Privacy Amplification of Iterative Algorithms via Contraction Coefficients

Shahab Asoodeh[†], Mario Diaz[*], and Flavio P. Calmon[†]
[†]Harvard University, [*]Universidad Nacional Autónoma de México

***Abstract*—We investigate the framework of privacy amplification by iteration, recently proposed by Feldman et al., from an information-theoretic lens. We demonstrate that differential privacy guarantees of iterative mappings can be determined by a direct application of contraction coefficients derived from strong data processing inequalities for f-divergences. In particular, by generalizing the Dobrushin's contraction coefficient for total variation distance to an f-divergence known as E_γ-divergence, we derive tighter bounds on the differential privacy parameters of the projected noisy stochastic gradient descent algorithm with hidden intermediate updates.**

I. Introduction and Motivation

Differential privacy (DP) [1, 2] has become the standard definition for designing privacy-preserving machine learning algorithms. One reason for its success is its *operational* significance, which can be best described in terms of binary hypothesis testing (see, e.g., [3, 4]). Nevertheless, it is often difficult to compute DP guarantees for applications where a high number of data accesses is needed for a single analysis [5, 6]. To obtain the DP parameters in such applications, which include machine learning models trained using stochastic gradient descent (SGD), one needs to resort to composition theorems which are often loose due to their generality. As a remedy, several variants of DP have been recently proposed [7–10] based on Rényi divergence. These variants enjoy better composition properties. Among these variants, Rényi DP (RDP) has proven to be effective in studying private deep learning algorithms [5] especially when paired with *sub-sampling* techniques [11].

Recently, the new framework of *privacy amplification by iteration* was proposed by Feldman et al. [12] as an alternative to privacy amplification by sub-sampling. This framework possesses several advantages which makes it well-suited for determining and enforcing privacy in distributed settings where data samples are stored locally by each user. Existing private algorithms based on sub-sampling require hiding the set of users participating in each update step of the model. This requirement, however, dictates either all data samples be stored centrally (i.e., no distributed setting) or all-to-all communication (i.e., excessive communication complexity). The new framework of privacy amplification by iteration relaxes these issues; it does not require the order of participating users to be random or hidden. On the other hand, it requires that all intermediate updates be hidden until a certain number of update steps are applied (e.g., not disclosing model update of SGD before a pre-specified step, say, n-th step).

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

Since the intermediate updates are assumed to be hidden, one can view an iterative process as a concatenation of channels. To see this, let $\{\psi_t\}_{t=1}^n$ be a sequence of mapping and the update rule be given by

$$Y_t = \psi_t(Y_{t-1}) + Z_t, \tag{1}$$

where $Y_0 = y_0 \in \mathbb{R}^d$ and $\{Z_t\}_{t=1}^n$ are i.i.d. copies of a noise distribution P_Z. Let $\{Y_t'\}_{t=1}^n$ be the output of the same process started at $Y_0' = y_0' \in \mathbb{R}^d$. Letting μ_t and ν_t be the distributions of Y_t and Y_t', the *strong* data processing inequality (SDPI) for f-divergences (see, e.g., [13, 14]) implies that

$$D_f(\mu_n\|\nu_n) \le D_f(\mu_1\|\nu_1)\prod_{t=1}^n \eta_f(\mathsf{K}_t), \tag{2}$$

where D_f is an f-divergence and $\eta_f(\mathsf{K}_t)$ is the *contraction coefficient* (also known as strong data processing constant) of the Markov kernel $\mathsf{K}_t(y) := P_{Y_t|Y_{t-1}=y} = P_{Z+\psi_t(y)}$ under f-divergence (see Section III for details). By exploiting the connection between DP and a certain f-divergence known as E_γ-divergence, we build upon (2) to obtain bounds for DP parameters of iterative processes. Specifically, we study the noisy stochastic gradient descent algorithm and obtain tighter bounds for its DP parameters than those provided currently in the literature [12, 15]. To do so, we obtain a closed-form expression for the contraction coefficient of Markov kernels under E_γ-divergence that generalizes the well-known Dobrushin's theorem [16].

Our approach is inspired by the original work of Feldman et al. [12]. They adopted RDP as the measure of privacy and proved the following SDPI result [12, Theorem 1] for the Rényi divergence of order $\alpha > 1$: For the iterative process described in (1) with P_Z the Gaussian distribution $\mathcal{N}(0, \sigma^2\mathrm{I}_d)$,

$$D_\alpha(\mu_n\|\nu_n) \le \frac{1}{n}D_\alpha(\mu_1\|\nu_1) = \frac{1}{n}\frac{\alpha\|y_0 - y_0'\|}{2\sigma^2}. \tag{3}$$

Despite its tractability, RDP lacks a clear operational interpretation. As a result, RDP guarantees are usually translated to DP guarantees via a transformation which is known to be loose, see, e.g., [7, Proposition 3].

As a special case of iterative processes, we consider the *noisy* SGD algorithm with Laplacian or Gaussian perturbation.

