha]\\ =&\sum_{\ell=0}^{\alpha}\binom{\alpha}{\ell}(1-\gamma)^{\alpha-\ell}\gamma^{\ell}\mathbb{E}_{\mu_0}[(\mu_1/\mu_0)^\ell].\end{aligned}$$

By definition $\mathbb{E}_{\mu_0}[(\mu_1/\mu_0)^\ell] = e^{(\ell-1)D_\ell(\mu_0\|\mu_1)}$. which the RDP bonud $\epsilon(\ell)$ is attained by μ_0,μ_1, then we have constructed one pair of p,q, which implies a lower bound for RDP of $\mathcal{M}\circ\mathsf{PoissonSample}$. □

Note that the only difference between the upper and lower bounds are a factor of 3 on the third summand in side the logarithm. In the regime when $\gamma\alpha e^{\epsilon(\alpha)} \ll 1$ (in which case the third summand is much smaller than the second), the upper and lower bound match up to a multiplicative factor of $1+O(\gamma\alpha e^{\epsilon(\alpha)})$. In all other regimes, the upper and lower bounds match up to an additive factor of $\frac{\log(3)}{\alpha-1}$. The results suggest that we can construct a nearly optimal moment accountant.

Remark 7 (Nearly optimal Moment Accountant)**.** *This implies that if* any algorithm *with the help of an oracle that calculates the exact RDP for $\mathcal{M}$ is able to prove an (ϵ,δ)-DP for the Poisson subsampled RDP mechanism, then the RDP upper bound we construct using Theorem 5 will lead to an $(\epsilon, 3\delta)$-DP bound for the same mechanism.*

Moreover, we show that for many randomized algorithms (including the popular Gaussian mechanism and Laplace mechanism) that satisfy an additional assumption, we can strengthen the upper bound further and *exactly match* the lower bound for all α.

Theorem 8 (Tight upper bound)**.** *Let $\mathcal{M}$ be a randomized algorithm with up to αth order RDP $\epsilon(\alpha)<\infty$. If for all adjacent data sets $X\sim X'$, and all odd $3\leq\ell\leq\alpha$,*

$$D_{\chi^\ell}(\mathcal{M}(X)\|\mathcal{M}(X')) := \mathbb{E}_{\mathcal{M}(X')}\left(\frac{\mathcal{M}(X)}{\mathcal{M}(X')}-1\right)^\ell \geq 0, \tag{4}$$

then the lower bound in Theorem 6 is also an upper bound.

The proof of this theorem is presented in Appendix A

In the theorem, $\frac{\mathcal{M}(X)}{\mathcal{M}(X')}-1$ is a linearized version of the privacy random variable $\log\frac{\mathcal{M}(X)}{\mathcal{M}(X')}$ and $D_{\chi^\ell}(\mathcal{M}(X)\|\mathcal{M}(X'))$ is the Pearson-Vajda χ^ℓ pseudo-divergence (Vajda, 1973), which has more recently been used to approximate any f-divergence in (Nielsen & Nock, 2014). The related $|\chi^\ell|$ version of this divergence is identified as the key quantity *natural* for studying subsampling without replacement (Wang et al., 2019).

The non-negativity condition requires, roughly speaking, the distribution of the linearized privacy loss random variable $\frac{\mathcal{M}(X)}{\mathcal{M}(X')}-1$ to be skewed to the right.

The following Lemma provides one way to think about it.

Lemma 9. *Let π,μ be two measures that are absolute continuous w.r.t. each other and let $\alpha\geq 1$.*

$$\mathbb{E}_\mu[(\pi/\mu-1)^\alpha] = \mathbb{E}_\pi[(\pi/\mu-1)^{\alpha-1}] - \mathbb{E}_\mu[(\pi/\mu-1)^{\alpha-1}].$$

Proof. $\mathbb{E}_\mu[(\pi/\mu-1)^\alpha] = \mathbb{E}_\mu[(\pi/\mu-1)(\pi/\mu-1)^{\alpha-1}] = \mathbb{E}_\pi[(\pi/\mu-1)^{\alpha-1}] - \mathbb{E}_\mu[(\pi/\mu-1)^{\alpha-1}]$ □

The lemma implies that (4) holds if an only if for all even $2\leq\ell\leq\alpha$

$$\mathbb{E}_{\mathcal{M}(X)}\left(\frac{\mathcal{M}(X)}{\mathcal{M}(X')}-1\right)^\ell \geq \mathbb{E}_{\mathcal{M}(X')}\left(\frac{\mathcal{M}(X)}{\mathcal{M}(X')}-1\right)^\ell$$

for all pairs of X, X'.