Authorized licensed use limited to: Harvard Library. Downloaded on November 18,2020 at 14:41:08 UTC from IEEE Xplore. Restrictions apply.

Our empirical analyses show that the DP parameters of noisy SGD obtained by our approach are smaller than that of [12, 15] (after applying the RDP to DP transformation). To capture common practice in machine learning applications, the input alphabet of the Markov kernels in this work are assumed to be compact. As a result, our analysis of contraction coefficients of such kernels is akin to the analysis of input-constrained channels performed by [17]. In the interest of space, all the proofs are delegated to its full version available online [18].

II. Background

In this section, we briefly review privacy mechanisms, f-divergences and contraction coefficients. We also review a relation between DP and E_γ-divergence.

A. Privacy Mechanisms

The following examples describe two typical privacy mechanisms used in machine learning.

Example 1. (Private Queries) Let $\mathcal{X}$ be an arbitrary alphabet. A query is a function f that takes a sample $\mathbb{D} \in \mathcal{X}^n$ and produces a *response* y in the space of responses $\mathcal{Y}$. In this setting, a privacy mechanism K takes a response $y \in \mathcal{Y}$ and produces another (random) response in the same space. In general, a privacy mechanism can be described by a Markov kernel $\mathsf{K} : \mathcal{Y} \to \mathcal{P}(\mathcal{Y})$, i.e., a channel with the same input and output space $\mathcal{Y}$, where $\mathcal{P}(\mathcal{Y})$ denotes the set of probability measures over $\mathcal{Y}$. Thus, the private query, say $\mathcal{M}$, is a random variable satisfying $\mathcal{M}(\mathbb{D}) \sim \mathsf{K}(f(\mathbb{D}))$.

Example 2. (Stochastic Optimization) Let $\mathcal{Y}$ denote a parameter space, e.g., the coefficients in a linear regression model. Given a dataset $\mathbb{D} = \{x_1, \dots, x_n\} \in \mathcal{X}^n$, typical stochastic optimization methods take an initial point $Y_0 \sim \mu_0 \in \mathcal{P}(\mathcal{Y})$ and further refine it through a random optimization process. The latter process typically depends on the dataset $\mathbb{D}$ and can be encoded by a Markov kernel $\mathsf{K}_\mathbb{D} : \mathcal{Y} \to \mathcal{P}(\mathcal{Y})$. Furthermore, in some cases it is of iterative form, e.g., stochastic gradient descent, and the kernel $\mathsf{K}_\mathbb{D}$ can be decomposed as $\mathsf{K}_\mathbb{D} = \mathsf{K}_{x_1} \cdots \mathsf{K}_{x_n}$. Here, the randomness of the initial point and the optimization process may provide some level of privacy.

Motivated by the previous examples, we model privacy mechanisms as random mappings taking a data set $\mathbb{D} \in \mathcal{X}^n$ as input and producing an element in a given set $\mathcal{Y}$ as output. Furthermore, we assume that any privacy mechanism, say $\mathcal{M}$, is a random variable satisfying

$$\mathcal{M}(\mathbb{D}) \sim \mu_0 \mathsf{K} \coloneqq \int \mu_0(\mathrm{d}y) \mathsf{K}(y),$$

where the measure $\mu_0 \in \mathcal{P}(\mathcal{Y})$ and the kernel $\mathsf{K} : \mathcal{Y} \to \mathcal{P}(\mathcal{Y})$ may depend on $\mathbb{D}$, i.e., $\mu_0 = (\mu_0)_\mathbb{D}$ and $\mathsf{K} = \mathsf{K}_\mathbb{D}$.

B. f-Divergence and Contraction Coefficients

Given a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, f-divergence between two probability measures μ and ν is defined in [19, 20] as

$$D_f(\mu \| \nu) \coloneqq \mathbb{E}_\nu \left[f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right) \right].$$

Let $\mathsf{K} : \mathcal{Y} \mapsto \mathcal{P}(\mathcal{Y})$ be a Markov kernel. Following the definition from Ahslwede and Gács [21], we define the *contraction coefficient* (or strong data processing coefficient) of K under f-divergence as

$$\eta_f(\mathsf{K}) \coloneqq \sup_{\substack{\mu,\nu: \\ D_f(\mu\|\nu) \neq 0}} \frac{D_f(\mu\mathsf{K} \| \nu\mathsf{K})}{D_f(\mu \| \nu)}.$$

This quantity has been studied for several f-divergences, e.g., KL-divergence for which $f(t) = t \log(t)$, χ^2-divergence for which $f(t) = (t-1)^2$, and also total variation distance for which $f(t) = \frac{1}{2}|t-1|$. In particular, Dobrushin [16] showed that

$$\eta_{\mathsf{TV}}(\mathsf{K}) = \sup_{y_1 \neq y_2} \mathsf{TV}(\mathsf{K}(y_1), \mathsf{K}(y_2)), \tag{4}$$

where $\mathsf{TV}(\mu, \nu)$ denotes the total variation distance between μ and ν. It is worth noting that (4) has been extensively used in information theory [17, 22], statistics [23] and graph theory [24, 25].

C. Differential Privacy and E_γ-Divergence

For an arbitrary alphabet $\mathcal{X}$, let $\mathcal{X}^n$ be the set of all datasets of size n. By definition, two datasets $\mathbb{D}$ and $\mathbb{D}'$ are *neighboring*, denoted as $\mathbb{D} \sim \mathbb{D}'$, if their Hamming distance is equal to one. Given a randomized mechanism $\mathcal{M}$, we let $P_\mathbb{D}$ be the distribution of $\mathcal{M}(\mathbb{D})$, the output of $\mathcal{M}$ with $\mathbb{D} \in \mathcal{X}^n$ as the input. For $\varepsilon \geq 0$ and $\delta \in [0, 1]$, a mechanism $\mathcal{M}$ is said to be (ε, δ)-differentially private (DP) if

$$P_\mathbb{D}(A) \leq e^\varepsilon P_{\mathbb{D}'}(A) + \delta, \tag{5}$$

for every measurable set $A \subset \mathcal{Y}$ and neighboring datasets $\mathbb{D} \sim \mathbb{D}'$. When $\delta = 0$, we simply say that $\mathcal{M}$ is ε-DP.

The definition of (ε, δ)-DP given in (5) can be reformulated in terms of E_γ-divergence, also known as hockey-stick divergence [26–28]. Given $\gamma \geq 1$, E_γ-divergence between two probability distributions P and Q is defined as

$$\begin{aligned} \mathsf{E}_\gamma(P \| Q) &\coloneqq \int_\mathcal{Y} [\mathrm{d}(P - \gamma Q)(y)]_+ \qquad (6) \\ &= \sup_{A \subset \mathcal{Y}} [P(A) - \gamma Q(A)] \\ &= P\left(\imath_{P\|Q} > \log\gamma\right) - \gamma Q\left(\imath_{P\|Q} > \log\gamma\right), \qquad (7) \end{aligned}$$

where $[b]_+ = \max\{0, b\}$ and $\imath_{P\|Q}(t) \coloneqq \log \frac{\mathrm{d}P}{\mathrm{d}Q}(t)$ denotes the *information density* between P and Q. The E_γ-divergence is in fact an f-divergence associated with $f(t) = (t-\gamma)_+$ and it also satisfies that $\mathsf{E}_1(P\|Q) = \mathsf{TV}(P, Q)$. The next theorem provides a relation between this divergence and (ε, δ)-DP.

Theorem 1 ([29, 30]). *A mechanism $\mathcal{M}$ is (ε, δ)-DP if and only if, for all $\mathbb{D} \sim \mathbb{D}'$,*

$$\mathsf{E}_{e^\varepsilon}(P_\mathbb{D} \| P_{\mathbb{D}'}) \leq \delta.$$

By relating DP to E_γ-divergence, this theorem enables us to invoke the SDPI relationship (2), specialized to E_γ-divergence, to obtain the DP parameters ε and δ of iterative processes. To do so, we first need to compute the contraction coefficient under E_γ-divergence, which is addressed in the next section.

Authorized licensed use limited to: Harvard Library. Downloaded on November 18,2020 at 14:41:08 UTC from IEEE Xplore. Restrictions apply.

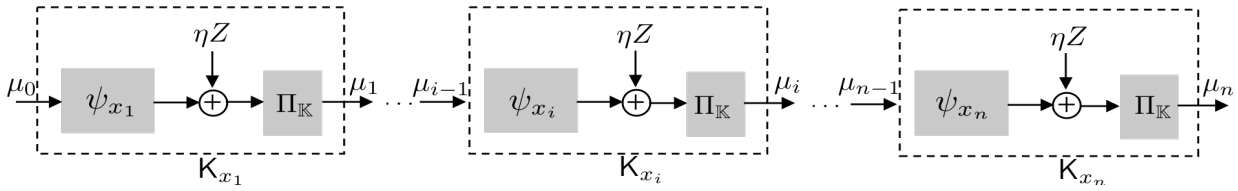

Fig. 1. The schematic representation of the projected noisy stochastic gradient descent described algorithm in Algorithm 1. Given μ_0 an arbitrary distribution on $\mathbb{K}$ and dataset $\mathbb{D} = \{x_1, \dots, x_n\}$, the i-th iteration is encoded by a projected additive kernel K_{x_i} given by $y \mapsto \Pi_{\mathbb{K}}(\psi_{x_i}(y) + \eta Z_i)$ where $\psi_{x_i}(y) = y - \eta \nabla_y \ell(y, x_i)$. The output distribution of kernel K_{x_i} is $\mu_i = \mu_0 \mathsf{K}_{x_1} \dots \mathsf{K}_{x_{i-1}}$.

III. Contraction of E_γ-Divergence

In this section we establish a closed-form expression for the contraction coefficient of kernels under E_γ-divergence that generalizes the Dobrushin's theorem in (4). We then instantiate this expression to introduce a family of practically-appealing kernels with compact input alphabet. For ease of notation, we let $\eta_\gamma(\mathsf{K}) := \eta_{\mathsf{E}_\gamma}(\mathsf{K})$.