This should intuitively be true for most mechanisms because we know from nonnegativity that $\frac{\mathcal{M}(X)}{\mathcal{M}(X')}-1\geq -1$, which poses a hard limit to which you can be skewed to the

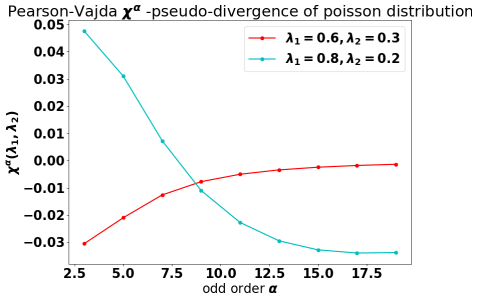


Figure 2. Negative χ^α divergence in Poisson distribution.

left. Indeed, we can show that the condition is true for the two most widely used DP procedure.

Proposition 10. *Condition* (4) *is true when $\mathcal{M}$ is the Gaussian mechanism and Laplace mechanism.*

The proof, given in Appendix B, is interesting and can be used as recipes to qualify other mechanisms for the tight bound. The main difficulty of checking this condition is in searching all pairs of neighboring data sets and identify one pair that minimizes the odd order moment. The convenient property of noise-adding procedure is that typically the search reduces to a univariate optimization problem of the sensitivity parameter.

One may ask, whether the condition is true in general for any randomized algorithm $\mathcal{M}$? The answer is unfortunately no. For example, Nielsen & Nock (2014) constructed an example of two Poisson distributions with negative χ^3-divergence (see Figure 2 for an illustration.) This also implies that for some $\mathcal{M}$, we can derive a lower bound that is greater than that in Theorem 6 by simply toggling the order of X and X'. As a result, if one needs to work out the tight bound, the condition needs to be checked for each $\mathcal{M}$ separately.

Finally, we address the computational issue of implementing our bounds in moments accountant. Naive implementation of Theorem 5 will easily suffer from overflow or underflow and it takes α calls to the RDP oracle $\epsilon_{\mathcal{M}}(\cdot)$ before we can evaluate one RDP of the subsampled mechanism at α. This is highly undesirable. The following theorem provides a fast approximation bound that can be evaluated with just 2τ calls to the RDP oracle of $\mathcal{M}$. The idea is that since either it is the first few terms that dominates the sum or the last few terms that dominates the sum, we can just compute them exactly and calculate the remainder terms with a more easily computable upper bound.

Theorem 11 (τ-term approximation)**.** *The expression in Theorem 6 (therefore Theorem 8) can be bounded by*

$$\epsilon_{\mathcal{M}\circ\text{PoissonSample}}(\alpha) \le \frac{1}{\alpha-1}\log\Big\{(1-\gamma)^\alpha(1-e^{-\epsilon(\alpha-\tau)})$$
$$+ e^{-\epsilon(\alpha-\tau)}(1-\gamma+\gamma e^{\epsilon(\alpha-\tau)})^\alpha$$
$$- \sum_{\ell=2}^{\tau}\binom{\alpha}{\ell}(1-\gamma)^{\alpha-\ell}\gamma^\ell(e^{(\ell-1)\epsilon(\alpha-\tau)} - e^{(\ell-1)\epsilon(\ell)})$$
$$+ \sum_{\ell=\alpha-\tau+1}^{\alpha}\binom{\alpha}{\ell}(1-\gamma)^{\alpha-\ell}\gamma^\ell(e^{(\ell-1)\epsilon(\ell)} - e^{(\ell-1)\epsilon(\alpha-\tau)})\Big\}.$$

A similar bound can be stated for the general upper bound in Theorem 5, which we defer to Appendix C.

Remark 12 (Numerical stability)**.** *The bounds in Theorem 5 and 8 can be written as the* log − sum − exp *form, i.e., softmax. The numerically stable way of evaluating* log − sum − exp *is well-known. The bound in Theorem 11, though, have both positive terms and negative terms. We choose to represent the summands in the* log *term by a sign, and the logarithm of its magnitude. This makes it possible for us to use* log − diff − exp *and compute the bound in a numerically stable way.*

4. Experiments and Discussion

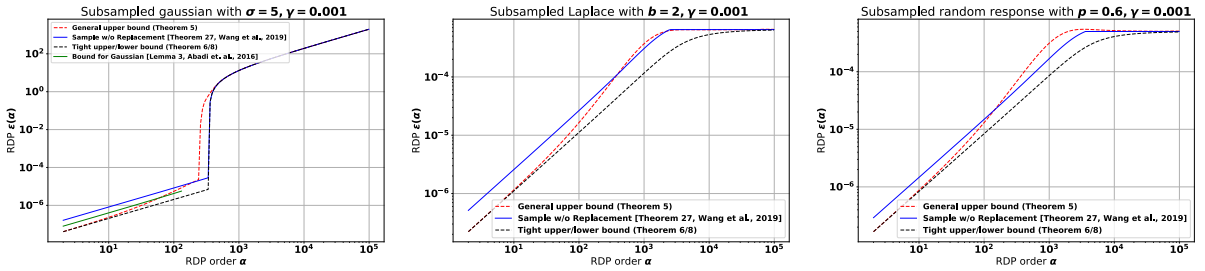
In this section, we conduct various numerical experiments to illustrate the behaviors of the RDP for subsampled mechanisms and showcasing its usage in moments accountant for composition. We will have three set of experiments. (1) We will just plot our RDP bounds (Theorem 5, Theorem 6) as a function of α. (2) We will compare how close the τ-term approximations approximate the actual bound.(5) (3) We will build our moments accountant and illustrate the stronger composition that we get out of our tight bound.

Specifically, for each of the experiments above, we replicate the experimental setup of which takes the base mechanism $\mathcal{M}$ to be Gaussian mechanism, Laplace mechanism and Randomized Response mechanism. Their RDP formula are worked out analytically (Mironov, 2017) below:

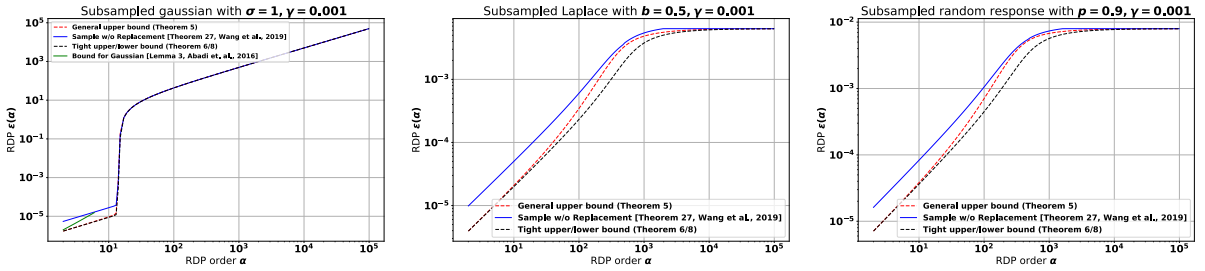
$$\epsilon_{Gaussian(\alpha)} = \frac{\alpha}{2\sigma^2},$$
$$\epsilon_{Laplace(\alpha)} = \frac{1}{\alpha-1}\log((\frac{\alpha}{2\alpha-1})e^{\frac{\alpha-1}{b}} + (\frac{\alpha-1}{2\alpha-1})e^{\frac{-\alpha}{b}} \text{ for } \alpha > 1,$$
$$\epsilon_{RandResp(\alpha)} = \frac{1}{\alpha-1}\log(p^\alpha(1-p)^{1-\alpha}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

where parameter σ, b, p are the standard parameters for Gaussian, Laplace and Bernoulli distributions.

Following Wang et al. (2019), we will have two sets of experiments with "high noise, high privacy" setting $\sigma = 5, b = 2$, and $p = 0.6$ and "low noise, low privacy" setting using $\sigma = 1, b = 0.5, p = 0.9$. These parameters are chosen such that the ϵ-DP or (ϵ, δ)-DP of the base mechanisms are roughly $\epsilon \approx 0.5$ in the high privacy setting or $\epsilon \approx 2$ in the low privacy setting.



(a) Subsampled Gaussian with $\sigma = 5$ (b) Subsampled Laplace with $b = 2$ (c) Subsampled Random Response with $p = 0.6$



(d) Subsampled Gaussian with $\sigma = 1$ (e) Subsampled Laplace with $b = 0.5$ (f) Subsampled Random Response with $p = 0.9$

Figure 3. The RDP parameter $(\epsilon(\alpha))$ of three subsampled mechanisms as a function of order α.We compare the general upper bound with other methods under high and low privcy regime, general upper bound is obtained through Theorem 5. the corresponding tight upper bound in possion subsample case is represented as the black curve.