Theorem 2. *For any $\gamma \geq 1$, we have*

$$\eta_\gamma(\mathsf{K}) = \sup_{y_1, y_2 \in \mathcal{Y}} \mathsf{E}_\gamma(\mathsf{K}(y_1) \| \mathsf{K}(y_2)). \tag{8}$$

Note that the Dobrushin's theorem [16] given in (4) corresponds to the special case of $\gamma = 1$ in Theorem 2. This theorem implies that in order to compute the contraction coefficient of a Markov kernel K under E_γ-divergence, one needs to compute E_γ-divergence between $\mathsf{K}(y_1)$ and $\mathsf{K}(y_2)$ for any $y_1, y_2 \in \mathcal{Y}$. The following lemmas are useful for such task. For $m \in \mathbb{R}$ and $v > 0$, let $\mathcal{L}(m, v)$ denote the Laplace distribution with mean m and variance $2v^2$.

Lemma 1. *For $m_1, m_2 \in \mathbb{R}$ and $v > 0$, we have*

$$\mathsf{E}_\gamma(\mathcal{L}(m_1, v) \| \mathcal{L}(m_2, v)) = \left[1 - e^{\frac{v \log(\gamma) - |m_1 - m_2|}{2v}}\right]_+. \tag{9}$$

For $m \in \mathbb{R}^d$ and $\sigma > 0$, let $\mathcal{N}(m, \sigma^2 \mathbf{I}_d)$ denote the multivariate Gaussian distribution with mean m and covariance matrix $\sigma^2 \mathbf{I}_d$.

Lemma 2. *For $\mathcal{N}_i = \mathcal{N}(m_i, \sigma^2 \mathbf{I}_d)$, $i = 1, 2$, we have*

$$\mathsf{E}_\gamma(\mathcal{N}_1 \| \mathcal{N}_2) = \mathsf{Q}\left(\frac{\log \gamma}{\beta} - \frac{\beta}{2}\right) - \gamma \mathsf{Q}\left(\frac{\log \gamma}{\beta} + \frac{\beta}{2}\right),$$

where $\mathsf{Q}(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2} du$ and $\beta = \frac{\|m_2 - m_1\|}{\sigma}$.

The previous lemma motivates the following definition.

Definition 1. *For $\gamma \geq 1$, we define $\theta_\gamma : [0, \infty) \to [0, 1]$ by*

$$\begin{aligned} \theta_\gamma(r) &:= \mathsf{E}_\gamma\left(\mathcal{N}(ru, \mathbf{I}_d) \| \mathcal{N}(0, \mathbf{I}_d)\right) \\ &= \mathsf{Q}\left(\frac{\log \gamma}{r} - \frac{r}{2}\right) - \gamma \mathsf{Q}\left(\frac{\log \gamma}{r} + \frac{r}{2}\right), \end{aligned} \tag{10}$$

where $u \in \mathbb{R}^d$ is any vector of unit norm.

With this definition at hand, we can write

$$\mathsf{E}_\gamma(\mathcal{N}(m_1, \sigma^2 \mathbf{I}_d) \| \mathcal{N}(m_2, \sigma^2 \mathbf{I}_d)) = \theta_\gamma\left(\frac{\|m_2 - m_1\|}{\sigma}\right). \tag{11}$$

It is worth pointing out that θ_γ has a similar role as the function R_α introduced by Feldman *et al.* in [12].

The additive Gaussian kernel $\mathsf{K} : \mathbb{R}^d \to \mathbb{R}^d$ is the kernel determined by $\mathsf{K}(y) = \mathcal{N}(y, \sigma^2 \mathbf{I}_d)$ for some $\sigma > 0$. This kernel models the privacy mechanisms which map $y \mapsto y + Z$ with $Z \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. An application of Theorem 2 and Lemma 2 shows that, under E_γ-divergence, the contraction coefficient of the additive Gaussian kernel is trivial[1], i.e., $\eta_\gamma(\mathsf{K}) = 1$. A similar conclusion holds for the additive Laplace kernel[2] which is determined by $\mathsf{K}(y) = \mathcal{L}(y, v)$ for some $v > 0$.

Fortunately, the input and output of kernels appearing in applications tend to be bounded. Think, for example, of the kernel which models the update of the weights of a neural networks during its training. In this case, the weights are bounded either by design or by regularization mechanisms. Motivated by this observation, we say that a kernel $\mathsf{K} : \mathbb{K} \to \mathbb{K}$ is the projected additive Gaussian kernel if it models the mechanism which maps $y \mapsto \Pi_{\mathbb{K}}(y + Z)$ where $\mathbb{K} \subset \mathbb{R}^d$ is compact and convex, $\Pi_{\mathbb{K}}$ is the projection operator onto $\mathbb{K}$ and $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ for some $\sigma > 0$. Similarly, we say that a kernel $\mathsf{K} : \mathbb{K} \to \mathbb{K}$ is the projected additive Laplace kernel if it models the mechanism which maps $y \mapsto \Pi_{\mathbb{K}}(y + L)$ where $L \sim \mathcal{L}(0, v)$ for some $v > 0$. These construction of kernels will be instrumental in the analysis of privacy guarantee of iterative processes in the next section.