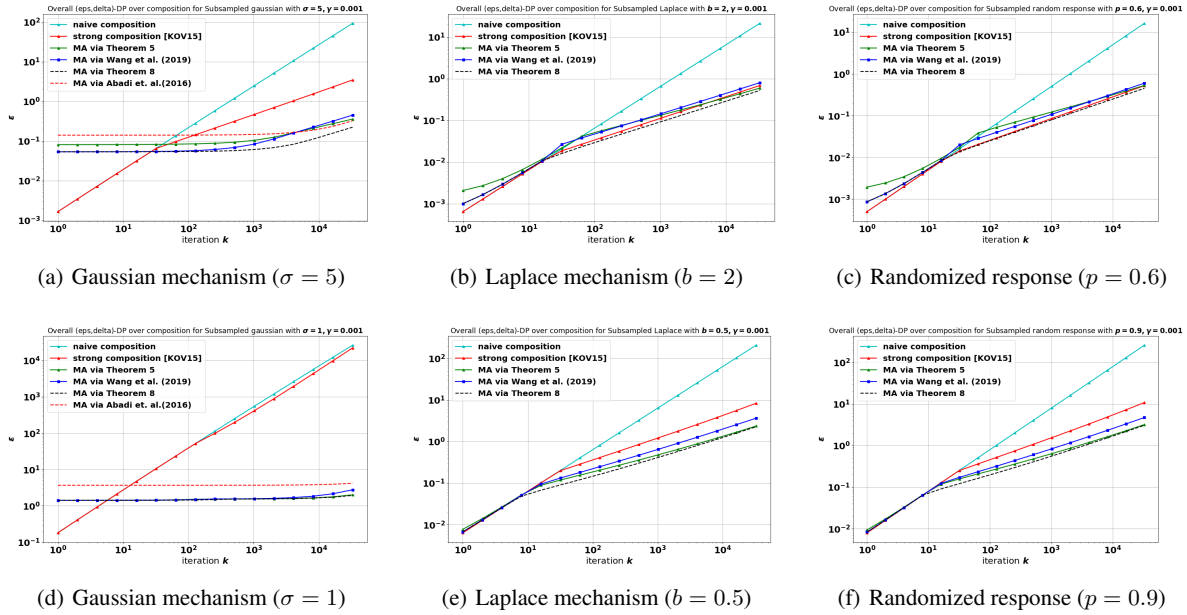
We will include benchmarks when appropriate. For example, we will compare to Lemma 3 of Abadi et al. (2016) whenever we work with Gaussian mechanisms. Also, we will compare to the upper bound of Wang et al. (2019) for subsample without replacements. Finally, we will include the more traditional approaches of tracking and composing privacy losses using simply (ϵ, δ)-DP. We will see that while the moments accountant approach does not dominate the traditional approach, it does substantially reduces the aggregate privacy loss for almost all experiments when we compose over a large number of rounds.

Comparing RDP bounds. The results on the RDP bounds are shown in the Figure 3. First of all, the RDP of subsampled Gaussian mechanism behaves very differently from that of the Laplace mechanism, There is a phase transition about the subsampled-Gaussian mechanism that happens around $\alpha\gamma e^{\epsilon(\alpha)} \approx \gamma^{-1}$. Before the phase transition the RDP is roughly $O(\gamma^2\alpha^2(e^{\epsilon(2)} - 1))$, the RDP quickly converges to $\epsilon(\alpha)$, which implies that subsampling has no effects. This kind of behaviors cannot be captured through CDP. On the other hand, for ϵ-DP mechanisms, the RDP increases linearly with α before being capped what the standard privacy amplification by Lemma 4. Relative to existing bounds, our tight bound closes the constant gap, while our general bound is also nearly optimal as we predict. It is worth noting that the bound in Abadi et al. (2016) only applies to up to a threshold of α.

τ-term approximation. Figure 5 illustrates the quality of approximation as we increase τ. With $\tau = 50$, the results nearly matches the RDP bound everywhere, except that in the Gaussian case, the phase-transition happened a little bit earlier.