IV. Analysis of Iterative Mechanisms via Contraction Coefficients

In this section, we consider iterative processes that can be decomposed into projected additive kernels. This constraint allows us to analyze the evolution of such processes through the lens of contraction coefficients.

A. General Setting

Recall the stochastic optimization setting in Example 2, where $\mathcal{Y} \subset \mathbb{R}^d$ is a parameter space and $\mathbb{D} = \{x_1, \dots, x_n\}$ is a dataset. In this context, an iterative stochastic optimization method is fully characterized by a set of kernels $\{\mathsf{K}_x : x \in \mathcal{X}\}$

[1]This is not surprising given the facts that $\eta_{\mathsf{TV}}(\mathsf{K}) = 1$ for any Gaussian channels K without input constraints [17] and $\eta_{\mathsf{TV}}(\mathsf{K}) = 1$ if and only if $\eta_f(\mathsf{K}) = 1$ for all non-linear functions f [31].

[2]The Euclidean norm of a d-dimensional Laplace noise vector is of order $d \log d$, see, e.g., [32, Thm. 2]. This asymptotic behavior makes Laplacian noise vectors highly inefficient for privacy purposes in the high dimensional setting. Therefore, in this paper we focus on the 1-dimensional case.

Authorized licensed use limited to: Harvard Library. Downloaded on November 18,2020 at 14:41:08 UTC from IEEE Xplore. Restrictions apply.

with $\mathsf{K}_x : \mathcal{Y} \to \mathcal{P}(\mathcal{Y})$. The following lemma provides an upper bound for the f-divergence between the parameters returned when using two neighboring datasets.

Lemma 3. *Let $\mu_0 \in \mathcal{P}(\mathcal{Y})$ and $\{\mathsf{K}_x : x \in \mathcal{X}\}$ be a family of kernels over $\mathcal{Y}$. If $\mathbb{D} = \{x_1, \ldots, x_n\}$ and $\mathbb{D}' = \{x'_1, \ldots, x'_n\}$ are neighbouring datasets with $x_i \neq x'_i$ for some $i \in [n]$, then*

$$D_f(\mu_0\mathsf{K}_{x_1} \cdots \mathsf{K}_{x_n} \| \mu_0\mathsf{K}_{x'_1} \cdots \mathsf{K}_{x'_n}) \leq D_f(\mu_{i-1}\mathsf{K}_{x_i} \| \mu_{i-1}\mathsf{K}_{x'_i}) \prod_{t=i+1}^{n} \eta_f(\mathsf{K}_{x_t}),$$

where $\mu_{i-1} \coloneqq \mu_0\mathsf{K}_{x_1} \cdots \mathsf{K}_{x_{i-1}}$.

While the previous lemma follows from a routine application of the strong data processing inequality, it provides a natural framework to study the privacy guarantees of iterative optimization methods. In the following, we use it to study the privacy properties of the projected noisy stochastic gradient descent algorithm and some of its variations.

B. Projected Noisy Stochastic Gradient Descent

We now apply Lemma 3 to study the projected noisy stochastic gradient descent (PNSGD) algorithm under two different noise densities: Laplacian and Gaussian.

Assume that $\mathbb{K}$ is a compact and convex subset of $\mathbb{R}^d$ and that $\ell : \mathbb{K} \times \mathcal{X} \to \mathbb{R}_+$ is a loss function differentiable in its first argument. In the literature it is customary to assume regularity conditions on the loss function [12, 15]. We make the following assumptions on the loss function:

- $y \mapsto \ell(y, x)$ is L-Lipschitz for all $x \in \mathcal{X}$,
- $y \mapsto \nabla_y \ell(y, x)$ is β-Lipschitz for all $x \in \mathcal{X}$,
- $y \mapsto \ell(y, x)$ is ρ-strongly convex for all $x \in \mathcal{X}$.

For a given a dataset $\mathbb{D} = \{x_1, \ldots, x_n\}$, the PNSGD algorithm starts from a given point $Y_0 \sim \mu_0 \in \mathcal{P}(\mathbb{K})$ (an arbitrary initial distribution) and then updates it according to the rule

$$Y_{t+1} = \Pi_{\mathbb{K}}(Y_t - \eta[\nabla_y \ell(Y_t, x_{t+1}) + Z_{t+1}]), \tag{12}$$

where $\Pi_{\mathbb{K}} : \mathbb{R}^d \to \mathbb{K}$ is the projection operator onto $\mathbb{K}$, $\eta > 0$ is the learning rate and $\{Z_t\}_{t=1}^n$ is a collection of i.i.d. noise variables sampled from a distribution absolutely continuous w.r.t. the Lebesgue measure. The PNSGD algorithm is summarized in Algorithm 1.