Usage in moments accountant. The experiments on moments accountant are shown in Figure 4. Our result are compared to the optimal strong composition (Kairouz et al., 2015) with parameters optimally tuned according to Wang et al. (2019). As we can see, all bounds based on the moments accountants eventually scales proportional to $\sqrt{k}$. Moments accountant techniques with the tight bound end up winning by a constant factor. It is worth noting that in the Gaussian case, moments accountant only starts to perform better than traditional approaches after composing for 1000 times. Also, the version of moments accountant using the theoretical bounds from Abadi et al. (2016) gave substantially worse results[3]. Finally the general RDP bound

[3]We implemented the bound from the proof of Abadi et al. (2016)'s Lemma 3 for fair comparison. According to Section 3.2 of Abadi et al. (2016), their experiments use numerical integration to approximate the moments. See more discussion on this in Appendix D.



(a) Gaussian mechanism ($\sigma = 5$) (b) Laplace mechanism ($b = 2$) (c) Randomized response ($p = 0.6$)

(d) Gaussian mechanism ($\sigma = 1$) (e) Laplace mechanism ($b = 0.5$) (f) Randomized response ($p = 0.9$)

Figure 4. Illustration of the use of our bounds in moments accountant. We plot the the privacy loss ϵ for $\delta = 1e-8$ (using (2)) after k rounds of composition. The x-axis is the number of composition k,and the y-axis is the privacy loss after k's composition. The green curve is based on general upper bound for all parametrized random mechanism obtained through Theorem 5. Short hand MA refer to "moment accountant". The upper three figures are in high privacy regime with parameter $\sigma = 5, b = 2, p = 0.6$, the lower three are in low privacy regime with $\sigma = 1, b = 0.5, p = 0.9$.

perform as well as the tight bound when k is large (thanks to the tightness for small α).

5. Conclusion

In this paper, we study the problem of privacy-amplification by poisson subsampling, which involves "add/remove" scheme instead of replacement strategy. Specifically, we derive a tight upper bound for $\mathcal{M} \circ$ PoissonSample for any mechanism satisfying that their odd order Pearson-Vajda χ^α-Divergences are nonnegative. We showed that Gaussian mechanism and Laplace mechanism have this property, as a result, finding the exact analytical expression for the Poisson subsampled Gaussian mechanism that has seen significant application in differentially private deep learning. Our results imply that we can completely avoid numerical integration in moments accounts and track the entire range of α without paying unbounded memory. In addition, we propose an efficiently τ-term approximation scheme which only calculates the first and last τ terms in the Binomial expansion when evaluating the RDP of subsampled mechanisms. This greatly simplifies the computation for computing ϵ given δ as is used in the moments accountant. The experiment result of τ-term approximate part reveals that approximate bound matches up the lower bound quickly even for a relative small τ.

Future work includes making use of the exact subsampled RDP bounds to tighten the existing results that made use of subsampled-mechanisms, coming up with more general recipe to automatically check the nonnegativity condition on the odd-order Pearson-Vajda χ^α-Divergences and design differentially private learning algorithms with more complex and hetereogenous building blocks.

Acknowledgements

YZ and YW were supported by a start-up grant of UCSB Computer Science Department and a Machine Learning Research Award from Amazon Web Services. The authors thank Borja Balle, Shiva Kasiviswanathan, Alex Smola, Mu Li and Kunal Talwar for helpful discussions.

References

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

Balle, B. and Wang, Y.-X. Improving gaussian mechanism for differential privacy: Analytical calibration and op-

timal denoising. *International Conference in Machine Learning (ICML)*, 2018.

Balle, B., Barthe, G., and Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems (NIPS-18)*, 2018.

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

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

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

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

Bun, M., Dwork, C., Rothblum, G. N., and Steinke, T. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 74–86. ACM, 2018.

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

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

Foulds, J., Geumlek, J., Welling, M., and Chaudhuri, K. 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, 2016.

Geumlek, J., Song, S., and Chaudhuri, K. Renyi differential privacy mechanisms for posterior sampling. In *Advances in Neural Information Processing Systems (NIPS-17)*, pp. 5295–5304, 2017.

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

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

Li, N., Qardaji, W., and Su, D. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *The 7th ACM Symposium on Information, Computer and Communications Security*, pp. 32–33. ACM, 2012.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR-18)*, 2018.

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

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

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations (ICLR-18)*, 2018.

Park, M., Foulds, J., Chaudhuri, K., and Welling, M. Variational bayes in private settings (vips). *arXiv preprint arXiv:1611.00340*, 2016.

Särndal, C.-E., Swensson, B., and Wretman, J. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

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

Vajda, I. χ^α-divergence and generalized fisher information. In *Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 223. Academia, 1973.

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

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

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. Subsampled rényi differential privacy and analytical moments accountant. In *International Conference on Artificial Intelligence and Statistics (AISTATS-19)*, 2019.