Algorithm 1 PNSGD Algorithm

Require: Dataset $\mathbb{D} = \{x_1, \ldots, x_n\}$, learning rate $\eta > 0$, initial point $Y_0 \sim \mu_0 \in \mathcal{P}(\mathbb{K})$ and i.i.d. copies $\{Z_t\}_{t=1}^n$ of a r.v. Z
for $t \in \{0, \ldots, n-1\}$ **do**
  $Y_{t+1} = \Pi_{\mathbb{K}}(Y_t - \eta[\nabla_y \ell(Y_t, x_{t+1}) + Z_{t+1}])$
end for
return Y_n

For any $x \in \mathcal{X}$, let $\psi_x(y) \coloneqq y - \eta\nabla_y \ell(y, x)$. Notice that $y \mapsto \Pi_{\mathbb{K}}(\psi_x(y) + \eta Z)$ is encoded by the projected additive Laplacian (resp., Gaussian) kernel if Z is Laplacian (resp., Gaussian) noise variable. Given the dataset $\mathbb{D} = \{x_1, \ldots, x_n\}$, one can therefore view the i-th iteration of the PNSGD algorithm as a projected kernel $\mathsf{K}_{x_i} : \mathbb{K} \to \mathcal{P}(\mathbb{K})$ that models the mapping

$$y \mapsto \Pi_{\mathbb{K}}(\psi_{x_i}(y) + \eta Z),$$

where Z is the common distribution of $\{Z_t\}_{t=1}^n$. If $Y_1, \ldots, Y_n$ are produced by the PNSGD algorithm with $Y_0 \sim \mu_0$, then, for all $t \in [n]$, we have

$$Y_t \sim \mu_t = \mu_0\mathsf{K}_{x_1} \cdots \mathsf{K}_{x_t}.$$

This allows us to express the PNSGD algorithm as a concatenation of n channels, as illustrated in Fig. 1.

Before delving into the privacy analysis of PNSGD, it is important to pause and adapt the definition of differential privacy to PNSGD setting. We recall from [12] that a mechanism $\mathcal{M}$ is (ε, δ)-DP for its ith input if $\mathsf{E}_{e^\varepsilon}(P_{\mathbb{D}} \| P_{\mathbb{D}'}) \leq \delta$ for any pair of datasets $\mathbb{D}$ and $\mathbb{D}'$ differing on the i-th coordinate. Specializing Lemma 3 to E_γ-divergence, we can say that the PNSGD algorithm is (ε, δ)-DP for its i-th input if

$$\mathsf{E}_{e^\varepsilon}(\mu_0\mathsf{K}_{x_1} \cdots \mathsf{K}_{x_n} \| \mu_0\mathsf{K}_{x'_1} \cdots \mathsf{K}_{x'_n}) \leq \mathsf{E}_{e^\varepsilon}(\zeta_{x_i} \| \omega_{x_i}) \prod_{t=i+1}^{n} \eta_{e^\varepsilon}(\mathsf{K}_{x_t}), \tag{13}$$

where $\zeta_{x_i} \coloneqq \mu_{i-1}\mathsf{K}_{x_i}$ and $\omega_{x_i} \coloneqq \mu_{i-1}\mathsf{K}_{x'_i}$. Assuming Z is either Laplacian or Gaussian, we can compute $\eta_{e^\varepsilon}(\mathsf{K}_{x_t})$.

C. Laplacian Projected Noisy Stochastic Gradient Descent

Here, we consider the PNSGD algorithm with Laplacian noise; i.e., $Z \sim \mathcal{L}(0, v)$ for some $v > 0$. The following theorem establishes the ε-DP property of such algorithm. For L, β, and ρ given in Section IV-B, define

$$M \coloneqq \sqrt{1 - \frac{2\eta\beta\rho}{\beta + \rho}}. \tag{14}$$

Theorem 3. *Assume that $\mathbb{K} = [a, b]$ for some $a < b$. Then PNSGD algorithm with Laplace noise is (ε, δ)-DP for its i-th input where $\varepsilon \geq 0$ and*

$$\delta = \left(1 - e^{\frac{\varepsilon}{2} - \frac{L}{v}}\right)_+ \left(1 - e^{\frac{\varepsilon}{2} - \frac{M(b-a)}{2\eta v}}\right)_+^{n-i}.$$

Consequently, we have $\delta = 0$ if $\varepsilon \geq \min\{\frac{2L}{v}, \frac{M(b-a)}{\eta v}\}$.

The use of Laplacian noise to provide ε-DP for SGD algorithms was extensively studied, see e.g., [32–34]. Unlike previous results, Theorem 3 is the first result regarding the privacy guarantees of PNSGD with Laplacian noise in the distributed-oriented framework proposed by Feldman et al. [12]. It is worth pointing out that our approach seems conceptually simpler than the approaches employed in [12, 15].

D. Gaussian Projected Noisy Stochastic Gradient Descent

Next, we assume that the noise distribution in the PNSGD algorithm is Gaussian, i.e., $Z \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$.

Authorized licensed use limited to: Harvard Library. Downloaded on November 18,2020 at 14:41:08 UTC from IEEE Xplore. Restrictions apply.

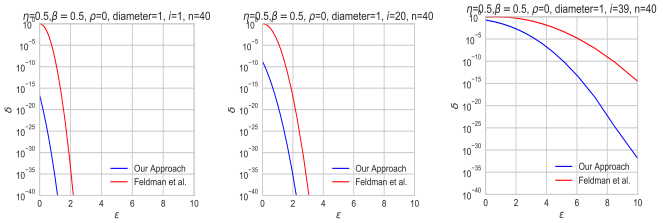

Fig. 2. The privacy parameters ε and δ of PNSGD with Gaussian noise having $\sigma = 2$ and loss function with parameter $L = 1$ and $\beta = 0.5$, computed using both Theorem 4 and Balle et al. [15, Theorem 5] for $i = 1$, $i = 20$, and $i = 39$ in dataset of size $n = 40$. Other parameters are as follows: $\eta = 0.5, \rho = 0$, and $D_{\mathbb{K}} = 1$.

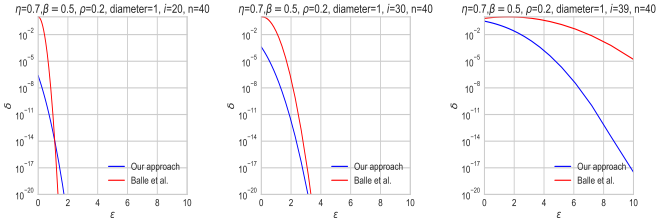

Fig. 3. The privacy parameters ε and δ of PNSGD with Gaussian noise having $\sigma = 1$ and strongly convex loss function ($\rho = 0.2$), computed using both Theorem 4 and Balle et al. [15, Theorem 5] for $i = 20$, $i = 30$, and $i = 39$ in a dataset of size $n = 40$. Other parameters are as follows: $\eta = 0.7, L = 1, \beta = 0.5$, and $D_{\mathbb{K}} = 1$.

Theorem 4. *Let $\mathbb{K} \subset \mathbb{R}^d$ be a compact and convex set. The PNSGD algorithm with Gaussian noise is (ε, δ)-DP for its i-th input where $\varepsilon \geq 0$ and*

$$\delta = \theta_{e^\varepsilon}\left(\frac{2L}{\sigma}\right)\theta_{e^\varepsilon}\left(\frac{MD_{\mathbb{K}}}{\eta\sigma}\right)^{n-i}.$$

Compared to Laplacian, the Gaussian perturbation has a better utility in high-dimensional setting, as illustrated in [32]. Hence, it has extensively appeared in DP literature as a *de facto* mechanism for providing privacy guarantees in training deep learning models [5]. Gaussian distribution is, in particular, appealing in the case of RDP as the Rényi divergence between two Gaussian distributions has a simple form (as opposed to E_γ-divergence). This intuition, among others, led Feldman et al. [12] and Balle et al. [15] to adopt RDP to examine the PNSGD algorithm with Gaussian noise in the framework of privacy amplification by iteration. While the former studied the problem for cases where $M = 1$ (i.e., $\rho = 0$), the latter assumed $M < 1$ (i.e., $\rho > 0$) and derived strictly better bounds for RDP guarantees. In fact, [15, Theorem 5] reduces to [12, Theorem 23] when $\rho = 0$. We wish to compare Theorem 4 with these results with or without strong convexity. To do so, we first need to convert the RDP guarantee given in [15, Theorem 5] to (ε, δ)-DP. This conversion is a standard practice in DP literature and follows from an straightforward application of [7, Proposition 3].

Proposition 1 (Adapted from [15])**.** *The PNSGD algorithm with Gaussian perturbation is $(\varepsilon, \tilde{\delta})$-DP for its i-th input where $\varepsilon > \kappa$ and*

$$\tilde{\delta} = e^{-\frac{1}{4\kappa}(\varepsilon-\kappa)^2}, \tag{15}$$

where $\kappa = \frac{2L^2}{(n-i)\sigma^2}M^{(n-i+1)}$ if $i \in [n-1]$ and $\kappa = 2\frac{L^2}{\sigma^2}$ if $i = n$.

Note that δ in Theorem 4 is given in terms of Q function and hence it is challenging to analytically compare δ with $\tilde{\delta}$. Nevertheless, we provide several numerical comparisons. In Fig. 2, we compare δ in Theorem 4 with $\tilde{\delta}$ in Proposition 1 for the first ($i = 1$), middle ($i = 20$) and the second to last ($i = 39$) individuals in a dataset of size $n = 40$ and $\sigma = 2$ with the assumption that the loss function is not strictly convex (i.e., $\rho = 0$). As clearly seen, our approach outperfoms [12] especially for the individuals whose records were processed later in the algorithm.

In Fig. 3, we focus on the effect of strong convexity parameter ρ on the privacy guarantee. We again depict δ and $\tilde{\delta}$ for the second half of the dataset: $i = 20$, $i = 30$, and $i = 39$ in a dataset of size $n = 40$ and $\sigma = 1$. Here, we assume that the loss function is strictly convex with parameter $\rho = 0.2$. As observed in this case, Theorem 4 provides better privacy in the high privacy region (i.e., small ε) as well as for the individuals who appear later in the dataset for all privacy region.

E. Randomly Stopped PNSGD Algorithm

We end this section by pointing out a potential shortcoming of Theorems 3 and 4 (and in general the privacy amplification by iteration framework): different individuals participating in the dataset experience different privacy guarantees; that is, individuals whose records were processed earlier experience higher privacy guarantee. This may not be justified in practice. To address this issue, we follow [12] to consider the *random stopping* for the PNSGD algorithm: namely, instead of iterating for n steps, we pick a random time T uniformly on $[n]$, stop the algorithm after T steps and then output Y_T. The following theorem illuminates that such algorithm in fact *uniformizes* the privacy guarantee among all individuals.

Theorem 5. *Let $\mathbb{K} \subset \mathbb{R}^d$ be a compact and convex set. The randomly-stopped PNSGD algorithm with Gaussian noise is (ε, δ)-DP with $\varepsilon \geq 0$ and*

$$\delta = \frac{1}{n}\theta_{e^\varepsilon}\left(\frac{2L}{\sigma}\right)\left(1 - \theta_{e^\varepsilon}\left(\frac{MD_{\mathbb{K}}}{\eta\sigma}\right)\right)^{-1}. \tag{16}$$

The randomly stopped PNSGD was first proposed by Feldman et al. [12] where they derived its RDP guarantee in [12, Theorem 26] *only* if σ satisfies a certain constraint. This constraint is due to the non-convexity of the map $(\mu, \nu) \mapsto D_\alpha(\mu\|\nu)$. In contrast, since $(\mu, \nu) \mapsto \mathsf{E}_\gamma(\mu\|\nu)$ is jointly convex (as for any other f-divergences), Theorem 5 holds for any σ.

Another approach to address the non-uniformity of privacy guarantees is to *permute* the dataset first, via a random permutation and then feed it to the PNSGD algorithm. We will examine this approach in our future work.

Authorized licensed use limited to: Harvard Library. Downloaded on November 18,2020 at 14:41:08 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

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

[2] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *EUROCRYPT*, S. Vaudenay, Ed., 2006, pp. 486–503.

[3] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375–389, 2010.

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

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

[6] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *NeurIPS*, 2018, pp. 6280–6290.

[7] I. Mironov, "Rényi differential privacy," in *Proc. Computer Security Found. (CSF)*, 2017, pp. 263–275.

[8] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," vol. abs/1603.01887, 2016. [Online]. Available: http://arxiv.org/abs/1603.01887

[9] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography*, 2016, pp. 635–658.

[10] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke, "Composable and versatile privacy via truncated cdp," in *STOC*, 2018, pp. 74–86.

[11] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, Jun. 2011.

[12] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," *FOCS*, pp. 521–532, 2018.

[13] M. Raginsky, "Strong data processing inequalities andϕ-sobolev inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.

[14] A. Makur and L. Zheng, "Bounds between contraction coefficients," 2018. [Online]. Available: http://arxiv.org/abs/1510.01844

[15] B. Balle, G. Barthe, M. Gaboardi, and J. Geumlek, "Privacy amplification by mixing and diffusion mechanisms," in *NeurIPS*, 2019, pp. 13 277–13 287.

[16] R. L. Dobrushin, "Central limit theorem for nonstationary markov chains. I," *Theory Probab. Appl.*, vol. 1, no. 1, pp. 65–80, 1956.

[17] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 35–55, Jan 2016.

[18] S. Asoodeh, M. Diaz, and F. P. Calmon, "Privacy amplification of iterative algorithms via contraction coefficients," *arXiv 2001.06546*, 2020.

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

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

[21] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, 12 1976.

[22] I. Sason and S. Verdú, "f -divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov 2016.

[23] A. Kontorovich and M. Raginsky, "Concentration of measure without independence: A unified approach via the martingale method," in *Convexity and Concentration*. Springer New York, 2017, pp. 183–210.

[24] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Mathematical Society, 2006.

[25] P. M. del Moral, M. Ledoux, and L. Miclo, "On contraction properties of markov kernels," *Probability Theory and Related Fields*, vol. 126, pp. 395–420, 2003.

[26] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

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

[28] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Commun. Inf. Theory*, vol. 1, no. 4, pp. 417–528, Dec. 2004.

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

[30] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *ICML*, vol. 80, 10–15 July 2018, pp. 394–403.

[31] J. Cohen, J. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices, with Applications in Information Theory, Statistics, Economics, and Population Sciences*. Birkhäuser, 1998.

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

[33] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization, revisited," in *ICML 2014 Workshop on Learning, Security and Privacy*, Beijing, China, 25 Jun 2014.

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

Authorized licensed use limited to: Harvard Library. Downloaded on November 18,2020 at 14:41:08 UTC from IEEE Xplore. Restrictions apply.